SPECIAL TOPICS IN COMPUTING AND ICT RESEARCH

Strengthening the Role of ICT in Development

Editors

Kizza, Joseph Migga
Jackson Muhirwe
Janet Aisbett
Katherine Getao
Victor W. Mbarika
Dilip Patel
Anthony J. Rodrigues

Volume III

FOUNTAIN PUBLISHERS

Kampala

Fountain Publishers
P.O. Box 488
Kampala
E-mail:fountain@starcom.co.ug
Website:www.fountainpublishers.co.ug

Distributed in Europe, North America and Australia by African Books Collective Ltd (ABC), Unit 13, Kings Meadow, Ferry Hinksey Road, Oxford OX2 0DP, United Kingdom.
Tel: 44(0) 1865-726686, Fax:44(0)1865-793298.
E-mail: abc@africanbookscollective.com
Website: www.africanbookscollective.com

First published 2007

ISBN 978-9970-02-730-9

Contents

Part 2: Information Systems

Part 3: Information Technology

Preface

The last ten years have been years of excitement in Africa as many African peoples saw and actually enjoyed unprecedented inputs of information technology. At last the dreams of many were coming true. After years of stagnation in sectors like communication, banking, and commerce, new frontiers are opening. One can be in any part of Africa, however remote and call somebody else within a few seconds. People are banking online, emailing and text messaging as if the "dark" continent has lost its bite. However exciting all this may be, there are still problems that must be addressed if Africa is to keep leap froging into the 22nd Century. Modern and most durable infrastructures must be built, a formidable task that requires individual, national, and multinational effort.

Driven by this need and others, the Makerere University Faculty of Computing and Information Technology (CIT) has been holding annual conferences, now, three years in a row, intended to build and sustain, and share in the gains attained so far in information communications technology (ICT) infrastructure and capacity-building throughout Africa and beyond. The goal is to create an international forum for discussion of ideas, proven solutions and best practices to Africa's development problems that can be easily adapted and utilised to build and sustain an ICT environment and infrastructure for Africa's development.

In its third year, the Makerere University Faculty of Computing and Information Technology's International Conference on Computing and ICT Research (SREC 07) brought together interested academicians, developers and practitioners for four days at the Makerere University campus to discuss, share ideas and experiences, and develop and recommend strategies for ICT-based development. The conference attracted more than 100 participants with over thirty scholarly papers presented covering a wide array of topics with relevance to ICT-based development in six tracks: Computer Science, Information Systems, Information Technology, Data Communications and Networks, Software Engineering, and Information Communication Technology for Sustainable Development.

Organising a conference like this one requires considerable effort and sacrifice. This is exactly what all the people on the organising committee did. We are grateful for their dedication and hard work, without which the conference and book would not have been possible. We are also grateful to the publishing team at Fountain Publishers, Kampala. Lastly special thanks to Jackson Muhirwe, Secretary to the conference and his team for their dedication, which made the conference the success that it was.

Joseph M. Kizza
Tennessee, USA.

Editors

Kizza, Joseph Migga received a BSc in Mathematics and Computer Science from Makerere University, Kampala, Uganda, a MSc in Computer Science from California State University, USA, an M.A. in Mathematics from the University of Toledo, Ohio, USA and a PhD in Computer Science from the University of Nebraska-Lincoln, Nebraska, USA. Dr Kizza is currently a professor of Computer Science and Director of the Center for Information Security and Assurance at the University of Tennessee-Chattanooga, Tennessee, USA. His research interests are in Social Computing and Computer Network Security in which he has so far published numerous conference and journal articles and more than seven books. He was appointed a UNESCO expert in Information Technology in 1994.

Jackson Muhirwe earned his BSc degree majoring in Mathematics and a MSc in Computer Science degree from Makerere University, Kampala, Uganda. Muhirwe is a PhD student in the Department of Computer Science, Faculty of Computing and IT, Makerere University and he is currently writing his thesis on Computational Analysis of Bantu language Morphology. His main areas of research interests are computational linguistics and open source software. Muhirwe is employed in the Faculty of Computing and IT, Makerere University, as a Research Coordinator and his main duty is to coordinate and manage research projects for (post) graduate students. In addition he coordinates all activities related to the International Conference on Computing and ICT Research and the International Journal of Computing and ICT Research.

Janet Aisbett has a PhD in Pure Mathematics from the University of Western Ontario. She worked in signal and image processing, before becoming interested in problems associated with information storage and retrieval. For the last decade her theoretical research has concerned effects of representation on information exchange and cognition. Her applied research has focused on integration and evaluation of health systems and electronic commerce. She has held various administrative posts within the University of Newcastle, including Deputy President of Academic Senate and membership of the University's Council, and within regional and national groups involved in encouraging IT skills development.

Katherine Getao is the Director of the School of Computing and Informatics, University of Nairobi. Her interests include artificial intelligence and natural language applications for Bantu languages. She supervises a number of undergraduate, Postgraduate students. She is a joint supervisor of the first author.

Victor W. Mbarika is on faculty in the College of Business at Southern University and A&M College. He is an information and communications technology (ICT) consultant with various governmental and private agencies. He holds a BSc in Management Information Systems from the US International University, a MSc in MIS from the University of Illinois at Chicago, and a PhD in MIS from Auburn

University. With over 80 academic publications, Dr. Mbarika's research on ICT diffusion in developing countries and his research on multimedia learning has appeared (or are forthcoming) in 32 academic journals. He was cited as being "in the forefront of academic research into ICT implementation in Africa, and has provided a theoretically-informed framework for understanding ICTs in less developed countries..." His scholarly publications have appeared in journals such as IEEE Transactions, IEEE Spectrum, Communications of the ACM, Journal of the Association for Information Systems, Information Systems Journal, The Information Society, Journal of the American Society for Information Sciences and Journal of Global Information Management. Most of his publications are on information technology transfer to developing countries. Additionally, he has published four book chapters and over 50 national and international conferences publications on Information Systems. He is author of 3 academic books. He serves as a senior board member (co-editor for the African region) of the International Journal of Health Information Systems & Informatics, as well as editorial board member of several other journals. He also serves as associate editor for IEEE Transactions on IT in Biomedicine and reviewer for IEEE Transactions on Engineering Management. Dr. Mbarika is Founder of the Cameroon Computer and Network Center (CCNC). He is a member of the Association of Information Systems (AIS), the Institute of Electrical and Electronics Engineers (IEEE), and the Information Resources Management Association (IRMA). He holds over 15 research, teaching and service excellence awards. Email: victor@mbarika.com; Website: www.mbarika.com.

Dilip Patel holds the chair of Information Systems and is the head of the Centre for Information Management and E-Business, which consist of four research groups: E-Business and The Digital Society, Health Informatics, Information Management and Modelling and Knowledge Based Systems. Professor Dilip Patel is currently on the editorial board for two international journals: The International Journal of Cognitive Informatics and Natural Intelligence (IJCiNi) and the International Journal of Information Technology and Web Engineering. He is also Editor-in-Chief of The International Journal of Computing and ICT Research.

Anthony J. Rodrigues is Professor of Computer Science at the School of Computing and Informatics, University of Nairobi. Research interests include Scientific Computing, Approximation Theory, Error Analysis and Systems Modelling. He has also been involved in the design and development of sustainable integrated management information systems at various universities and is currently studying the developmental impact of various ICT policies and strategies (or the lack thereof) in the region.

Contributors

Julianne Sansa holds a BSc (Maths, Computer Science) and MSc (Computer Science) from Makerere University. Since 2001 she has worked with the Faculty of Computing and IT of Makerere University in various capacities. She is currently registered as a PhD Student at the University of Groningen, in the Netherlands and her research interests are Quality of Service and Network protocols.

Baryamureeba, Venansius is the Dean and Professor of Computer Science at the Faculty of Computing and Information Technology, Makerere University, Uganda. Baryamureeba holds a PhD in Computer Science. He is currently the Executive President of Uganda National Academy of Computing and Information Technology, an association that brings together all computing and information technology faculties/ institutes in all public and private universities to address common objectives and challenges. Baryamureeba is also Chairman and Managing Director of ICT Consults Ltd (http://www.ict.co.ug), one of the prestigious ICT consultancy firms in Africa. He is a distinguished lecturer, research, administrator and manager. Baryamureeba lectures courses in Computing at all levels including Doctoral level. He has supervised several students at masters and PhD level in the area of Computing. Baryamureeba has published extensively in refereed journals and books.

John Ngubiri is an Assistant Lecturer at the Department of Computer Science, Faculty of Computing and Information Technology, Makerere University. His research interests are in performance evaluation, system optimization and parallel and distributed processing. He is currently a PhD Student at Radboud University Nijmegen – The Netherlands. He is in the Information Retrieval and Information Systems (IRIS) research group of the Nijmegen Institute for Informatics and Information Science (NIII), Faculty of Science, Mathematics, Computing Science. His promoter is Prof. dr. Mario van Vliet.

Jackson Muhirwe earned his BSc degree majoring in Mathematics and a MSc in Computer Science degree from Makerere University, Kampala, Uganda. Muhirwe is a PhD student in the Department of Computer Science, Faculty of Computing and IT, Makerere University and he is currently writing his thesis on Computational Analysis of Bantu language Morphology. His main areas of research interests are computational linguistics and open source software. Muhirwe is employed in the Faculty of Computing and IT, Makerere University, as a Research Coordinator and his main duty is to coordinate and manage research projects for (post) graduate students. In addition he coordinates all activities related to the International Conference on Computing and ICT Research and the International Journal of Computing and ICT Research.

Benjamin Kanagwa is an Assistant Lecturer and PhD student at the Faculty of Computing and Information Technology, Makerere University. His research interests are in the area of software architectures, service oriented software engineering, components and software reuse.

Narcis T. Rwangoga obtained his Masters Degree in Computer Science in 2004 from Central South University, Hunan Province, P. R. China. Mr. Rwangoga also holds a Post Graduate Diploma in Management from Uganda Management Institute (2000) and a Bachelors Degree in Agricultural Economics from Makerere University (1994). Mr. Rwangoga is currently employed at the position of Assistant Lecturer in the Computer Science Department at the Faculty of Computing and IT, Makerere University. He has been involved in teaching and research supervision for students in areas of Artificial Intelligence and Expert Systems, IT Strategic Planning and Management, Networked Systems and Information Security Management. He has professional certifications in different computer application fields, especially in Databases and Networking Technologies. Mr. Rwangoga's current research interests lie in the areas of Distributed Systems (Mobile Agents Technology, Distributed Databases), and Artificial Intelligence (Expert Systems, Neural□šetworks and Genetic Algorithms).

Ismail Wadembere holds a BSc Surveying degree from Makerere University and a MSc in planning - Geographical Information Systems (GIS) degree from University of Science Malaysia. Wadembere is a PhD student in the Department of Information Systems, Faculty of Computing and IT, Makerere University. His main areas of research interest are GIS, Service-Oriented developments and Online Information Systems. Wadembere is employed in Kyambogo University, where he handles GIS, Computing for Surveying and Mapping.

Agnes F. Namulindwa Lumala is an assistant lecturer at the faculty of computing & IT, Makerere University. She is a PhD candidate at the same university. Her research interests include: Composition of Software, Decision Support for Component-Based Software Development.

Daudi Jjingo holds an undergraduate degree in Biochemistry from Makerere University (2002) and a Masters degree (distinction) in Bioinformatics and Biological Computing from the University of Leeds, UK (2004). In the Laboratory of Prof. Wojchiech Makalowski at Penn State University in the US, he did research on the computational prediction of Transposable elements in the mouse genome. He is now in the Department of IT, where he lectures Compiler design theory to the MSc. Computer Science group and Information Systems Security to Undergraduates. He is a member of the Faculty's Health Informatics group, where he has strong research interests in Bioinformatics and Biological Computing, especially as it applies to African disease pathogens and other African health problems. Currently, he's doing work on prediction and characterization of Binding hotspots on putative drug targets in the Malaria proteome. His work

in this area has been selected for the All Africa Bioinformatics Conference, and his earlier research on mouse transposable elements has been selected for publication in the International Journal of Computing and ICT research. His research interests include Bio-informatics and computational biology of Infectious Diseases.

Wojciech Makalowski obtained his versatile education in Poland. He received masters degrees in biology and philosophy in 1983 and 1988, respectively, and PhD in molecular biology from A. Mickiewicz University in Poznan in 1991. Although, without a formal training in computer science, he started to use computational tools in his biological research early in the scientific career. In the late 1980s he worked on a sequence pattern recognition including tRNA-gene finding in genomic sequences and regulatory signals for gene transcription. After a short postdoc at the University de Montreal, he worked at the National Center for Biotechnology Information. In 2001 he joined Penn State University with the primary appointment at the Department of Biology. He serves on the editorial board of Genome Research, RNA Biology and International Journal of Biology. His current research is focused on comparative genomics and evolutionary analyses of eukaryotic genes. This includes development of new algorithms and software for evolutionary biology and genomics.

Rehema Baguma is a PHD student in Information Systems researching on Web accessibility. She holds a BLIS Hons, a PGD CS and an MSc in Computer Applications. She is employed in the Faculty of Computing and IT, Makerere University as an Assistant Lecturer and has skills in development informatics, networking and system administration

Nazir Ahmad Suhail is registered as a PhD (Information Technology) student at Faculty of Computing and Information Technology, Makerere University, Kampala-Uganda in August 2005. He is holder of B.Sc, M.Sc (Mathematics) degrees from The University of The Punjab, Lahore-Pakistan and a PGD(Computer Science) from Makerere University. He is also lecturing at Kampala University in the Departments of Computer Science and Mathematics and has also taught at Department of Mathematics Makerer e University , some time back.

John Aogon holds a Masters degree in Computer Science from Makerere University. He has worked in the telecommunications industry for the last eight years. He previously worked with Celtel Uganda and currently works in the Information Technology department with MTN Uganda, a leading telecommunications company in Uganda.

Suresh Shanmugasundaram is a PhD Research Scholar in the Department of Computer Science and Engineering at Government College of Technology, Coimbatore, India. He is currently involved in developing an Algorithm for Gridlock avoidance. He was employed as the Lecturer, Department of Information Technology and Placement Coordinator at VLB Janakiammal College of Engineering and Technology (www.vlbjcet.ac.in). Suresh graduated his

Bachelor of Engineering in the discipline of Computer Science and Engineering at Sri Ramakrishna Engineering College. Then he graduated his Master of Science in Computer Networks at Middlesex University, London. His Masters thesis was a technique for Secured Group Key Agreement. He has lectured Object Oriented Analysis and Design, Programming in C, Computer Networks, Management Information System, Enterprise Resource Planning, Software Engineering Methodologies, Network Security both for UG and PG Engineering students. In addition, as a Placement Coordinator he was responsible for liaisons with corporate. Currently he has been offered a position as Senior Faculty at NIIT, Botswana.

Paul Ssemaluulu obtained a FTC in Telecommunications from Uganda Polytechnic Kyambogo in 1985. He also holds a BBA from Makerere University in 2001, a MSc in Computer Science in 2006 and is a PhD student in Information Systems at Faculty of Computing and IT, Makerere University.

Ddembe Williams Is a Visiting Senior Lecturer in Information Systems at the Faculty of Computing and Information Technology (CIT), Makerere University, Uganda. He is also the Deputy Dean for Graduate Studies and Research. For the last seven years, Ddembe has been a Senior Lecturer in Decision Sciences and Head of Centre for Advanced Systems Modelling and Simulation within the Faculty of Business, Computing and Information Management (BCIM) at London South Bank University (LSBU) in the United Kingdom. Ddembe has an MSc in Advanced Information Technology Management and a PhD in Computer Science/ System Dynamics from London South Bank University. Ddembe's practical and professional work in teaching and consultancy focus on applied computing to systems requirements engineering, process analysis, problem structuring and model-based decision support systems. His theoretical/research work centres on the contribution that system dynamics/information systems can make to the advancement of systems modelling and simulation. Ddembe has published over 20 refereed conference and journal papers.

Gilbert Maiga Holds the following academic qualifications: MSc. Computer Science (Makerere University), MSc. Animal Physiology (UON), Bachelor of Veterinary Science (Makerere University), and Diploma in Computer Systems Design (WLC), Certificate in Computer Programming (WLC). Networking Basics Certificate (CISCO). The MSc. Computer Science project was on "Information systems integration in HIV-AIDS Management: A web-based database approach." He is registered for a PhD in the Department of Information Systems, Faculty of Computing and IT, Makerere University in August 2005, where he is also an assistant lecturer. Bioinformatics is his main area of interest for research. PhD Research Topic: "An Ontology-based Framework for Integrating Large Scale Biological and Clinical Information".

Patrick Ogao has been involved in visualization research since 1996. He has a PhD (Utrecht University, the Netherlands) and Msc (ITC Enschede, The Netherlands) focusing on visualization in information systems and dynamic geo-phenomena. Prior to joining Makerere University, he worked with the research group in scientific visualization and computer graphics in the department of computing and mathematics of the University of Groningen, The Netherlands. He is currently a visiting senior lecturer and acting Head of Information Systems Department in the Faculty of Computing and Information Technology, at Makerere University. He is the Long Term Expert (on secondment by the University of Groningen) in a NUFFIC/NPT project in ICT training capacity building in four public universities in Uganda. His research interests are in visualization applications in bio-informatics, geo-informatics, and software engineering and in developing e-strategies for developing countries.

Asifiwe RG Collins' career in education includes secondary teaching in government. His advocacy for computers in the classroom began in 2005; a short time after enrolling for the Masters course in Information and Communication Technology and has as a result carried out training sessions in ICT in various areas especially in secondary schools. He is currently a student at the University of Makerere doing his research as his final part of his Masters degree in Educational Information and Communication Technology. Mr. Asifiwe has had a teaching experience of more than five years in a number of schools and for the past three years to date he has been into active research with a number of research consultancies, both academic and social-economic. In his advocacy, he is committed to preparing young people to succeed in a world changing rapidly to that of the present day and sees the development of ICT for staff and students as fundamental in creating schools and institutions for the future. My wish is placed on developing staff capabilities and effective network and technical support infrastructure as precursors to inspiring teacher enthusiasm for integrating ICT into the curriculum.

Joseph Muliaro Wafula obtained BSc (Kenyatta University),M.Sc (University of Nairobi) and M.Phil (University of Cambridge UK). He is currently a lecturer at the Institute of Computer Science and Information Technology at Jomo Kenyatta University of Agriculture and Technology, Kenya and a Ph.D student researching on ICT policies and e-strategies.

Lorenzo Dalvit (Laurea in Sociology UNITN, Italy; MA Linguistics Rhodes, South Africa) is currently a Doctoral candidate in the Department of Education at Rhodes University. His main area of academic interest is access to Education for members of marginalised communities, with a particular focus on language issues and ICT Education. Mr Dalvit has published several articles and has presented several conference papers both in South Africa and internationally.

Hyppolite Muyingi (PhD Electrical Engineering Brussels, Belgium) is a Professor in the Department of Computer Science at the University of Fort Hare. His

main areas of academic interest are power utilities communications and ICT for development. Prof Muyingi has a long experience in teaching and research and he is the Head of the Fort Hare Centre of Excellence in Developmental eCommerce, one of the prime sponsors of the project.

Mamello Thinyane (MSc Computer Science Rhodes, South Africa) is currently a Doctoral candidate in the Department of Computer Science at Rhodes University. His main areas of academic interest are the creation of Internet services (including elearning), wireless networking and the integration of indigenous knowledge networks into ICT. Mr Thinyane has been one of the driving forces of the project and he helped to supervise most of the research work.

Richard Sekibule is an Assistant Lecturer in the Department of Computer Science, Faculty of Computing and Information Technology, Makerere University. He holds a BSc (Computer Science, Mathematics) from Makerere University Kampala, and an MSc (Computer Science) from Radboud University Nijmegen. His research interests are in the area of security in open distributed computing research.

Bulega Tonny Eddie is currently a PhD candidate and a research assistant of the communications and information systems Lab, Network and communication centre at South China University of Technology. He received his BSc Eng in 2000 and Msc (MEng/MIS) in 2004 respectively. Bulega won the distinguished international foreign students award for his research on ICT in 2005. He is currently a member of the consulting committee on the on the National fiber backbone between Chinese government and Uganda. Currently he is also working with the design and research centre of telecommunications of Guangdong. His research interests are; broadband access, Wireless LAN's, network security and multimedia networks and next generation networks.

Gang Wei is currently a professor, Dean and chair of Communications and Information Systems Engineering at South China University of Technology (SCUT). Besides these positions, his industry experience spans China Telecom, China Unicom, and China mobile. He has authored over 50 publications in the form of books, patents, and papers in refereed journals and conference proceedings. He has authored the textbook Information theory and source coding. He holds several patents in the area of wireless and nextgeneration networking. He has served on the editorial boards of IEEE Personal Communications and as conference chair for many international conferences and workshops. He received his PhD degree in electrical and computer engineering from University of South Califonia in 1997, and his Msc and Bsc from south China University of technology.

Fang-Jiong Chen is a vice professor and vice chair of communications and information systems at south China University of Technology. His research spans technologies such as CDMA, OFDM and WiMax. He received the 2005 Guangdong science award in ICT. He is currently a visiting researcher at Hong Kong University of science and technology working on smart antennas

Tom Wanyama received BSc. and MSc. Degrees in Electrical Engineering from Makerere University – Uganda, and the University of New South Wales – Australia in 1993 and 1995 respectively. In addition, he received a postgraduate certificate in university teaching in 2005, and a PhD in Electrical and Computer Engineering (majoring in Software Engineering) in 2006 from the University of Calgary – Canada. Tom Wanyama has over ten years of university teaching experience in Electrical, Communications and Computer Engineering Courses such as, Power Systems, Electrical Machines, Microwave Engineering, Data Communication and Computer Networks, Application Development in C#, Turbo Pascal, C++ and Java, Component-Based Software Development, and MATLAB simulations. Moreover, he has carried out research and worked in various areas of Electrical and Computer Engineering for over ten years. Currently, Tom Wanyama is a Lecturer in the Faculty of Technology – Makerere University, and a Research Advisor in the faculty of Computing and Information Technology – Makerere University.

Geoffrey Ocen is employed at the London Borough of Haringey as Head of Strategy. Areas of interests include ICT, community and economic development. The author previously worked as Director of Economic Development at a London Development Trust, Training Manager in London and finally as a Lecturer at Makerere University (Uganda) in the 1980s.

Michael Niyitegeka is an Assistant Lecturer in the Department of Information Systems, Faculty of Computing & IT, Makerere University. He is specifically interested in Technology Management as a research area and is currently working on his PhD proposal in the area of ICT and SME business development.

Yirsaw Ayalew is a lecturer at the Department of Computer Science, University of Botswana. He holds a PhD in computer science from the University of Klagenfurt, Austria and teaches both undergraduate and graduate courses in software engineering and database systems. His research interests are mainly in the area of software engineering; particularly software quality, requirements engineering, and application of soft computing techniques in software engineering.

Francis Otto holds a master of Engineering in Computer Apllication Technology from Central South University, China. He worked as an assistant lecturer at Uganda Christian University and was previously with the School of Information Science and Engineering, Central South University in Professor Ouyang Song's distributed and intelligent systems research group. Otto has authored Improving Search in Unstructured P2P Systems: Intelligent Walks (I - Walks) in LNCS 4224, pp. 1312-1319, Springer, 2006 and Enhanced master-slave time synchronization architecture for wireless sensor networks in ATNAC2006, Melbourne, Australia December. His research Interests include Intelligent distributed systems, P2P technology, Wireless Sensor Networks.

Drake Patrick Mirembe holds a Master of Science degree in Computer Science (major Security of Systems) from Radboud University, Nijmegen, the Netherlands. While pursuing his Masters, Drake worked with the Nijmegen Health Initiative and the Security of Systems groups under the supervision of Dr. Martijn Oostijdk and Dr. Perry Groot on the development of a secure framework for the implementation of telemedicine, e-health and wellness services. Drake is a co-author of "Enhanced master-slave time synchronization architecture for wireless sensor networks (ATNAC2006 Melbourne, Australia). Today, he works with the Department of Networks, Faculty of Computing & IT, Makerere University as an assistant Lecturer and researcher. His research interest include; Security of Systems, Health Informatics (HI), Peer-to-Peer Computing, Artificial Intelligence, E-voting, and Wireless communication technologies. Contacts: Email dmirembe@cit.mak.ac.ug. Tel: +256-712-844343; URL: www.drake.silwarriors.com

Maybin Muyeba is a member of the Intelligent and Distributed Systems (IDS) Research Laboratory. He was awarded a British council scholarship for best graduating student for BSc Mathematics in 1991 and was also awarded a scholarship for his PhD Degree at UMIST, in Data Mining. Dr Muyeba holds a BSc Honors (Real analysis, abstract algebra and complex analysis), University of Zambia and Master of Science in Computing, University of Hull. Prior to joining Liverpool Hope University, Dr Muyeba has worked for seven years as a lecturer and researcher abroad and in the UK. He spent five years at University of Zambia, a year each at UMIST, Computation Department as assistant lecturer and as full lecturer at the University of Bradford, Informatics Department. At UMIST, Dr Muyeba contributed to the development of new machine learning algorithms for discovering logic patterns in databases.

Agnes Nakakawa holds a Masters Degree in Computer Science (Computer Information Systems) of Makerere University. Under the supervision of Dr. Patrick Ogao, she carried out a research project entitled "Spatial Decision Support Tool For Landfill Site Selection For Municipal Solid Waste Management". She also holds a Bachelors Degree in Statistics (Statistical Computing) of Makerere University. Her research interests are in SDSSs, GISs, MCE and Collaborative Spatial Decision Making. She is an employee in the Department of Information Systems, the Faculty of computing and IT of Makerere University. Contacts: email: anakakawa@cit.mak.ac.ug. Tel: +256-772-690247

Mariam Sensalire is a PhD student in software Engineering researching in software visualization. She holds a BSc (Hons) and a MSc in Computer Science Degree. She is employed in the Faculty of Computing and IT, Makerere University as an Assistant Lecturer and has skills in networking and system administration.

Joel Lakuma is an Assistant Lecturer in the Department of Computer Science, Faculty of Computing and Information Technology, Makerere University. He holds a BSc (Telecom Engineering) from Trinity University, Terra Haute, IA,

USA; and an MSc (Computer Science) from Makerere University, Kampala, Uganda. His research interests are in Algorithms, and Computing Complexity.

Doreen Tuheirwe is currently a Systems and Software Engineering master student at Rijksuniversiteit Groningen. Her research interests include Systems and Software Engineering, Intelligent Systems, Mobile Communications and their Architecture and Infrastructure.

Wakabi, Peter Patrick Waiswa earned a Bachelor's degree in Statistics and Applied Economics and a Postgraduate Diploma in Computer Science from Makerere University and he holds a MSc. in Computer Science from National University of Science & Technology from the National University of Science and Technology (Zimbabwe). He is also a Certified Network Administrator (CNA). Wakabi is a PhD student in the Department of Computer Science, Faculty of Computing and IT, Makerere University and he is currently writing his thesis on Association Rule Mining Using Evolutionary Computing. His main areas of research interest are evolutionary computing, algorithms and complexity, artificial intelligence, data mining and knowledge discovery. Peter is currently employed by Uganda Electricity Transmission Company Limited as the Head of IT where his main duty is the strategic positioning of the company using ICTs. He previously worked as a lecturer at ICS Makerere University and Uganda Christian University; as Commonwealth IT Technical Adviser to the NCC in Tanzania and as an IT Consultant with PricewaterhouseCoopers, among others.

Paul Muyinda is lecturer of Information Technology in the Department of Distance Education, Institute of Adult and Continuing Education, Makerere University. He is also charged with the responsibility of pedagogically integrating ICTs into the distance learner student support system. He holds a Bachelor's Degree in Statistics from Makerere University obtained in 1995 and a Master of Computer Science and Applications degree from Shanghai University, People's Republic of China, obtained in 1998. Muyinda is a Certified Netware Administrator (CNA). His Master's Dissertation was titled "A National Passport IntranetWare: A Case of Uganda National Passport Office". He is currently a registered PhD (Information Systems) student at the Faculty of Computing and IT, Makerere University with research interest in e-learning. His thesis is titled "A Framework for Instantiating Learning Objects Applications for Mobile and Fixed Communication Systems".

Kitoogo Fredrick Edward is a holder of a Bachelor of Statistics degree from Makerere University, Kampala, Uganda and a MSc in Computer Science degree from the National University of Science and Technology (NUST), Bulawayo, Zimbabwe. Kitoogo is a PhD student in the Department of Computer Science, Faculty of Computing and IT, Makerere University and he is currently writing his thesis on Improved Use of Machine Learning Techniques in Named Entity Recognition. His main areas of research interests are Machine Learning, Natural Language Processing and Business Intelligence & Data Warehousing. Kitoogo

is formally employed in the Courts of Judicature, as the Principal Information Technology Officer where he heads the Information & Communications Technology Section. In addition he lectures a course on Business Intelligence & Data Warehousing at the Faculty of Computing & Information Technology, Makerere University, Kampala, Uganda.

Agnes Rwashana Semwanga is an Assistant Lecturer in the Information Systems Department, Faculty of Computing and Information Technology, Makerere University. She is currently undertaking PhD research (final year) and her main research interests include Simulation and Modeling, System Dynamics and Health Informatics.

Acknowledgements

We are grateful to those who contributed peer reviewed and edited papers to this book. The papers were accepted for presentation at the third International Conference on Computing and ICT Research (SREC 07), held 5 - 8 August 2007, at the Faculty of Computing and Information Technology, Makerere University, Kampala Uganda.

The editor would like to recognise and also give special thanks to the following:

- The different researchers around the world who reviewed the papers anonymously;
- The Conference Chair and the Chairpersons of the conference technical committees;
- Members of the conference organising committee and conference technical committees;
- The Conference Secretariat for supporting the conference activities and putting together this book;
- Makerere University Management and the University of Groningen for their overwhelming support;
- The Netherlands Organisation for International Cooperation in Higher Education for providing financial support for the Conference and publication of this book.
- Lastly but not least our publishers, Fountain Publishers, for a job well done.

Editors

Joseph Kizza
Jackson Muhirwe
Janet Aibett
Katherine Getao
Victor Mbarika
Dilip Patel
Anthony Rodrigues

Introduction

Joseph M. Kizza

The computer revolution, which many take to be more profound than its two preceding cousins, the agricultural and industrial revolutions, is touching every aspect of humanity irrespective of where one lives. In the last fifty years of this revolution, there have been predictions, although many have passed without incident, of the transformation of human societies. There is evidence that some of this transformation is actually taking place. In many countries, more so in some than in others, this transformation is life changing. There is a lot for everyone to learn, something that would probably be impossible without the help of computers. Africa, although a late comer to the sharing table, is and must share in the benefits of this revolution in order to speed up the acquisition of capacity needed to gainfully exploit its natural resources that have, for generations, been taken at will by non-Africans.

This generation of the needed capacity for development must be spearheaded by the African peoples themselves. The process is being helped by mastering the development and use of the tools of the information communication technology (ICT). ICT knowledge, like all new technologies, is expensive to acquire, especially in the African environment with rare and constrained resources. The quickest and easiest way to do this under the given environment is to create academic knowledge base sharing market places or bazaars. Conferences are the foremost of the knowledge sharing bazaars.

The Faculty of Computing and Information Technology (CIT) of Makerere University in Uganda annually holds one of these bazaars on the continent. This is the third in the series under the title" International Conference on Computing and ICT Research (SREC 07)". From each one of these annual conferences, a proceedings book of papers presented at the conference is produced. This year is no exception.

The papers in this book all focus on the theme of "Strengthening the Role of ICT in Development". The International Conference on Computing and ICT Research (SREC 07), out of which this book grew, was a platform and a forum where ICT-based development was discussed over a course of three days. Papers were presented in six tracks: Computer Science, Information Systems, Information Technology, Data Communications and Networks, Software Engineering, and Information Communication Technology for Sustainable Development.

There were six keynote papers to headline the conference. Joseph M. Kizza in "The Diminishing Private Network Security Perimeter Defense" discusses the changing defensive posture of the Enterprise networks in the face of the rapidly changing and miniaturizing technology creating a dangerous situation in

the security of computer networks. Anthony J. Rodrigues in "Does Interactive Learning Enhance Education: For Whom, In What Ways and In Which Contexts?" explores a balanced view of the body of research during the past three decades that has investigated interactive learning in a variety of forms ranging from the earliest days of mainframe-based computer-assisted instruction to contemporary multimedia learning environments accessible via the World Wide Web. He summarizes what is known and what is not known about interactive learning, and describes the strengths and weaknesses of various approaches to interactive learning research. Dilip Patel et al in "An Ontological Approach to Domain Modeling for MDA-oriented Processes" contend that there exists a gap between objects which exist in the real world and the elements which represent them in a software system and they use an ontological approach to address this gap with a domain modeling process able to construct, manage, and negotiate the appropriate domain concepts within a domain model. In "Efficient Splice Site Prediction with Context-Sensitive Distance Kernels" Bernard Manderick et al discuss four types of context-sensitive kernel functions: linear, polynomial, radial basis, and negative distance functions for doing splice site prediction with a support vector machine. P. Mahanti et al in "A Framework for Adaptive Educational Modeling: A Generic Process" look at the education system with an approach of facing challenges of globalization through e-education, and discuss the possibilities of using existing IT developments in the field of education so as to enable it to evolve new paradigms of developmental education. In their second paper, " Design Space Exploration of Network on Chip: A system Level Approach", Mahanti et al propose a model for the design of space exploration of a mesh based Network on Chip architecture at system level. They do this to find the topological mapping of intellectual properties (IPs) into a mesh-based Network on Chip(NoC), to minimize energy and maximum bandwidth requirement using a heuristic technique based on multi-objective genetic algorithm.

In Computer science, eight papers were presented. Lorenzo Dalvit et al in "The Deployment of an E-Commerce Platform and Related Projects in a Rural Area in South Africa" describe the development and deployment of an ecommerce platform in Dwesa, a rural area in the former homeland of Transkei in South Africa, designed to promote tourism and advertise local arts, crafts and music, and it entails a number of related projects. In "Properties of Preconditioners for Robust Linear Regression" V. Baryamureeba and T. Steihaug show that by combining an inexact Newton method and an iteratively re-weighted least squares method preconditioned conjugate gradient least square algorithm they can solve large, sparse systems of linear equations efficiently.

In a second paper, "A Methodology for Feature Selection in Named Entity Recognition", Baryamureeba and Frederick Kitoogo present a methodology for feature selection in named entity recognition. Unlike traditional named entity recognition approaches which mainly consider accuracy improvement as the sole objective, their approach uses a multi-objective genetic algorithm which

is employed for feature selection based on various aspects including error rate reduction and time taken for evaluation, and also demonstrating the use of Pareto optimization. In "Extraction of Interesting Association Rules Using Genetic Algorithms", Peter P. Wakabi-Waiswa and V. Baryamureeba present a Pareto – based multi – objective evolutionary algorithm rule mining method based on genetic algorithms using confidence, comprehensibility, interestingness, and surprise as objectives of the association rule mining problem. Daudi Jjingo and Wojciech Makalowski in "Computational Identification of Transposable Elements in the Mouse Genome" discuss the identification of Transposable elements (TEs) within the mouse transcriptome based on available sequence information on mouse cDNAs (complementary DNAs) from GenBank and how using various bioinformatics software tools as well as tailor made programming, they can establish: the absolute location coordinates of the TEs on the transcript, the location of the IRs with respect to the 5'UTR, CDS and 3'UTR sequence features, the quality of alignment of the TE's consensus sequence on the transcripts where they exist, the frequencies and distributions of the Tes on the cDNAs, and the descriptions of the types and roles of transcripts containing TEs. Jackson Muhirwe in "Towards Human Language Technologies for Under-resourced Languages" discusses and presents strategies for improving the human language technologies for under-resourced languages, which form the majority of world languages, yet in contrast to developed languages which enjoy highly developed language technologies, have not attracted much attention from researchers probably due to economical and political reasons. In his second paper "Computational Analysis of Kinyarwanda Morphology: The Morphological Alternations", Muhirwe discusses efforts being made in morphological alternations of Kinyarwanda, a Bantu language spoken in East Africa, one of those under-resourced languages still lacking two of the most essential components of most natural language applications: the morphological analyzer and a machine-readable lexicon. J.Poornaselvan1 et al in "Efficient IP Lookup Algorithm" present a new order O (w/k + k) efficient algorithm which uses the Longest Prefix Matching to speed up the IP address lookup. The algorithm provides a better search time over the existing Elevator-Stairs Algorithm, by accomplishing a two-way search in the trie.

In Information Systems, nine papers were presented. In "Geometrical Spatial Data Integration in Geo-Information Management", Ismail Wadembere and Patrick J. Ogao present a conceptual mechanism for geospatial data integration that can identify, compare, determine differences and adjust spatial geometries of objects for information management. In "A Visualation Framework for Discovering Prepaid Mobile Subscriber Usage Patterns", John Aogon and Patrick J. Ogao, present findings from a visualization framework they developed for discovering subscriber usage patterns. The framework is evaluated using call data with known knowledge obtained from the local telecommunication operator. Further, Ogao and Agnes Nakakawa in "A Spatial Decision Support Tool for Landfill Site Selection: A Case for Municipal Solid Waste Management" present a Spatial Decision Support Tool

that can be used to solve the complex process of landfill site selection for municipal solid waste management in Kampala and the neighboring Wakiso districts. In "Web Accessibility in Uganda: A study of Webmaster Perceptions", Rehema Baguma et al examine the practice and perceptions of webmasters in Uganda on the accessibility of government websites and what can be done to improve the situation. Geoffrey Ocen in "Organisational Implementation of ICT: Findings from NGOs in the United Kingdom and Lessons for Developing Countries" reports on website adoption process in 5 small and medium-sized voluntary sector organisations and identify Technology, Organisational and People (TOP) imperatives that provide new conceptual framework for facilitating organisational implementation of ICT and the relevancy of this framework to NGOs in developing countries. In "Enhancing Immunization Healthcare Delivery through the Use of Information Communication Technologies", Agnes Semwanga Rwashana and Ddembe Williams discuss the challenges in the current immunization system and show how information communication technologies can be used to enhance immunization coverage. They also present a framework to capture the complex and dynamic nature of the immunization process, to enhance the understanding of the immunization health care problems and to generate insights that may increase the immunization coverage effectiveness. Also in "Complexity and Risk in IS Projects- A System Dynamics Approach", Paul Ssemaluulu and Ddembe Williams discuss the research efforts where they are using system dynamics methodology to create a better understanding of the link between information quality and customer satisfaction. In "Towards a Reusable Evaluation Framework for Ontology based biomedical Systems Integration" Gilbert Maiga and Ddembe Williams propose a framework to extend existing Information systems and ontology evaluation approaches with the potential to relate ontology structure to user objectives in order to derive requirements for a flexible framework for evaluating ontology based integrated biological and clinical information systems in environments with changing user needs and increasing biomedical data. Finally in "Knowledge Management Technologies and Organizational Business Processes: Integration for Business Delivery Performance in Sub Saharan Africa", Asifiwe Rubanju presents a discussion using practical examples of knowledge management theories in order to understand business performance contending that knowledge management together with mindful application of the business technology can be of extreme importance for performance in organizations.

In Information Technology, four papers were presented. In " M-Learning: The Educational Use of Mobile Communication Devices", Paul Birevu Muyinda et al explore the use of mobile communication devices in teaching and learning focusing on the need to blend mobile with fixed communication devices in order to bridge the digital divide in electronic learning. In "Implementation of E-Leaning in Higher Education Institutions in Low bandwidth Environments: A Blended Learning Approach", Nazir Ahmad Suhail and Ezra K Mugisa review various factors and processes in e-learning with an emphasis on university settings and

after analyzing, synthesizing and making a comparative study of the frameworks and models, they propose a gradual transition model for implementation of e-learning in Higher Educations Institutions in least developed countries followed by a comprehensive framework adaptable in low bandwidth environment using a blended learning approach. Implementation process of the framework is also explained. In "Standards-based B2B e-commerce Adoption", Moses Niwe, presents preliminary findings from an explorative study of Industry Data Exchange Association (IDEA) concerning challenges, benefits and problems in adoption of standard based business-to-business e-commerce. Michael Niyitegeka in "Towards a website Evaluation Framework for Universities: A Case Study of Makerere University", proposes an evaluation framework that could be used by management at the faculties and the University.

In data Commmunication and Computer Networks, five papers were presented. Bulega Tommy et al in " Improving QoS with MIMO-OFDM for the Future Broadband Wireless", provide an analytical overview and performance analysis on the key issues of an emerging technology known as multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) wireless that offers significant promise in achieving high data rates over wireless links. In "Analysis of Free Haven Anonymous Storage and Publication System", Drake Mirembe and Francis Otto evaluate the design of a distributed anonymous storage and publication system that is meant to resist even attacks of the most powerful adversaries like the government. They also present a discussion of whether the assumptions held in the design of Free Haven System (FHS) are achievable in the real world and the potential implementation hurdles. In "A model for Data Management in Peer-to-peer System", Francis Otto et al propose a modularized data management model for a P2P system that cleanly separates the functional components, enabling the implementation of various P2P services with specific quality of service requirements using a common infrastructure. Tom Wanyama in "Subscriber Mobility Modeling in Wireless Networks" presents a simplified model for user mobility behavior for wireless networks. The model takes into account speed and direction of motion, the two major characteristics of user mobility. Julianne Sansa in "An Evaluation Study of Data Transport Protocols for e-VLBI", compares TCP-like data transport algorithms in the light of e-VLBI requirements and proposes HTCP with bandwidth estimation (HTCP-BE) as a suitable candidate by simulating its behaviour in comparison with seven existing TCP variants: HighSpeed TCP for large Congestion Window (HSTCP), Scalable TCP (STCP), Binary Increase TCP (BI-TCP), Cubic TCP (CUBIC-TCP), TCP for Highspeed and long-distance networks (HTCP), TCP Westwood+ (TCPW) and standard TCP (TCP).

In Software Engineering, six papers were presented. In „ A Comparison of Service Oriented Architecture with otherAdvances in Software Architectures", Benjamin Kanagwa and Ezra K Mugisa discuss Service Oriented Architecture (SOA) as they investigate and show the relationship between SOA and other

advances in software architecture. They also advance the view that SOA's uniqueness and strength do not lie in its computational elements but in the way it enables and handles their interaction. In "Towards Compositional Support for a Heterogeneous Repository of Software Components", Agnes F. N. Lumala and Ezra K. Mugisa present research work to define composition within an environment of heterogeneous software artifacts, and later propose a strategy to handle the research. Tom Wanyama and Agnes Lumala in "Decision Support for the Selection of COTS", describe an agent-based Decision Support System (DSS), which integrates various COTS selection DSA and repositories to address a variety of COTS selection challenges. Besides managing the COTS selection DSA and repositories, the agents are used to support communication and information sharing among the COTS selection stakeholders. Miriam Sensalire and Patrick Ogao in "Not all Visualizations are Useful: The Need to Target User Needs when Visualizing Object Oriented Software" present and discuss results from observing expert programmers use 3 visualization tools. The results show that if tools are developed without user consultation, they may fail to be useful for the users. In "A User-Centered Approach for Testing Spreadsheets", Yirsaw Ayalew presents an approach for checking spreadsheets on the premises that their developers are not software professionals.

In Information Communication Technology for Sustainable Development, there were seven papers presented. In "ICT as an Engine for Uganda's Economic Growth: The Role of and Opportunities for Makerere University", Venansius Baryamureeba, discusses the role of Makerere University and suggests opportunities for Makerere University in ICT led –economic growth of Uganda. In "Conceptual ICT tool for Sustainable Development: The Community Development Index (CDI)", Joseph Muliaro Wafula et al present the CDI concept that can be used to develop a decision support tool for policy makers, leaders and their people. In "Computational Resource Optimization in Ugandan Tertiary Institutions", Richard Ssekibuule et al present ways (limited) computer resources can be optimally used in a financially constrained setting, as well as ways for building an environment for providing high computing power for learning and research in a cost effective way. Richard Sssekibuule in "Security Analysis of Remote E-Voting", analyzes security considerations for a remote Internet voting system based on the system architecture of remote Internet voting. Tom Wanyama and Venansius Baryamureeba in "The Role of Academia in Fostering Private Sector Competitiveness in ICT Development" present the prerequisite for producing computing graduates who have the skills required to foster private sector competitiveness in information and communications technology development. They further discuss the steps the Faculty of Computing and Information Technology at Makerere University has taken to ensure that graduates are of high quality and have the computing skills needed by the private sector and other potential employers. In „E-government for Uganda: Challenges and Opportunities", Narcis T. Rwangoga and Asiime Patience Baryayetunga look at Uganda government ICT flagship programmes and

discuss underlying challenges faced by these ICT initiatives in Uganda making recommendations for planners when designing future ICT initiatives.

Part 1

Computer Science

1
Efficient Splice Site Prediction with Context-Sensitive Distance Kernels

Bernard Manderick, Feng Liu and Bram Vanschoenwinkel

This paper presents a comparison between different context-sensitive kernel functions for doing splice site prediction with a support vector machine. Four types of kernel functions will be used: linear-, polynomial-, radial basis function- and negative distance-based kernels. Domain-knowledge can be incorporated into the kernels by incorporating statistical measures or by directly plugging in distance functions defined on the splice site instances. From the experimental results it becomes clear that the radial basis function-based kernels get the best accuracies. However, because classification speed is of crucial importance to the splice site prediction system, this kernel is computationally too expensive. Nevertheless, in general incorporating domain knowledge does not only improve classification accuracy, but also reduces model complexity which in its turn again increases classification speed.

1. Introduction

An important task in bio-informatics is the analysis of genome sequences for the location and structure of their genes, often referred to as gene finding. In general, a gene can be defined as a region in a DNA sequence that is used in the production of a specific protein. In many genes, the DNA sequence coding for proteins, called exons, may be interrupted by stretches of non-coding DNA, called introns. A gene starts with an exon, is then interrupted by an intron, followed by another exon, intron and so on, until it ends in an exon. Splicing is the process by which the introns are subtracted from the exons.

Hence we can make a distinction between two different splice sites: i) the exon-intron boundary, referred to as the donor site and ii) the intron-exon boundary, referred to as the acceptor site. Splice site prediction, an important subtask in gene finding systems, is the automatic identification of those regions in the DNA sequence that are either donor sites or acceptor sites [?].

Because splice site prediction instances can be represented by a context of a number of nucleotides before and after the candidate splice site, it is called a context-dependent classification task. In this paper we do splice prediction with support vector machines (SVMs) using kernel functions that take into account the information available at different positions in the contexts. In this sense the kernel functions are called contextsensitive. This is explained in Sections ?? and ??.

More precisely, in a support vector machine, the data is first mapped non-linearly from the original input space X to a high-dimensional Hilbert space called the feature space F and then separated by a maximum-margin hyperplane, i.e. linearly, in that space F. By making use of the *kernel trick*, the mapping $\varnothing : X \rightarrow F$ remains implicit, and as a result we avoid working in the high-dimensional feature space F. Moreover, because the mapping is non-linear, the decision boundary which is linear in F corresponds to a non-linear decision boundary in the input space X. One of the most important design decisions in SVM learning is the choice of kernel function $K : X \times X \rightarrow \mathbb{R}$ because the maximum-margin hyperplane is defined completely by inner products of vectors in the Hilbert feature space F using the kernel function K. Since K takes elements x and y from the input space X and calculates the inner products of $\varnothing$ (x) and $\varnothing$(y) in the Hilbert feature space F without having to represent or even to know the exact form of the elements $\varnothing$(x) and $\varnothing$(y). As a consequence the mapping $\varnothing$ remains implicit and we have a computational benefit [?]. In the light of the above it is not hard to see that computational efficiency of K is crucial for the success of the classification process. We refer to Section ?? for more on theoretical background concerning the SVM.

As a result, the learning process can benefit a lot from the use of special purpose similarity or dissimilarity measures in the calculation of K [?,?,?,?]. However, incorporating such knowledge in a kernel function is non-trivial since a kernel function K has to satisfy a number of properties that result directly from the definition of an inner product. In this paper we will consider two types of kernel functions that can make direct use of distance functions defined on contexts themselves: i) *negative distance kernels* and ii) *radial basis function kernels*. This is explained in Section ??.

Furthermore, because classification speed is of crucial importance to a splice site prediction system the used kernel functions should be computationally very efficient. For that reason, in related work on splice site prediction with a SVM a linear kernel is chosen in favor of computational efficiency but at the cost of some accuracy [?]. In this light most of the kernels presented here will probably be too expensive, therefore we also show results for context-sensitive linear kernels and from these results it can be seen that the classification speed can be further increased while at the same time precision and accuracy of the predictions are a little higher. This is discussed in Section ??.

2. Context-Dependent Classification

In this paper we consider classification tasks where it is the purpose to classify a focus symbol in a sequence of symbols, based on a number of symbols before and after the focus symbol. The focus symbol, together with the symbols before and after it, is called a *context* and applications that rely on such contexts will be called context-dependent. Splice site prediction is a typical example of a *context-dependent* classification task. Here, each symbol is one of the four nucleotides {A,C,G, T}.

2.1 Contexts

We start with a definition of a context followed by an illustration in the framework of splice site prediction.

Definition 1. *A context* s_p^{-q} *is a sequence of symbols si ∈ D with p symbols before and q symbols after a focus symbol* s_p *at position p as follows*

$$s_p^{-q} = (s_0, \ldots, s_p, \ldots s_{p+q} \quad (1)$$

with (p + q) + 1 the length of the context, with D the dictionary of all symbols, with D| = m and with p the left context size and with q the right context size.

Example 1. Remind from the introduction that in splice site prediction it is the purpose to automatically identify those regions in a DNA sequence to be donor sites or acceptor sites. Essentially, DNA is a sequence of nucleotides represented by a four character alphabet or dictionary D = {A,C,G, T}. Moreover, an acceptor site always contains the AG dinucleotide and a donor site always contains the GT dinucleotide. In this light splice site prediction instances can be represented by a context of a number of nucleotides before and after the AG/GT dinucleotides. More precisely, given a fragment of a DNA sequence, . . . CCATTGGTGGCAGCCAG . . . the candidate donor site given by the dinucleotide GT can be represented by a context in terms of Definition ?? as

$$s_p^{-q} = \left(\underbrace{A\ T, T\ G,}_{S_0 \ldots\ldots, S_{P-1}} \underbrace{G\ T}_{S_P} , \underbrace{G\ G\ C}_{S_{P+1} \ldots\ldots, S_{P+Q}} \right)$$

with $p = 4$ the left context size and $q = 3$ the right context size and with $(p + q) + 1 = 8$ the total length of the context.

Furthermore, for splice site prediction there is no need to represent the AG/GT dinucleotides, because two separate classifiers are trained, one for donor sites and one for acceptor sites. In this light the only possible symbols occurring in the contexts are given by the dictionary D.

Note that, for reasons of computational efficiency, in practice the symbols in the contexts will be represented by an integer, more precisely by assigning all the symbols that occur in the training set a unique index and subsequently using that index in the context instead of the symbols themselves.

2.2 The Overlap Metric

The most basic distance function defined on contexts is called the overlap metric, it simply counts the number of mismatching symbols at corresponding positions in two contexts.

Definition 2. *Let* $\mathbb{S}^n$ *be a set with contexts* $\bar{s}_p^q$ *and* $\bar{t}_p^q$ *with* $n = (p + q) + 1$ *the length of the contexts, with symbols si, ti 2 D the dictionary of all distinct symbols with* $|D| = m$ *and let* $w \in \mathbb{R}^n$ *be a context weight vector. Then the* overlap metric $d_{OM} : \mathbb{S}^n \times \mathbb{S}^n \rightarrow \mathbb{R}^+$ is defined as

$$\mathbb{S}^n \times \mathbb{S}^n \rightarrow \mathbb{R}^+ : d_{OM}\ (\bar{s}, \bar{t}) = \sum_{i=0}^{n-1} w_{i\_}\, \delta(s_i, t_i)$$

with $\delta : \mathbb{S} \times \mathbb{S} = \mathbb{R}^+$ *defined as*

$$\delta\,(s_i, t_i) = \begin{cases} w_i & \text{if } s_i\ t_i \\ 0 & \text{else} \end{cases} \tag{3}$$

with $w_i \geq 0$ *a context weight for the symbol at position i.*

Next, we make a distinction between two cases: i) if all $w_i = 1$ no weighting takes place and the metric is referred to as the *simple overlap metric* d_{SOM} and ii) otherwise a position dependent weighting does take place and the metric is referred to as the weighted overlap metric d_{WOM}. A question that now naturally rises is: what measures can be used to weigh the different context positions?

Information theory provides many useful tools for measuring statistics in the way described above. In this work we made use of three measures known as i) information gain [?], ii) gain ratio [?] and iii) shared variance [?]. For more details the reader is referred to the related literature.

2.3 The Modified Value Difference Metric

The *Modified Value Difference Metric* (MVDM) [?] is a powerful method for measuring the distance between sequences of symbols like the contexts considered here. The MVDM is based on the *Stanfill-Waltz Value Difference Metric* introduced in 1986 [?]. The MVDM determines the similarity of all the possible symbols at a particular context position by looking at co-occurrence of the symbols with the target class. Consider the following definition.

Definition 3. *Let* $\mathbb{S}^n$ *be a set with contexts* $\bar{s}_p^q$ *and* $\bar{t}_p^q$ *with* $n = (p + q) + 1$ *the length of the contexts as before, with components* s_i and $t_i \in D$ *the dictionary of all distinct symbols with* $|D| = m$. *Then the modified value difference metric* $d_{MVDM} : \mathbb{S}^n \times \mathbb{S}^n \rightarrow \mathbb{R}^+$ *is defined as*

$$\mathbb{S}^n \times \mathbb{S}^n \rightarrow \mathbb{R}^+ : d_{MVDM}\,(\bar{s}, \bar{t}) \sum_{i=0}^{n-1} \delta\,(s_{i,}\, t_i)^r \tag{4}$$

with r a constant often equal to 1 or 2 and with $: D \times D \rightarrow \mathbb{R}^+$ *the difference of the conditional distribution of the classes as follows:*

$$\delta(\mathbf{s}_i,\mathbf{t}_i)^r = \sum_{j=1}^{M} |p(y_j \backslash \mathbf{s}_i) - p(y_j \backslash \mathbf{t}_i)|^r \quad (5)$$

with y_j the class labels and with M the number of classes in the classification problem under consideration.

3. Context-Sensitive Kernel Functions

In the following section we will introduce a number of kernel functions that make direct use of the distance functions d_{SOM}, d_{WOM} and d_{MVDM} defined in the previous section. In the case of d_{WOM} and d_{MVDM} the kernels are called context-sensitive as they take into account the amount of information that is present at different context positions as discussed above.

3.1 Theoretical Background

Remind that in the SVM framework classification is done by considering a kernel induced feature mapping Φ that maps the data from the input space X to a high dimensional Hilbert space F and classification is done by means of a maximum-margin hyperplane in that space F. This is done by making use of a special function called a *kernel*.

Definition 4. A kernel $K : X \times X \to \mathbb{R}$ *is a symmetric function so that for all* $\mathbf{x}$ *and* $\mathbf{x}'$ *in X,* $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ *where* ϕ *is a (non-linear) mapping from the input space X into the Hilbert space F provided with the inner product* $\langle ., . \rangle$.

However, not all symmetric functions over X × X are kernels that can be used in a SVM, because a kernel function needs to satisfy a number of conditions imposed by the fact that it calculates an inner product in F. More precisely, in the SVM framework we distinguish two classes of kernel functions: i) positive semi-definite kernels (PSD) and ii) conditionally positive definite (CPD) kernels.

Whereas a PSD kernel can be considered as one of the most simple generalizations of one of the simplest similarity measures, i.e. the inner product, CPD kernels can be considered as generalizations of the simplest dissimilarity measure, i.e. the distance $\| \mathbf{x} - \mathbf{x}' \|$ [?,?,?].

One type of CPD kernel that is of particular interest to us is given in [?] from which we quote the following two theorems.

Theorem 1. *Let X be the input space, then the function* $K : X \times X \to \mathbb{R}$ *:*

$$K_{nd}(\mathbf{x} - \mathbf{x}') = \| \mathbf{x} - \mathbf{x}' \|^{\beta} \text{ with } 0 < \beta \leq 2 \quad (6)$$

is CPD. The kernel K defined in this way is referred to as the negative distance kernel.

Another result that is of particular interest to us relates a CPD K to a PSD kernel $\tilde{K}$ by plugging in K into the exponent of the standard radial basis function kernel, this is expressed in the following theorem [?]:

Theorem 2. *Let X be the input space and let $K : X \times X \to \mathbb{R}$ be a kernel, then K is CPD if and only if*

$$K_{rbf}(\mathbf{x} - \mathbf{x}') = exp(\gamma K(\mathbf{x} - \mathbf{x}') \quad (7)$$

is PSD for all $\gamma > 0$. The kernel K_{rbf} defined in this way is referred to as the radial basis function kernel.

For Theorem ?? to work, it is assumed that $X \subseteq \to \mathbb{R}^n$ where $\mathbb{R}^n$ is a normed vector space. But, for contexts in particular and sequences of symbols in general one can not define a norm like in the RHS of Equation ??. More precisely, given the results above, if we want to use an arbitrary distance dX defined on the input space X in a kernel K, we should be able to express it as $d_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$ from which it then automatically follows that $-d_x$ is CPD by application of Theorem ??.

In our case however, since the input space $X \subseteq \mathbb{S}^n$ the set of all contexts of length n, the distances d_{SOM}, d_{WOM} and d_{MVDM} we would like to use can therefore not be expressed in terms of Theorem ??. Nevertheless, in previous work it has been shown that $-d_{SOM}$, $-d_{WOM}$ and $-d_{MVDM}$ are CPD [?,?,?,?], this will be briefly explained next. For more details the reader is referred to the literature.

More precisely, for the overlap metric defined on the contexts it can be shown that it corresponds to an orthonormal vector encoding of those contexts [?,?,?]. In the orthonormal vector encoding every symbol in the dictionary D is represented by a unique unit vector and complete contexts are formed by concatenating these unit vectors. Notice that this is actually the standard approach to context-dependent classification with SVMs [?,?] and in this light the non-sensitive linear, polynomial, radial basis function and negative distance kernels employing the simple overlap metric (i.e. the unweighted case) presented next, are actually equivalent to the standard linear, polynomial, radial basis function and negative distance kernel applied to the orthonormal vector encoding of the contexts.

Finally, for the MVDM with $r = 2$ it can be shown that it corresponds to the Euclidean distance in a transformed space, based on a probabilistic reformulation of the MVDM presented in [?,?].

3.2 A Weighted Polynomial Kernel

The first kernel we will define here is based on Equation ?? of the definition of the overlap metric from Definition ??. In the same way as before, we make a distinction between the unweighted non-sensitive case and the weighted context-sensitive case, for more details the reader is referred to [?,?,?].

Definition 5. *Let $X \subseteq \mathbb{S}^n$ be the input space with contexts $\overset{-q}{s}_p$ and $\overset{-q}{t}_p$ with $n = (p+q)+1$ the length of the contexts and s_i, $t_i \in D$ the symbols at position i in the contexts as before, and let $w \in \mathbb{R}^n$ be a context weight vector, then we define the simple overlap kernel $KSOK : X \times X \to \mathbb{R}$ as*

$$K_{SOK}\left(\bar{s},\bar{t}\right)=\left(\sum_{i=0}^{n-1}\delta(s_i,t_i)+c\right)^d \tag{8}$$

with $c \geq 0$, $d > 0$ *and* $w = 1$, *the weighted overlap kernel* $K_{WOK} : X \times X \to \mathbb{R}$ is defined in the same way but with a context weight vector $w \neq 1$.

3.3 Negative Distance Kernels

Next, we give the definitions of three negative distance kernels employing the distances d_{SOM}, d_{WOM} and d_{MVDM}, for more details we refer to [?,?,?].

We start with the definition of two negative distance kernels using the overlap metric from Definition ??. In the same way as before, we make a distinction between the unweighted, non-sensitive case d_{SOM} and the weighted, context-sensitive case d_{WOM}.

Definition 6. *Let* $X \subseteq \mathbb{S}^n$ *be the input space with contexts* $\bar{s}_p^q$ *and* $\bar{t}_p^q$ *with* $n = (p+q)+1$ *the length of the contexts and* s_i, $t_i \in D$ *the symbols at position* i *in the contexts as before, and let* $w \in \mathbb{R}^n$ *be a context weight vector, then we define the negative overlap distance kernel* $K_{NODK} : X \times X \to \mathbb{R}$ *as*

$$K_{NODK}\left(\bar{s},\bar{t}\right) = -d_{SOM}(\bar{s},\bar{t})^{\frac{1}{2}\beta} \tag{9}$$

with $0 < \beta \leq 2$ *and* $w = 1$ *as before, the negative weighted distance kernel* $K_{NWDK} : X \times X \to \mathbb{R}$ *is defined in the same way but substituting* d_{WOM} *for* d_{SOM} *in the RHS of Equation ??, i.e. with a context weight vector* $w \neq 1$.

Similarly, for the MVDM from Definition ?? we can define a negative distance type kernel as follows.

Definition 7. *Let* $X \subseteq \mathbb{S}^n$ *be the input space with contexts* $\bar{s}_p^q$ *and* $\bar{t}_p^q$ *with* $n = (p+q)+1$ *the length of the contexts and* s_i, $t_i \in D$ *the symbols at position* i *in the contexts as before, then we define the negative modified distance kernel* $K_{NMDK} : X \times X \to \mathbb{R}$ *as*

$$K_{NMDK}\left(\bar{s},\bar{t}\right) = -d_{MVDM}(\bar{s},\bar{t})^{\frac{1}{2}\beta} \tag{10}$$

with $0 < \beta \leq 2$*as before.*

However, it should be noted that for r = 1 in the definition of the MVDM $-d_{MVDM}$ is not CPD and thus for r = 1 the kernel K_{NMDK} will also not be CPD. Nevertheless, given the good empirical results we will use K_{NMDK} with d_{MVDM} and r = 1 anyway.

3.4 Radial Basis Function Kernels

Next, we give the definitions of three radial basis function kernels employing the distances d_{SOM}, d_{WOM} and d_{MVDM}, for more details we refer to [?,?,?].

We start with the definition of two radial basis function kernels employing the overlap metric from Definition ??. In the same way as before, we make a distinction between the unweighted non-sensitive case d_{SOM} and the weighted context-sensitive case d_{WOM}.

Definition 8. *Let $X \subseteq \mathbb{S}^n$ be the input space with contexts $\vec{s}_p^{\,q}$ and $\vec{t}_p^{\,q}$ with $n = (p+q)+1$ the length of the contexts and $s_i, t_i \in D$ the symbols at position i in the contexts as before, and let $\boldsymbol{w} \in \mathbb{R}^n$ be a context weight vector, then we define the overlap radial basis function kernel $K_{ORBF} : X \times X \to \mathbb{R}$ as*

$$K_{ORBF}\left(\vec{s}, \vec{t}\right) = exp\left(-\gamma\, d_{SOM}\left(\vec{s}, \vec{t}\right)\right) \tag{11}$$

with $\gamma > 0$ as before, with $\boldsymbol{w} = \mathbf{1}$ and the weighted radial basis function kernel $K_{WRBF} : X \times X \to \mathbb{R}$ is defined in the same way but substituting d_{WOM} for d_{SOM} in the RHS of Equation ??, i.e. with a context weight vector $w \neq 1$.

Similarly, for the MVDM from Definition ?? we can define a radial basis function type kernel as follows.

Definition 9. *Let $X \subseteq \mathbb{S}^n$ be the input space with contexts $\vec{s}_p^{\,q}$ and $\vec{t}_p^{\,q}$ with $n = (p+q)+1$ the length of the contexts and $s_i, t_i \in D$ the symbols at position i in the contexts as before, then we define the modified radial basis function kernel $K_{MRBF} : X \times X \to \mathbb{R}$ as*

$$K_{MRBF}\left(\vec{s}, \vec{t}\right) = exp\left(-\gamma\, d_{MVDM}\left(\vec{s}, \vec{t}\right)\right) \tag{12}$$

with $\gamma > 0$ as before.

It should however be noted that, with respect to the discussion above, i.e. that for $r = 1$ the distance $-d_{MVDM}$ and corresponding kernel K_{NMDK} are not CPD and therefore here for $r = 1$ the kernel K_{MRBF} will not be PSD. Nevertheless, given the good empirical results we used it anyway.

4. Experiments

We have done a number of experiments, first of all we wanted to validate the feasibility of our approach and compare our kernel functions that operate on contexts directly and see whether they are doing at least as well and hopefully better than the traditional kernels. In these experiments the left and right context length was set to 50. Second, we set up some experiments to find the optimal left and right context length for each classifier. Third, we also looked at di- and trinucleotides to find out whether this gave better performance than the single nucleotide case.

In the next sections, we describe the software and the data sets used in our experiments, we discuss how we have set the different parameters for the SVM, we present and discuss the results obtained and finally we give an overview of related work.

4.1. Software and Data

We did the experiments with LIBSVM [?], a Java/C++ library for SVM learning. The dataset we use in the experiments is a set of human genes, which is referred to as HumGS [?]. Each instance is represented by a fixed context size of 50 nucleotides before and 50 nucleotides after the candidate splice site based on the initial design strategy in [?]. Since, we train one classifier to predict donor sites and another classifier to predict acceptor sites, separate training and test sets are constructed for donor and acceptor sites. For the purpose of training the classifiers, we constructed balanced training sets.

For testing however we want a reflection of the real situation and keep the same ratio as given in the original set HumGS. This is shown in Table ??.

4.2. Parameter Selection and Accuracy

Parameter selection is done by 5-fold cross validation on the training set. For the ORBF, WRBF and MRBF, there are two free parameters that need to be optimized: the SVM cost parameterC (which is a trade-off for the model complexity and the model accuracy) and the radial basis function parameter γ.

Table 1. Overview of the data sets that have been used for the splice site prediction experiments.

data set	genes	GT+	GT−	AG+	AG−
HumGS	1115	5733	484714	5733	655822
training	/	4586	4586	4586	4586
testing	/	1147	96943	1147	131165

We performed a fine grid search for values of C and between 2^{-16} and 2^{5}. For the NODK, NWDK and NMDK only the cost parameter C has to be optimized because we choose fixed to 1 as this gives very good results, more precisely for β

= 2 results are not good at all, other values have not been tried. Again, values for C between 2^{-16} and 2_5 have been considered.

The LINEAR kernel in the results below refers to the the overlap kernel from Definition ?? with $d = 1$ and $c = 0$. For the SOK and the WOK we take $d = 2$ and $c = 0$ as previous work pointed out that higher values for d actually leads to bad results, while taking values for $c > 0$ does not have a significant impact on the results.

As a weighting scheme for the weighted kernels, we used three different weights: Information Gain (IG), Gain Ratio (GR) and Shared Variance (SV). For more details the reader is referred to [?].

Splice site prediction systems are often evaluated by means of the percentage of FP classifications at a particular recall rate. This measure is referred to as FP% [?] and is calculated as follows:

$$\text{FP\%} = \frac{\#\text{false positives}}{\#\text{ false positives} + \#\text{ true negatives}} \text{ x } 100$$

We used this evaluation measure for a recall rate of 95%, in this case the measure is referred to as FP95%, i.e. the FP95% measure gives the percentage of the predictions falsely classified as actual splice site at a level where the system has found 95% of all actual splice sites in the test set. Note that it is the purpose to have FP95% as low as possible.

4.3. Results

Table ?? gives an overview of the final FP95% results and model complexity in terms of the number of support vectors of the different kernels on the splice site prediction task. Note that the confidence intervals have been obtained by bootstrap resampling, at a confidence level = 0.05 [?]. A FP95% rate outside of these intervals is assumed to be significantly different from the related FP95% rate at a confidence level of = 0.05. In addition to the final FP95% results we also give as an illustration two FP% plots, for donor sites, comparing the context-sensitive kernels with those kernels that are not context-sensitive. Figure ?? does this for the negative distance kernel making use of the MVDM and Figure ?? does this for the radial basis function kernel making use of the WOK with GR, IG and SV weights.

Table 2. Splice site prediction, results for all kernels, for donor sites and for acceptor sites.

		donor sites	acceptor sites	
Kernel and Weights	**FP95%**	**#S Vs**	**FP95%**	**#S Vs**
LINEAR	8.18 ± 0.92	2398	12.78 ± 1.45	3621
LINEAR/GR	7.86 ± 1.05	1902	11.50 ± 1.55	2499
LINEAR/IG	7.92 ± 1.08	1905	11.78 ± 1.82	2494
LINEAR/SV	7.88 ± 1.02	2072	11.60 ± 2.01	2500
SOK	7.19 ± 0.70	3414	10.00 ± 1.22	3635
WOK/GR	6.51 ± 0.72	2126	9.06 ± 1.17	2698
WOK/IG	6.38 ± 0.80	2151	9.04 ± 1.11	2647
WOK/SV	6.43 ± 0.67	2156	9.07 ± 1.23	2695
NODK	7.97 ± 1.02	3372	11.36 ± 1.44	3696
NWDK/GR	6.43 ± 0.68	2803	9.71 ± 1.52	3143
NWDK/IG	6.40 ± 0.66	3009	9.66 ± 1.57	3380
NWDK/SV	6.38 ± 0.70	3169	9.76 ± 1.55	3252
NMDK (r = 1)	6.26 ± 0.75	2995	10.70 ± 1.46	2902
NMDK (r = 2)	6.38 ± 0.59	2625	12.63 ± 1.46	3146
ORBF	7.46 ± 0.77	4327	10.50 ± 1.66	4927
WRBF/GR	6.25 ± 0.72	2346	8.60 ± 1.65	2881
WRBF/IG	6.21 ± 0.57	2348	8.49 ± 1.73	2906
WRBF/SV	6.27 ± 0.75	2440	9.06 ± 1.43	2703
MRBF (r = 1)	5.81 ± 0.68	2696	10.36 ± 1.28	3652
MRBF (r = 2)	6.40 ± 0.78	2364	12.19 ± 1.67	2836

From the results it can be easily seen that in all cases the context-sensitive kernels making use of the WOM with IG, GR and SV weights and the MVDM always outperform their simple non-sensitive counterparts both in accuracy and in model complexity. Moreover in almost all cases this happens with a significant difference. There is however one exception, i.e. the MRBF with $r = 2$ for acceptor sites performs worse than its non-sensitive counterpart. Another overall observation is that the difference in the results between different context weights is not significant at all. Finally, it can be seen that the best result for donor sites is obtained by the MRBF with $r = 1$, but for acceptor sites this is the WRBF with IG weights. Therefore it is clear that the success of the used metric and the used weights depends, for a great deal, on the properties of the data under consideration so that it is worthwhile trying different metrics and different context weights to see which one gives the best result.

Finally, if one would like to use the LINEAR kernel in favour of classification speed but at the cost of some accuracy, it can be seen from the results that the weighted LINEAR kernel outperforms its unweighted counterpart, although the difference is not significant at a confidence level = 0.05. Nevertheless, it can be seen that the number of support vectors is significantly lower than for the unweighted LINEAR kernel and this will result in faster classification, because classification of a new instance happens by comparing it with every support vector in the model through the kernel function K.

Next, we look at the experiments to find the optimal left and right context length for each classifier. Then, we look at di- and trinucleotides to find out whether this gave better performance than the single nucleotide case. For these experiments we used the WRBF/GR kernel, this choice was based on the fact that WRBF performs second best for donor sites and best for acceptor sites. Moreover, since IG(information gain), GR(gain ratio) and SV(shared variance) were not significantly different in our experiments we used GR (gain ratio) as the weighting scheme. This follows from the experiments described above. The results are shown in Table ??.

Table 3. Splice site prediction, for the WRBF/GR kernel, for donor sites and for acceptor sites.

	donor sites	acceptor sites
nr. nucleotides	FP95% left context right context	FP95% left context right context
Single nucleotide	6.54 ± 0.92 60 40	8.79 ± 1.45 40 100
Dinucleotide	5.52 ± 0.70 60 40	7.46 ± 1.22 80 100
Trinucleotide	5.87 ± 1.02 60 60	8.46 ± 1.44 80 100

4.4. Related Work

The number of papers on splice site prediction and the related problem of gene finding is enormous and hence it is impossible to give an exhaustive overview. We will give some popular references (according to citeseer) and discuss some recent work.

The problem of recognizing signals in genomic sequences by computer analysis was pioneered by Staden [?] and the recognition of splice sites using neural networks was first addressed by Brunak et al. [?]. They trained a backpropagation feedforward neural network with one layer of hidden units to recognize donor and acceptor sites, respectively. The input consist of a sliding window centered on the nucleotide for which a prediction is to be made. The window is encoded as a numerical vector. The best results were obtained by combining a neural network to recognize the consensus signal at the splice site with another one that predicted coding regions based on the statistical properties of the codon usage and preference. This tool is available online at http://www.cbs.dtu.dk/services/NetGene2/.

Kulp et al. [?] and Reese et al. [?] build upon the work of Brunak by explicitely taking into account the correlations between neighboring nucleotides around a splice site by using dinucleotides as input features instead of single nucleotides. This tool is available online at http://www.fruitfly.org.

Genesplicer [?] uses a combination of a hidden Markov model and a decision tree. They obtained good performance compared to other leading splice site detectors at that time.

Rätsch and Sonnenburg [?] use a SVM with a special kernel to classify nucleotides as either donor or acceptor sites. There is one SVM for donor sites and one for acceptor sites. The system predicts the correct splice form for more than 92% of these genes. This approach is quite similar to ours but the kernel is different.

Finally, a list of online tools for splice site prediction and gene finding is available at http://www.cbs.dtu.dk/biolinks/pserve2.php.

5. Conclusions

In this article it was shown how different statistical measures and distance functions can be included into kernel functions for SVM learning in context-dependent classification tasks. The purpose of this approach is to make the kernels sensitive to the amount of information that is present in the contexts. More precisely, the case of splice site prediction has been discussed and from the experimental results it became clear that the sensitivity information has a positive effect on the results.

So far, this was shown on only one data set because the SVM is computationally very expensive but we have shown that kernel functions that operate on contexts directly gives additional benefits. At the moment, we are running experiments on a number of other data sets to show that the increased performance is not due to bias to the data sets. Apart from that, we are running experiments with more complex features based on the improved design strategy in [?], where a FP95% rate of 2.2% for donor and 2.9% for acceptor sites is obtained. In this light it remains to be seen whether the positive effect of the sensitivity information will still be significant in a system that already performs at very high precision without such information. Finally, we plan to compare our results with the ones obtained by other classifiers on the same data sets.

2
Design Space Exploration of Network on Chip: A system Level Approach

P.Mahanti* and Rabindra Ku Jana**

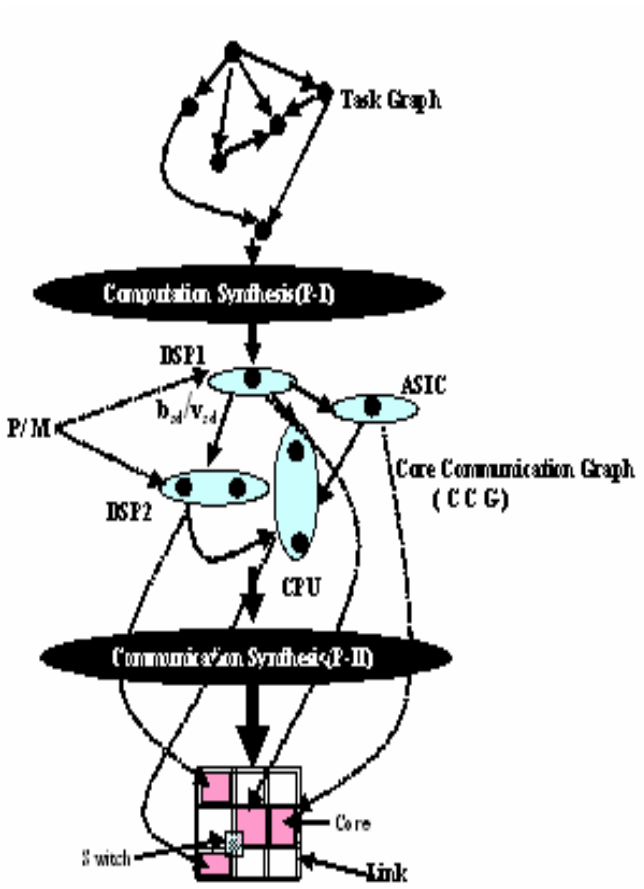
In this paper, we have proposed a model for design space exploration of a mesh based Network on Chip architecture at system level. The main aim of the paper is, to find the topological mapping of intellectual properties (IPs) into a mesh-based Network on Chip(NoC), to minimize energy and maximum bandwidth requirement. A heuristic technique based on multi-objective genetic algorithm is proposed to obtain an optimal approximation of the pareto-optimal front. We used "many-many" mapping between switch and cores (IPs) instead of "one-one" mapping. The experiments are performed on randomly generated benchmarks and a real application (a M-JPEG encoder) is shown to illustrate the efficiency, accuracy and scalability of the proposed model.

1. Introduction:

Network on Chip (NoC) has been proposed as a solution for the communication challenges like propagation delays, scalability, infeasibility of synchronous communication etc. in a nano scale regime [1-2]. To meet these challenges under the strong time-to-market pressure, it is essential to increase the reusability of components and system architectures in a plug and play fashion [3]. Simultaneously, the volume of data and control traffic among the cores grows. So, it is essential to address the communication-architecture synthesis problem through mapping of cores onto the communication architecture [4]. Therefore this paper focuses on communication architecture synthesis to minimize the energy consumption and communication delay by minimizing maximum link bandwidth.

The proposed communication synthesis task has solved in two phases as shown in figure-1. The first phase (P-1) is called computational synthesis. The input to P-I of is a task graph. The task graph consists up tasks as vertices and direct edges represent volume of data flowing between two vertices and their data dependencies.

Figure-1: Mappings for NoC synthesis problems



The output of P-I is a core communication graph (CCG) characterized by library of interconnection network elements and performance constraints. The core communication graph consists of processing and memory elements are shown by P/M in the figure-1. The directed edges between two blocks represent the communication trace. The communication trace characterized by bandwidth ($b_{sd)}$ and volume (v_{sd}) of data flowing between different cores. The second phase (P-II) is basically called as communication synthesis. The input to the P-II communication synthesis problem is the CCG. The output of the P-II is the energy and throughput synthesizes NoC back bone architecture as shown in Figure-1.

In this paper we address the problem of mapping the core onto NoC architecture to minimize energy consumption and maximum link bandwidth. Both of the above stated objectives are inversely proportional to each other. The stated problem is an NP-hard problem and genetic algorithm is a suitable candidate for solving the multi-objective problem [6]. The optimal solution obtained by our approach saves more than 15% of energy on average in compare to other existing approaches. Experimental result shows that our proposed model is superior in terms of quality of result and execution time in compare to other approaches.

The paper is organized as follows. We review the related work in Section 2. Section 3 and Section 4 describes the problem definition and the energy model assumed in this paper. Section 5 represents the multi-objective genetic algorithm formulation for the problem. Section 6 discusses the basic idea and problem formulation for the proposed approach. Experimental results are discussed in Section 7. Finally, a conclusion is given in Section 8.

2. Related Work

The problem of synthesis in mesh-based NoC architectures has been addressed by different authors. Hu and Marculescu [2] present a branch and bound algorithm for mapping IPs/cores on a mesh-based NoC architecture that minimizes the total amount of power consumed in communications. De Micheli [7] addresses the problem under the bandwidth constraint with the aim of minimizing communication delay by exploiting the possibility of splitting traffic among various paths. Lei and Kumar [8] present an approach that uses genetic algorithms to map an application on a mesh-based NoC architecture. The algorithm finds a mapping of the vertices of the task graph on the available cores so as to minimize the execution time. However these papers do not solve certain important issues. The first relates to the evaluation model used. In most of the approaches the exploration model decides the mapping to explore the design space without taking important dynamic effects of the system into consideration. Again in the above mentioned works, in fact, the application to be mapped is described using task graphs, as in [8], or simple variations such as the core graph in [7] or the application characterization graph (APCG) in [2]. These formalisms do not, however, capture important dynamics of communication traffic.

The second problem relates to the optimization method used. It refers in all cases to a single performance index (power in [2], performance in [7-8].So the optimization of one performance index may lead to unacceptable values for another performance index (e.g. high performance levels but unacceptable power consumption). Recently, jena and sharma [18] proposed a model that consider “many-many” mapping between core and tiles using multi-objective genetic algorithm. But they used core communication graph as the input to their model. We therefore think that the problem of mapping can be more usefully solved in a multi-objective environment starting from the higher level of input as compared to the model discussed in[18]. The contribution we intend to make in this paper is to propose a multi-objective approach to solving the synthesis problem on a mesh-based NoC architecture, where we take the specification graph as input. The approach will use evolutionary computing techniques based on genetic algorithm to explore the mapping space with the goal to optimize maximum link bandwidth and energy consumption (both computational and communication).

3. Problem Definition

3.1. A Task Graph (TG) is a digraph, G (V, E), where each vertex v ∈V represent task and each edge e ∈ E is a weighted edge, where weight signifies the volume of data flowing through the edge. Every edge also represents the data dependency between the connecting vertices.

3.2. A Core Communication Graph (CCG) is a digraph, G(V,E), where each vertex v ∈ V represent core and e ∈ E is a communication edge having two attributes, denoted by b_{sd} and v_{sd} are the required bandwidth and total volume of communication respectively.

3.3 Communication Structure

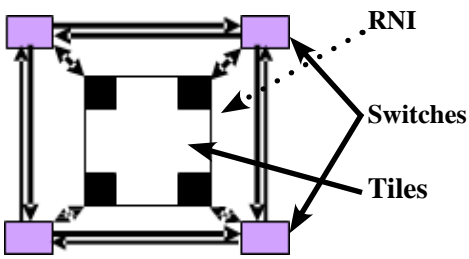
The 2-D mesh communication architecture has been considered for its several desire properties like regularity, concurrent data transmission and controlled electrical parameters[9][3]. Figure-2 shows a 2-D mesh NoC architecture. Each tile is surrounded by switches and links. Each switch is connected to its neighboring switches via two unidirectional links. Each IP in a tile can be connected to '4' neighboring switches as shown in Figure-2. To prevent the packet loss due to the multiple packets approaching to the same output port, each switch has small buffers (registers) to temporarily store the packets. Each resource has Resource Network Interfaces (RNIs) to connect to the network via switches. RNIs are responsible for packetizing and depacketizing the communication data. We implement static XY wormhole routing in this paper because:

i) it is easy to implement in switch.
ii) it doesn't required packet ordering buffer at the destination.
iii) it is free of deadlock and live lock [7][10].

4. Energy Model

Energy minimization is the one of the major challenging task in NoC design. In [11], Ye et al. first define the bit energy metric of a router as the energy consumed when a single bit of data goes through the router. In [2], Hu et al. modify the bit energy model so that it is suitable for 2D mesh NoC architecture.

Fig 2: Communication Structure



They derives mathematical expression for bit energy consume , when data transfer from switch 'i' to switch 'j' is given by

$$E_{i,j\,bit} = (h_{ij} + 1)\, E_{Sbit} + h_{ij}\, E_{Lbit} \text{ ---------- (1)}$$

Where E_{Sbit} and E_{Lbit} are the energy consumed in the switches and links respectively. The variable h_{ij} represent the number of links on the shortest path. As per the expression, the energy consumption is depend on the hop distance (h_{ij}) between switch 'i' and 'j' because E_{Sbit} and E_{Lbit} constant. Note E_{Sbit} is the energy consumption due to switches is depending on the number of ports in the switches. But in our case the total energy is the sum of communication and computation energy, i.e

$$E_{i,j\,bit} = (h_{ij} + 1)\, E_{Sbit} + h_{ij}\, E_{Lbit} + E_{Comp} \text{-------(2)}$$

E_{Comp} is the computational energy consumption. The following sections discuss the basic ideas of problem formulation using multi-objective optimization paradigm.

5. Multi-Objective Optimization

Definition: A general multi-objective optimization problem is defined as:

Minimize $f(x) = (f_1(x),\ldots.., f_k(x))$ subject to $x \in X$, where x represents a solution and X is a set of feasible solutions.

The objective function vector f (x) maps a solution vector 'x' in decision space to a point in objective space.

In general, in a multi-objective optimization problem, it is not possible to find a single solution that minimizes all objectives, simultaneously. Therefore, one is interested to explore a set of solutions, called the pareto optimal set, which are not dominated by any other solution in the feasible set. The corresponding objective vectors of these Pareto optimal points, named efficient points, form the Pareto front on the objective space.

Definition: We say x dominates x* iff $i \in \{1,\ldots..,k\}$
$f_i(x) \leq f_i(x^*)$ and there exists at least one $i \in \{1,\ldots., k\}$ such that $f_i(x) < f_i(x^*)$.

The most traditional approach to solve a multi-objective optimization problem is to aggregate the objectives into a single objective by using a weighting mean. However this approach has major drawbacks. It is not possible to locate the non-convex parts of the pareto front and it requires several consecutive runs of the optimization program with different weights. Recently, there has been an increasing interest in evolutionary multi-objective optimization. This is because of the fact that evolutionary algorithms (EAs) seem well-suited for this type of problems[12], as they deal simultaneously with a set of possible solutions called population. This allows us to find several members of the pareto optimal set in a single run of the algorithm. To solve the synthesis problem as discussed in Section 4, we used the multi-objective genetic algorithm.

5.1. A Multi-Objective Genetic Algorithm

In order to deal with the multi-objective nature of NoC problem we have developed genetic algorithms at different phases in our model. The algorithm starts with a set of randomly generated solutions (population). The population's size remains constant throughout the GA. Each iteration, solutions are selected, according to their fitness quality (ranking) to form new solutions (offspring). Offspring are generated through a reproduction process (Crossover, Mutation). In a multi-objective optimization, we are looking for all the solutions of best compromise, best solutions encountered over generations are fled into a secondary population called the "Pareto Archive". In the selection process, solutions can be selected from this "Pareto Archive"(elitism). A part of the offspring solutions replace their parents according to the replacement strategy. In our study, we used elitist non-dominated sorting genetic algorithm NSGA-II by Deb et al. [13].

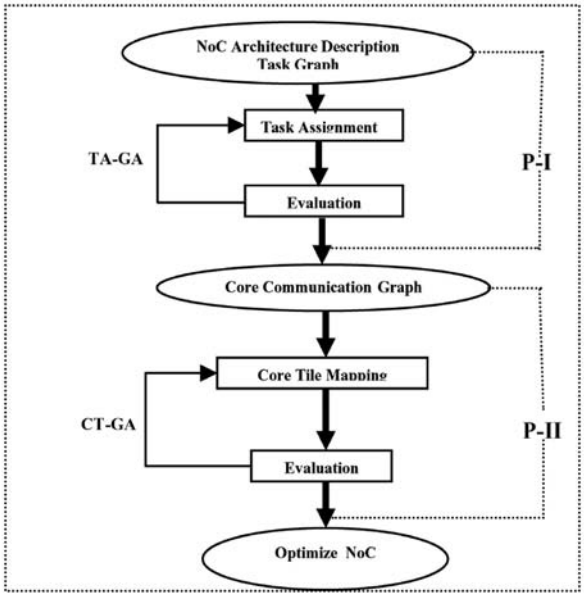
6. Problem Formulation

6.1 Basic Idea

Like other algorithms in area of design automation, the algorithm of NoC communication architecture is a hard problem. Our attempt is to develop an algorithm that can give near optimal solution within reasonable time. Genetic algorithms have shown the potential to achieve the dual goal quite well [8,14,15,18].

As shown in Figure-3 and discussed in Section-1, the problem is solved in two phases. The first phase (P-I) is basically a task assignment problem (TA-GA). The input to the problem is a TG. We assume that all the edge delays are a constant and equal to Average Edge Delay (AED) [10].The output of the first phase is a Core Communication Graph (CCG). The task of the second phase is Core-Tile Mapping using genetic algorithm (CT-GA). The next section discusses each phases in detail.

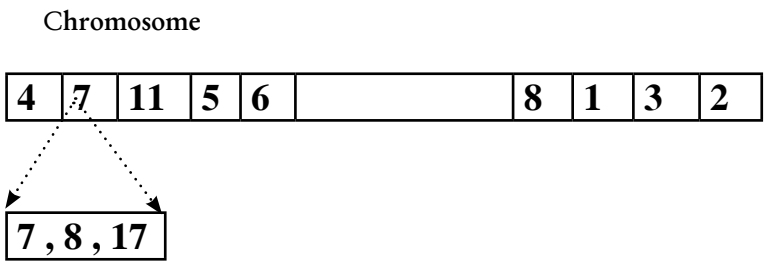
Fig 3: An overall design flow



6.1.1 Task Assignment Problem (TA-GA)

Given a task graph TG with all edge delay are constant and equal to average edge delay and IPs with specifications matrix containing cost and computational energy. The main objectives of this phase are to assign the tasks from the task graph to the available IPs in order to: (i) minimize the computational energy by reducing the power consumption. (ii) Minimize the total cost of the resources. The above said problem is a NP-hard multi-objective problem. We proposed a multi-objective genetic algorithm based on principle of (Non-Dominated Sorting Genetic Algorithm(NSGA)-II. Generally, in genetic algorithm, the chromosome is the representation of solution to the problem. In this case length of each chromosome is proportional to the number of nodes in a task graph. The i-th gene in the chromosome identifies the IP which is assigning the i-th node in the task graph. One example of chromosome encoding is given in Figure-5. Each gene (node in TG) in the chromosome contains an integer which represents an IP. Every IP chosen from the list of permissible IPs for that task. As shown in the Figure-4 the task number '2' in the task graph is assign to IP number '7' which is chosen from set of IPs {7, 8, and 17}. We consider a single point crossover to generate the offspring's. As for mutation operation, we consider the mutation by substitution i.e. at a time a gene in a chromosome is chosen with some random probability and the value in the gene is substitute by one of the best permissible value (i.e the index value of a IP) for the gene. The aim is to assign more tasks to a particular IP to reduce the communication between IPs i.e to minimize the number of IPs used for a task graph.

Figure-4: Chromosome encoding for task assignment.



6.1.2 Core-Tile Mapping (CT-GA)

After optimal assignment of tasks to the IPs, we get a Core Communication Graph (CCG) as shown in the figure-4. The input to this mapping task CT-GA is a CCG and a structure of NoC back bone. In our case it is a $n \times m$ mesh. The objectives of the mapping are (i) to reduce the average communication distance between the cores (i.e to reduce number of switches in the communication path). (ii) to maximize throughput(i.e minimize the maximum link bandwidth) under the communication constraint.

Core-tile mapping is a multi-objective mapping. So we use genetic algorithm based on NSGA-II. Here the chromosome is the representation of the solution to the problem, which in this case describe by the mapping. Each tile in mesh has an associated gene which identified the core mapped to the tile. In $n \times m$ mesh, for example the chromosome is formed by $n \times m$ genes. The i-th gene identified the core in the tiles in row ($\lceil (i / n) \rceil$) and column (i % n). The crossover and mutation operators for this mapping have been defined suitably as follows:

Crossover:

The crossover between two chromosomes C_1 and C_2 is generated a new chromosome C_3 as follows. The dominant mapping between C_1 and C_2 is chosen. Its hot core (the hot core is the IP required maximum communication) is remapped to a random tile in the mesh, resulting a new chromosome C_3.

Algorithm Crossover (C_1, C_2)

```
{
If (C1 dominate C2)
  C3 = C1;
   e l s e
  C3 = C2;
Swap (C3, Hot (C3), random ( {1,2,3,.....m×n}));
Return (C3);
}
```

The function Swap(C, i, j) exchange the i-th gene with j-th gene in the chromosome C.

Mutation:

The mutation operator act on a single chromosome (C) to obtained a muted chromosome C^0 as follows. A tile T_s from chromosome C is chosen at random. Indicating the core in the tile T_s as c_s and c_t as the core with which c_s communicates most frequently, c_s is remapped on a tile adjacent to T_s so as to reduce the distance between c_s and c_t . thus obtaining the mutated chromosome C^0. Algorithm, given below describes the mutation operator. The RandomTile(C) function gives a tile chosen at random from chromosome C. The MaxCommunication(c) function gives the core with which *c* communicates most frequently. The Row(C, *T*) and Col(C, *T*) functions respectively give the row and column of the tile *T* in chromosome C. Finally, the Uper, Lower, Left, Right(C, *T*) functions give the tile to the north, south, east and west of the tile *T* in chromosome C.

Mutate (C)

```
{
Chromosome C⁰ = C;
Tile T_s = Random Tile (C⁰);
Core c_s = C^0-1(T_s);
Core c_t = MaxCommunication (c_s);
Tile T_t = C⁰ (c_t);
i f ( Row(C⁰ , T_s ) < Row(C , T_t ) )
        T ⁰_s = Upper (C⁰ , T_s ) ;
else if ( Row(C⁰ , T_s ) > Row(C⁰ , T_t ) )
        T ⁰_s = Lower(C⁰ , T_s ) ;
          e l s e i f ( Col(C⁰ , T_s ) < Col(C⁰ , T_t ) )
        T ⁰_s = Left(C⁰ , T_s ) ;
else
   T ⁰_s = Right(C⁰ , T_s ) ;
      Swap (C⁰ , T_s , T ⁰_s ) ;
Return ( C⁰ );
}
```

7. Experimental Results

This section presents the results of our multi-objective genetic formulation (MGA). The final results i.e the result obtained after completion of CT-GA are compared with PBB algorithm [2] and MGAP algorithm [18]. For TA-GA, we consider NSGA-II multi-objective evolutionary algorithm technique with crossover probability 0.98 and mutation probability 0.01. For CT-GA, we consider NSGA-II with our introduced new crossover and mutation operator. For Table 1 shows the bit-energy value of a link and a switch (4×4) assuming 0.18 μm technology.

Table 1: Bit energy values for switch and link

E_{Lbit}	E_{Sbit}
5.445pJ	0.43pJ

The value of E_{Lbit} is calculated from the following parameters.

(1) length of link (2mm) (2) capacitance of wire (0.5fF/ μm) (3)voltage swing (3.3V)

In our experiment, we consider three random applications, each consisting of 9, 14 and 18 cores respectively. After P-I, we found that the CCG of all three benchmarks consists up less than 9 cores. So, which can be mapped on to a 3×3 mesh NoC architecture. It has been seen that the required bandwidth of an edge connect two different nodes is uniformly distributed over the range [0, 150Mbytes]. The traffic volume of an edge also has been uniformly distributed over the range [0, 1Gbits]. Figure-5 shows the maximum link bandwidth utilization of three benchmarks. It is clear from the figure that our approach (MGA) saves more than 15% link bandwidth as compare to PBB around 5% in compare to MGAP. Figure-6 shows that our approach saves 10%-20%(on average) of energy consumption.

The real time application is a modified *Motion- Joint Photographic Experts Group* (M-JPEG) encoder. Which differs from traditional encoders in three ways: (i) it only supports lossy encoding while traditional encoders support both lossless and lossy encodings.(ii) it can operate on YUV and Red-Green-Blue(RGB) based video data whereas traditional encoders usually operate on the YUV format, and (iii) it can change quantization and Huffman tables dynamically while the traditional encoders have no such behavior. We omit giving further details on the M-JPEG encoder as they are not crucial for the experiments performed here. Interested readers may refer to [15].

Figure-7 shows the bandwidth requirements and energy consumptions for M-JPEG encoder application. From the figure it is clear that our approach out perform other approaches. Figure-8 shows the behavior of NSGA-II with respect to number of generations. Figure-9, shows the speedup of our algorithm in compare to PBB. We see that speedup factor increases with increase of the size of the system.

Fig 5: Maximum Link Bandwidth comparisons for three random benchmarks

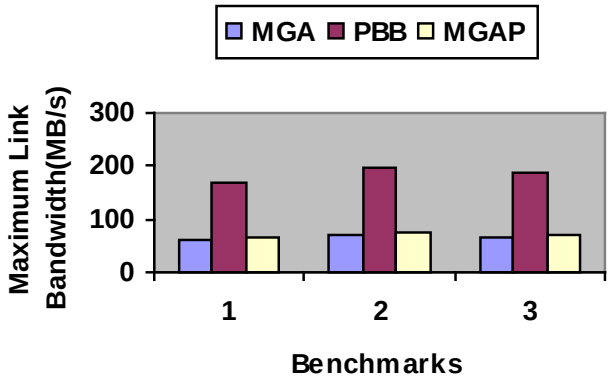


Fig 6: Energy comparisons for three random benchmarks

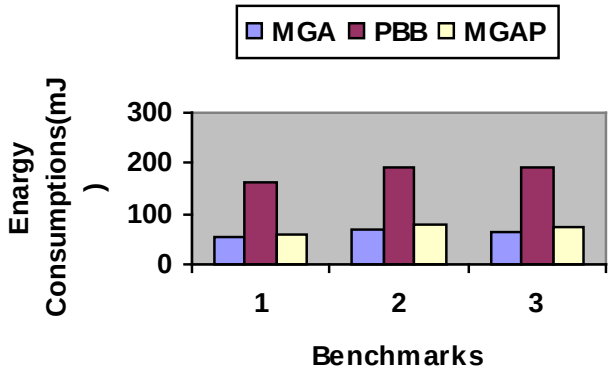


Fig 7: Maximum Link Bandwidth and Energy comparisons for M-JPEG

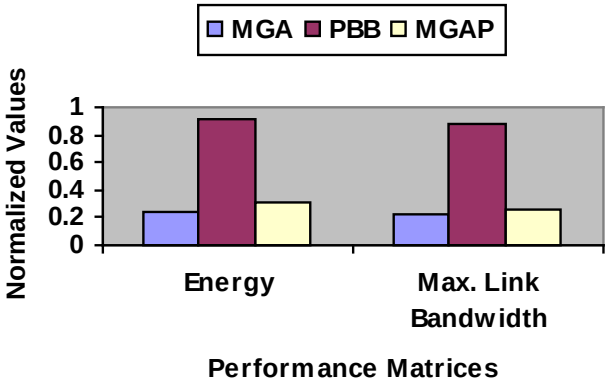


Fig 8 : M-JPEG Encoder performance using NSGA-II

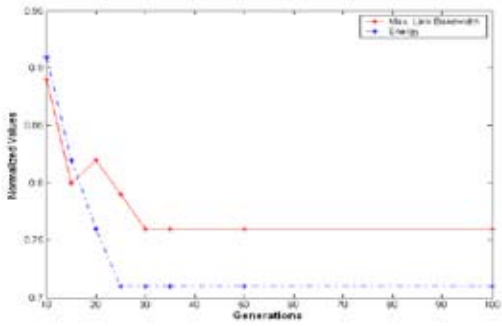
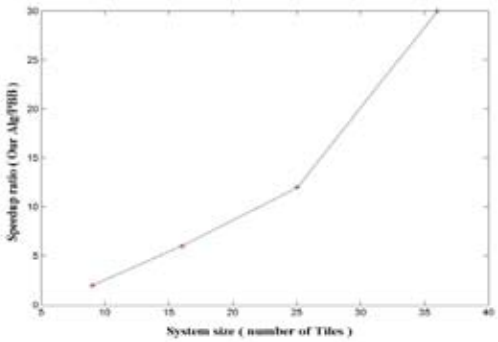


Fig 9: Speedup Comparison between random set of benchmarks



8. Conclusion

This paper proposed a model for topological mapping of IPs/cores in a mesh-based NoC architecture using "many-many" communication between IPs and switches. The approach uses heuristics based on multi-objective genetic algorithms (NSGA-II) to explore the mapping space and find the pareto mappings that optimize Maximum link bandwidth and performance and power consumption. The experiments carried out with three randomly generated benchmarks and a real application (M-JPEG encoder system) confirms the efficiency, accuracy and scalability of the proposed approach. Future developments will mainly address the definition of more efficient genetic operators to improve the precision and convergence speed of the algorithm. To conclude, evaluation may be made possible of optimizing mapping by acting on other architectural parameters such as routing strategies, switch buffer sizes, etc.

Reference

E. Zitzler and L. Thiele. "Multi-objective evolutionary algorithms: A comparative case study and the strength pareto approach". *IEEE Transactions on Evolutionary Computation*, 4(3), pp.257–271, Nov. 1999,.

J. Hu and R. Marculescu. " Energy-aware mapping for tile-based NoC architectures under performance constraints". In *Asia & South Pacific Design Automation Conference*, Jan. 2003.

J. Hu and R. Marculescu "Exploiting the Routing Flexibility for Energy/Performance Aware Mapping of Regular NoC Architectures," in *Proc. DATE'03*pp. 688-693, 2003.

K. Lahiri, A. Raghunathan, and S. Dey, "Efficient Exploration of the SoC Communication Architecture Design Space", in *Proc. IEEE/ACM ICCAD'00*, pp. 424-430, 2000.

M. R. Garey and D. S. Johnson, "intractability: a guide to the theory of NP-Completeness.", Freeman and Company, 1979.

Luca Benini and Giovanni De Micheli, "Networks on Chips: A New SoC Paradigm", *IEEE Computer*, pp. 70–78, January 2002.

S. Murali and G. D. Micheli. "Bandwidth-constrained mapping of cores onto NoC architectures.", In *Design,Automation, and Test in Europe*, IEEE Computer Society, pp. 896–901, Feb. 16–20 2004.

T. Lei and S. Kumar, "A two-step genetic algorithm for mapping task graphs to a network on chip architecture.", In *Euro micro Symposium on Digital Systems Design*, Sept. 1–6, 2003.

S. Kumar *et al*., "A Network on Chip Architecture and Design Methodology," in *Proc. ISVLSI'02*, pp. 105-112, April 2002.

N. Banerjee, P. Vellanki, and K. S. Chatha, "A power and performance model for network-on-chip architectures.", In *Design, Automation and Test in Europe*, pp. 1250–1255, Feb. 16–20, 2004.

T. T. Ye, L. Benini, and G. D. Micheli, "Analysis of Power Consumption on Switch Fabrics in Network Routers," in *Proc. DAC'02*, pp.524-529 June, 2002.

C. A. Coello Coello, D. A. Van Veldhuizen, and G. B. Lamont. Evolu-tionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers, New York, 2002.

Deb, K. (2002). Multi-Objective Optimization using evolutionary Algorithms, John Wiley and Sons Ltd, 2002, pp. 245-253.

K. Srinivasan and Karam S. Chatha, " ISIS : A Genetic Algorithm based Technique for Custom On-Chip Interconnection Network Synthesis", Proceedings of the 18th International Conference on VLSI Design (VLSID'05),2005.

A. D. Pimentel, S. Polstra, F. Terpstra, A. W. van Halderen, J. E. Coffland, and L. O. Hertzberger., Towards efficient design space exploration of heterogeneous embedded media systems. In E. Deprettere, J. Teich, and S. Vassiliadis, editors, Embedded Processor Design Challenges: Systems, Architectures, Modeling, and Simulation, volume 2268 of LNCS, Springer-Verlag, pp.7–73, 2002.

C. J. Glass and L. M. Ni, "The Turn Model for Adaptive Routing," in *Proc.19th Ann. Int'l Symp. Computer Architecture*, May 1992, pp. 278-287.

William J. Dally and Brian Towles. "Route Packet, Not Wires: On-Chip Interconnection Networks",In Proceedings of *DAC*, June 2002.

Jena, R.K, Sharma, G.K, "A multi-objective Optimization Model for Energy and Performance Aware Synthesis of NoC architecture", In Proceedings of IP/SoC,pp477-482, December, 6-7,2006.

3
The Deployment of an E-commerce Platform and Related Projects in a Rural Area in South Africa

Lorenzo Dalvit, Hyppolite Muyingi, Alfredo Terzoli, Mamello Thinyane

In our paper we describe the development and deployment of an ecommerce platform in Dwesa, a rural area in the former homeland of Transkei in South Africa. The system is designed to promote tourism and advertise local arts, crafts and music, and it entails a number of related projects. Deployment of infrastructure, technical support, promotion of the initiative and teaching of computer literacy take place during monthly visits of approximately one week, and involve young researchers from two universities (one previously disadvantaged; the other historically privileged). This ensures a synergy between technical expertise and understanding of the local context. Findings so far emphasise the importance of contextualising the intervention to suit local needs and adjust to the local context. The platform is currently being extended to include e-government, e-learning and e-health capabilities. If proven successful, this model can be exported to similar areas in South africa and in the rest of Africa. This could open up potential opportunities for the still unexplored market for ICT in rural Africa.

Introduction

In this paper we describe the deployment of ICT in Dwesa, a rural community in South Africa. This involves the implementation of an ecommerce platform which can make a contribution to rural development and poverty alleviation in the area. The novelty of our approach is its sensitivity to the context and its emphasis on the promotion of active participation of the community and sustainability. We will start with a description of Dwesa and of the various stages of deployment. We will then describe the e-commerce platform and its related projects. We will end with a discussion of future developments and possible implementations.

Background

Dwesa is a rural community located on the Wild Coast of the former homeland1 of Transkei, in the Eastern Cape Province of South Africa (see Figure 1). In many ways, it is representative of many rural realities in South Africa and Africa as a whole. As noted by Palmer, Timmermans and Fay (2002), its 15000 inhabitants are traditionally subsistence farmers who depend on the land for their livelihood. The region features a coastal nature reserve and it was the site of one of the first restitution projects in postapartheid South Africa.

The region has a high potential for both eco and cultural tourism due to the rich cultural heritage and the marine conservation project undertaken at the nature reserve.

Figure 1: Transkei region and Dwesa

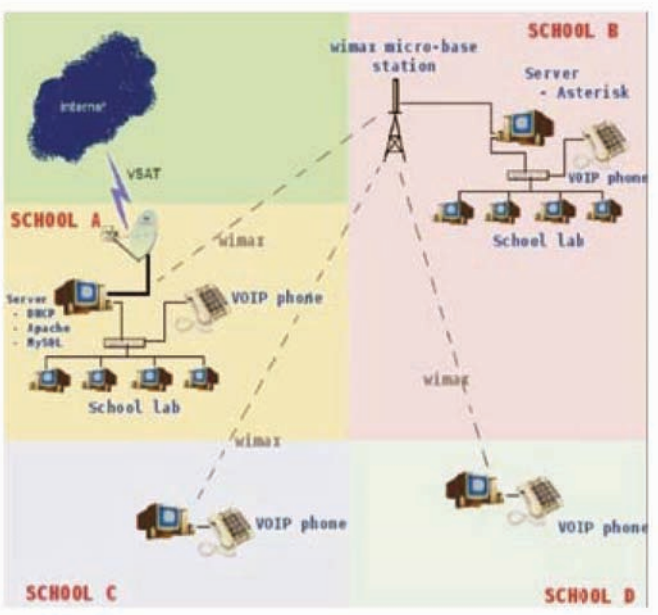
The nature reserve is a catalyst for tourism, which together with government subsidies, is the main source of money for the local community. Tourism is seasonal (almost exclusively during school holidays) and visitors are mostly South Africans. At the moment, the revenues are redirected to the administrative government offices in Bisho and do not benefit the community directly. The only way tourism benefits the community is by promoting local arts and crafts. There are a number of activities ranging from basketmaking to woodcarving. Moreover, Transkei is the site of conservation of the traditional culture of the Xhosa people, and Dwesa has much to offer to tourists and to the outside world in terms of preservation of traditional customs and ceremonies, dances and, especially, music.

Unfortunately, like many rural areas, Dwesa is characterised by lack of infrastructure in terms of road and electricity, widespread poverty, lack of services and isolation (Human Sciences Research Council, 2005). Isolation is probably the main reason for young people leaving Dwesa for the cities, a typical phenomenon in rural areas (Salt, 1992). This deprives the community of fresh energies and of the primary force for change and innovation. Even worse than the physical isolation is isolation in terms of knowledge and information. Flor (2001) discusses and highlights the centrality of the link between access to information (or lack of it) and poverty. Access to relevant information has become one of the discriminating factors between the rich and the poor communities in the world. Some popular views have it that uptodate and readilyavailable information is not a crucial

concern for communities which struggle to satisfy more basic needs such as clean water and electricity. Government sources (Department of Education and Department of Communication, 2001) have highlighted that ICT has a key role to play in contributing to improving the situation of communities which are already disadvantaged in so many other ways.

The project was initiated at the beginning of 2006, and it is partly sponsored by Telkom South Africa. It is a joint venture of the Telkom Centres of Excellence at the University of Fort Hare (a previously disadvantaged university) and Rhodes University (an historically "white" institution). The driving force of the project is a group of young researchers from both universities. This kind of cooperation, still relatively rare in South Africa, ensures a combination of technical expertise and understanding of the territory. Members of the group pay regular monthly visits to Dwesa, and stay there for approximately one week each time. The points of presence so far are four schools: Mpume, Ngwane, Ntokwane and Nodobo (see Figure 2). The equipment deployed is:

Figure 2: Project infrastructure



Mpume (school A): this is the first school that was identified as a point of presence for the project. Due to its location and the availability of electricity, it was selected as the ideal site for the installation of the VSAT[2] connection (satellite dish, indoor unit and cabling). Other equipment include:

- WiMAX[3] (Alvarion Breezemax CPE outdoor unit, and CPE indoor unit, wall mounting)
- A server (LTSP, HTTP, MySQL)

- 6 client PCs
- An 8 port DLink switch
- A VoIP[4] phone

Ngwane (school B): this school is in line of sight from Mpume and has been identified as one of the points of WiMAX installation. Unlike other schools, Ngwane managed to source their own lab (approximately 20 PCs) and a printer. The additional equipment deployed there was:

- WiMAX (Alvarion Breezemax Microbase station, antenna and wall mounting)
- Switch rack (wall mounted) and a 24 port switch
- A server (Asterisk, DHCP)
- A VoIP phone

Ntokwane and Nondobo (schools C and D): these schools are also in line of sight from Mpume, which allows us to provide VoIP and Internet to them. The equipment deployed in each school is minimal:

- WiMAX (Alvarion Breezemax CPE outdoor unit, and CPE indoor unit)
- A client PC
- A VOIP phone

Client PCs and servers alike have 2.67 Celeron processors, 512 MB of RAM, a 40 GB hard disk and CD ROM reader (Thinyane, Slay, Terzoli and Clayton, 2006). The two prerequisites for the ecommerce platform to be effective are Internet connection and basic training in computer skills. Internet connectivity is provided to Mpume via VSAT and then extended to the other schools via WIMAX. Computer training has been run by the members of the team at the various schools.

Ecommerce Platform and Related Projects

The ecommerce platform is designed as a web application that allows local entrepreneurs to engage in trade activities. The platform is designed and developed by members of the research group (i.e. it does not use offtheshelf applications). This allows for customisation to meet the specific needs and profiles of the local users. The platform is designed as a sort of virtual mall in which local entrepreneurs can manage virtual stores. There are three levels of usage. The administrator is responsible for the overall maintenance of the platform and the subsequent management of stores on the emall. The store owners have access to the management of their individual stores. Customers can access the platform over the Internet to purchase goods and services (Njeje, Terzoli and Muyingi 2006).

The ecommerce platform is flexible enough to accommodate a variety of goods and services. The areas where Dwesa has most to offer are tourism, arts and crafts. Different businesses can spin off from tourism: services (e.g. accommodation and catering) and entertainment (e.g. cultural tourism, horse riding, etc.). Recorded

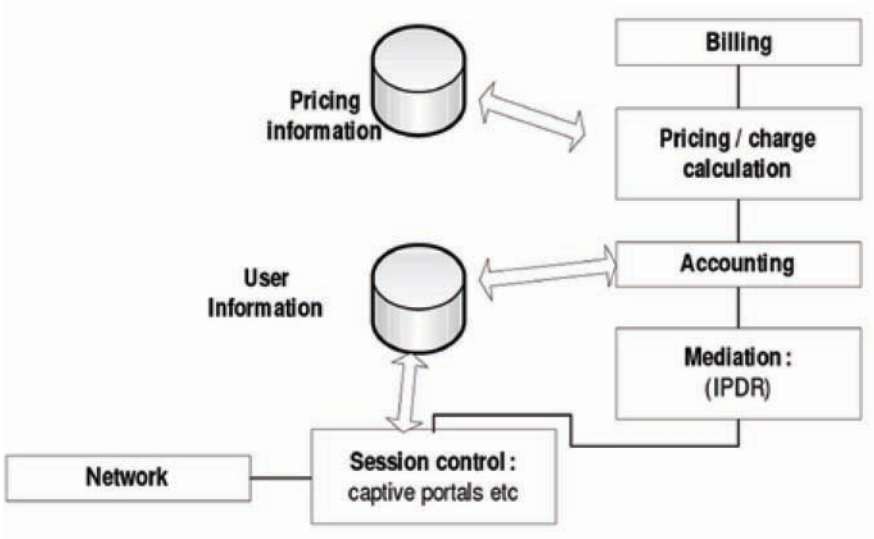
music and traditional artefacts can also be made available for purchase on the ecommerce platform.

Requirements elicitation is the study of the needs of the prospective users of the system. This was important in order to inform the development of the ecommerce platform. In a rural African context, new methodologies and techniques need to be experimented with. The novelty is to adapt established practices to a new and partly unresearched context. This involved extensive interaction with the local community. In many cases, educating the prospective users in ways in which ICT could be integrated into their activities and used to promote their businesses was an integral part of the elicitation process (Isabirye, Roets and Muyingi, 2006).

Recording and online selling of local music: Transkei is the heartland of traditional Xhosa music, and Dwesa has a rich and long musical tradition. An experimental music studio has been set up to record locally produced music, synthesise it and store it in electronic format. Music samples can be made freely available on the ecommerce platform to wet the appetite of tourists and potential customers. Music can then be purchased and downloaded on demand (Ndlovu, Terzoli and Pennels, 2006).

Determining a community costsharing model: One of the main considerations in the implementation of a development project of this nature is to ensure longterm sustainability. The running costs of this project include electricity, Internet connection, maintenance and equipment. These are all costs that have to be shared among the different users of the solution. The billing system collects usage information from the network as well as the amount of sales from the electronic shopping mall's database. The essential system components, presented in Figure 3, are:

Figure 3: Cost sharing system structure



- Network Layer: captures, tracks and records usage of resources (metrics);
- Session control layer: ensures Authentication, Authorization and Accountability on network resources;
- Mediation Layer: formats data provided by meters and forwards it to the accounting layer IPDR [8];
- Accounting Layer: sorts the collected data into accounting records;
- Charging / Pricing Layer: assigns prices to records to come up with costs of usage;
- Billing Layer: compiles all charges for a customer over specific period.

The end of this research is to propose a costsharing framework for the Dwesa community based on the consideration of the different costs and revenue streams (Tarwireyi, Terzoli and Muyingi 2006).

Experimenting with a multilingual approach: Many projects involving the implementation of ICT in rural areas in Africa have failed because of the language barrier posed by the use of English. In our project we experiment with the use of both English and isiXhosa, the local African language. This includes three interventions:

1. The localisation of the user interface of both the ecommerce platform and the computers in the laboratories into isiXhosa;
2. The development and use of teaching material in isiXhosa, to be used alongside the existing material in English;
3. The implementation of a text to speech and screenreader in isiXhosa, as an attempt to tackle the problem of illiteracy or low literacy.

The future plan is to work together with Translate.org.za, an NGO dedicated to the localisation of free/libre open source software (FLOSS) in the 11 South African languages (Translate.org.za 2006), as a test site. Our hope is that this will guarantee access to a larger number of users, and contribute to a sense of ownership in the community.

Implementation of wireless networking: the ICT infrastructure that is to be deployed in Dwesa is built on top of networks that support the interaction between the different stakeholders in Dwesa. So far, four schools in strategic locations have been connected using WIMAX. Community members can access the network from the computer laboratories, through WIFi hotspots linked to the WIMAX base or directly through WIMAX. The new development in wireless networking in terms of range and bandwidth has made it possible to link up and connect regions that would otherwise be too remote to access with traditional wireline technologies (Mandioma, Rao, Terzoli and Muyingi, 2006). This research provides the infrastructural basis for the effective, efficient and far reaching deployment of the ecommerce platform.

Deploying a VoIP PBX system: the availability of a local network will allow for the deployment of a VoIP (Voice over Internet Protocol) telephone system. An experimental version has already been developed and deployed at Rhodes University (Penton and Terzoli, 2004). Its name is iLanga ("sun" in isiXhosa) and it is basically an internal phone system, like the ones used by firms and academic institutions. Four VoIP phones have been deployed at the four schools, and an experimental asterisk server (on top of which iLanga is run) has been deployed at Ngwane. The purpose is to improve communication within the community and with the outside world (Wittington and Terzoli, 2006). Its Web interface has already been localised in isiXhosa (Dalvit, Alfonsi, Mini, Murray and Terzoli 2006) but further customisation is needed to adapt it to the Dwesa context.

Best practices for teaching computer literacy in rural communities: this research is concerned with finding new and effective ways of teaching members of a rural community how to use computers. As pointed out by Czerniewicz (2004), the digital divide begins when one is connected to the Internet, since connectivity and availability of infrastructure by themselves do not guarantee access. In order for the inhabitants of Dwesa to enjoy the advantages of the Internet connection and of the new technologies in general, it is essential to develop a basic level of computer literacy in the area. At the beginning we focused especially on the training of teachers, hoping they would help to promote computer literacy in the school and in the community.

Figure 4: Computer laboratory training

Our software of choice is Edubuntu (Figure 5), the educationally flavoured version of the Ubuntu Linux distribution (Alfonsi and Dalvit, 2006). Edubuntu is specifically designed for use in schools, and offers a capturing user interface as well as the best available free and open educational software.

Figure 5: Edubuntu operating system

Our starting point was the OpenICDL manual, which was considerably shortened and summarised. The three main lessons learnt were:

1. We had to rethink our classification of what constitutes beginners' level material and what is advanced. This entailed a constant process of contextualising and, in some cases, rewriting the teaching material.
2. Some basic technical expertise in loco had to be created. In case of system failure, someone in Dwesa had to have sufficient technical understanding to be able to explain what the problem was over the phone, and possibly be able to tackle the more trivial issues.
3. Motivating and activating the teachers and the community at large was a key point. From the very beginning, we tried to present our intervention as a two way stream, encouraging the community to take ownership of the initiative. The idea was that, building on African rural tradition, everybody will be looking out for what is perceived as common property.

Investigating resistance to the adoption of Information and Communication Technologies (ICT) in rural communities: This project seeks to identify the specific issues that hinder the proliferation of the deployment of ICT solutions in rural communities. In particular, it focuses on the resistance to the adoption of ICT among old people. This group is particularly interesting because younger people tend to leave to go and work in the cities, and old people are the driving economic force of Dwesa. An interesting finding was that, unlike in many other settings, women are the primary driving force in the adoption of ICT.

This might be an attempt to acquire status in a patriarchal society. Another finding was that level of education greatly influences adoption. Uneducated people find it hard to find meaningful ways of integrating ICT in their daily lives. Moreover, given the status associated with technology, it is generally accepted that computers were meant mainly for educated people. This generated some power dynamics which could potentially hamper the adoption of ICT in the area (Mapi, Dalvit and Terzoli 2006).

Future Plans

Expansion of the system: the system is currently being expanded to include egovernment, ehealth and elearning capability. As far as egovernment is concerned, the transfer of personal data will demand extensive work on security. We will experiment with smart cards to try and solve this problem (Nkomo, Terzoli and Muyingi, 2006).

The delivery of health information is a crucial problem in rural areas. The use of ICT might contribute to cut the cost of printing and circulating paperbased material. Elearning could make a great contribution to rural areas. Since schools are our primary point of presence and teachers are one of the driving forces of the project, we feel that focus on elearning is very important.

Additional projects: while some of the subprojects (i.e. investigation of adoption of ICT in rural communities) have been concluded, some others (i.e. deployment of a VoIP system) are still at the experimental stage. The extension of the system with e-government, e-health and e-learning capability mentioned above will result in several additional projects. Java mobile and 3G applications will be developed for deployment in the infrastructure.

A BingBee kiosk (Slay and Wentworth, 2006) will be deployed outside some of the computer laboratories. The distinctive feature of this kiosk is a lowcost touch pad keyboard which enables users to operate a computer kept behind glass. This makes it possible to access the Internet even when laboratories are closed with no security risks.

Additional steps need to be taken towards contextualising the system. One of the many ways of doing this is through more extensive use of multimedia, adding more audio and visual components. This could contribute to make the use of computers more meaningful also to people with low levels of literacy.

Involving the community: as mentioned above, promoting ownership of the project among the community is a crucial factor. Strategies to include the community range from organising public meetings to promote the initiative to running computer literacy training. One possibility would be to institute a committee, elected by the community itself, to promote and organise the initiative locally.

Conclusions

In this paper we described a developmental ecommerce platform and related projects deployed in Dwesa, a rural community in South Africa. The novelty of our approach is its strong emphasis on adjusting to the local context and involving the community. Responsiveness to local needs will be the guiding principle in the future developments of the project.

We hope that our intervention will prove to be a successful example of ICT implementation in a deepl rural setting in Africa. Our goal is to come up with a pervasive and highly distributed ICT solution whose model might possibly be replicated in other parts of Africa.

Acknowledgements

This work was undertaken in joint venture between the Telkom Centre of Excellence in Developmental e-Commerce of the University of Fort Hare and the Telkom Centre of Excellence in Distributed Multimedia of Rhodes University, with financial support from the respective sponsors.

Funding of individual researchers came from the South African National Research Foundation, the Andrew Mellon Foundation and the Opera Universitaria of Trento (Italy). We would like to thank all the members of the Dwesa research team, both from Rhodes and Fort Hare. Most importantly, we would like to thanks the Dwesa community for their enthusiastic support.

References

Alfonsi R.M. and L. Dalvit (2006). "Edubuntu Linux nelle scuole del Sudafrica", *Scienza e Pace, University of Pisa, Pisa*, February 2006. (Accessed 15 June 2006): http://www.scienzaepace.unipi.it/modules.php?na me=News&file=article&sid=203

Czerniewicz L. (2004). "Cape of Storms or Cape of Good Hope? Educational technology in a changing environment". *British Journal of Educational Technology*, Vol. 35, Issue 2, Pp. 145158.

Dalvit, L., R, Alfonsi, B. Mini, S. Murray and A. Terzoli (2006). "The localisation of iLanga, a next generation PBX". SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Dalvit, L., R. Alfonsi, N. Isabirye, S. Murray, A. Terzoli and M. Thinyane (2006). "A case study on the teaching of computer training in a rural area in South Africa". 22nd CESE (Comparative Education Society of Europe) conference, Granada, Spain,

Department of Education (DoE) and Department of Communication (DoC) (2001), Strategy for Information and Communication Technology in Education, Government Printer, Pretoria.

Dwesa Project. (Accessed 15 June 2006): http://dwesa.coe.ru.ac.za/wiki

Edubuntu Linux Operating System. (Accessed 15 June 2006): http://www.edubuntu.org/ Flor A.G. (2001), "ICT and poverty: The indisputable link", 3rd Asian development forum on 'regional economic cooperation in Asia and the pacific', Asian development bank, Bangkok.

Human Sciences Research Council (2005). Emerging Voices: A Report on Education in South African Rural Communities. Cape Town: HSRC Press/Nelson Mandela Foundation.

iLanga. (Accessed 15 June 2006): http://pbx.ict.ru.ac.za/

Isabirye, N., R. Roets and H. Muyingi (2006). "Eliciting user requirements for ecommerce applications for entrepreneurs in rural areas". SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Mandioma, M., K. Rao, A. Terzoli and H. Muyingi (2006). "A feasibility study of WiMax implementation at DwesaCwebe rural areas of Eastern Cape of South Africa", Submitted to IEEE TENCON 2006.

Mapi, T., L. Dalvit, and A. Terzoli (2006). "Adoption of Information Communication Technologies and computer education in a rural area in South Africa". Kenton conference, Southern Cape, South Africa.

Mathias C. (2006), Mobile home for open source, Electronic Engineering Times, Issue 1426, pp. 6060. (Accessed 15 June 2006):http://search.epnet.com/login.aspx?direct=true&db=aph&an=21156235

Ndlovu, G., A. Terzoli and G. Pennels (2006). "Indigenous knowledge: a pathway towards developing sustainable ICT solutions". SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Njeje, S., A. Terzoli and H. Muyingi (2006). "Software Engineering of a Robust, Costeffective E-commerce Platform for Disadvantaged Communities", SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Nkomo, P., A. Terzoli and H. Muyingi (2006). "An open smart card infrastructure for South African eservices". SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Palmer R., H. Timmermans and D. Fay (2002), "From conflict to negotiation: naturebased development on South Africa's Wild Coast". Pretoria: Human Sciences Research Council; Grahamstown: Institute of Social & Economic Research, Rhodes University.

Penton J. and A. Terzoli (2004). "Ilanga, A next generation VOIPbased, TDMenabled PBX", SATNAC conference, Drakensberg, Spier, Stellenbosch, South Africa. (Accessed 15 June 2006):http://www.ru.ac.za/research/pdfs/RESEARCH_REPORT_2004.pdf

Salt J. (1992), The Future of International Labour Migration, International Migration Review, Vol. 26, No. 4. pp. 10771111. (Accessed 15 June 2006): http://links.jstor.org/sici?sici=01979183% 28199224%2926%3A4%3C1077%3ATFOI LM%3E2.0.CO%3B2H

Slay, H. and P. Wentworth (2006). "BingBee: A contactless information kiosk for social enablement". User Interface Software and Technology Conference, Montreaux, Switzerland.

Tarwireyi, P., A. Terzoli and H. Muyingi (2006). "Design and Implementation of an Internet Access Cost Management system for disadvantaged communities", SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Thinyane, M., H. Slay, A. Terzoli and P. Clayton (2006). "A preliminary investigation into the implementation of ICT in marginalised communities". SATNAC conference, Cape Town, South Africa. ISBN 0620370432.

Translate.org.za. (Accessed 19 December 2006): http://translate.org.za

Whittington, B and A. Terzoli (2006), "A low cost, IPbased access loop for consumer telephony in rural communities", ICTe Africa conference, Nairobi, Kenya.

Endnotes

[1]A homeland is the concept of the territory to which one belongs; usually, the country in which a particular nationality was born. [...] In apartheid South Africa the concept was given a different meaning. The white government transformed the 13% of its territory that had been exempted from white settlement into regions of homerule. Then they tried to bestow independence on these regions, so that they could then claim that the other 87% was white territory. See Bantustan.

[2] According to Wikipedia, a *Very Small Aperture Terminal* (VSAT), is a 2way satellite ground station with a dish antenna that is smaller than 3 meters. VSATs are most commonly used to transmit credit card or RFID data for point of sale transactions, and for the provision of Satellite Internet access to remote locations.

[3] WiMAX is defined by Wikipedia as Worldwide Interoperability for Microwave Access by the WiMAX Forum, formed in June 2001 to promote conformance and interoperability of the IEEE 802.16 standard, officially known as WirelessMAN. The Forum describes WiMAX as "a standardsbased technology enabling the delivery of last mile wireless broadband access as an alternative to cable and DSL".

[4] According to Wikipedia, Voice over Internet Protocol, also called VoIP, IP Telephony, Internet telephony, Broadband telephony, Broadband Phone and Voice over Broadband is the routing of voice conversations over the Internet or through any other IPbased network.

4
Computational Identification of Transposable Elements in the Mouse Genome

Daudi Jjingo, Wojciech Makalowski

Repeat sequences cover about 39 percent of the mouse genome and completion of sequencing of the mouse genome [1] has enabled extensive research on the role of repeat sequences in mammalian genomics. This research covers the identification of Transposable elements (TEs) within the mouse transcriptome, based on available sequence information on mouse cDNAs (complementary DNAs) from GenBank [28]. The transcripts are screened for repeats using RepeatMasker [23], whose results are sieved to retain only Interspersed repeats (IRS). Using various bioinformatics software tools as well as tailor made programming, the research establishes: (i) the absolute location coordinates of the TEs on the transcript. (ii) The location of the IRs with respect to the 5'UTR, CDS and 3'UTR sequence features. (iii) The quality of alignment of the TE's consensus sequence on the transcripts where they exist, (iv) the frequencies and distributions of the TEs on the cDNAs, (v) descriptions of the types and roles of transcripts containing TEs. This information has been collated and stored in a relational database (MTEDB) at http://warta.bio.psu.edu/htt_doc/MTEDB/homepage.htm).

1.0 Introduction

1.1 Review

Transposable elements (TEs) are types of repetitive sequences that can move from one place to another and are interspersed throughout most eukaryotic genomes. Their sequence based classification (e.g. SINE [Short Interspersed Element], LINE [Long Interspersed Element], LTR [Long Terminal Repeat]) is frequently based on what is known as a repeat consensus sequence. A repeat consensus sequence [4] is an approximation of an ancestral active TE that is reconstructed from the multiple sequence alignments of individual repetitive sequences. Libraries of such consensus sequences have been compiled and stored in databases like RepBase [5]. Repetitive sequence identifying software like RepeatMasker [3], REPuter [6] and the like rely on such libraries to act as sources of reference sequences against which repetitive sequences from a query sequence(s) are identified. Of these repetitive sequences, transposon-derived IRs or TEs form the largest percentage. In comparison to the human genome, the mouse genome has a higher number of recent TEs that diversify more rapidly than in the human [1]. TEs don't merely

have a pronounced presence in eukaryotic genomes, they also influence these genomes' evolution, structure and functioning in many varied ways; they act as recombination hotspots, facilitate mechanisms for genomic shuffling, and provide ready-to-use motifs for new transcriptional regulatory elements, polyadenylation signals and protein-coding sequences [8]. Though the effects of the mobility of TEs are mostly neutral, in some cases they lead to undesirable mutations, resulting in diseases like Haemophilia B, B-cell lymphoma and other cancers, Neurofibromatosis and many others. These far reaching effects of TEs are an important part of the motivation behind this research. Aswell, some work has been done on inferred molecular functional associations of repeats in mouse cDNAs [7]. This paper concentrates on only TEs in the mouse transcriptome and its coverage is not limited to TEs with some evidence of functionality, but to all TEs in the transcriptome.

1.2 Research objective

With respect to the scope specified above, the research sought, at a minimum to fulfil seven aims and objectives which include establishing; (i) the absolute location coordinates of the TEs on each transcript, (ii) the location of the TEs with respect to the 5'UTR, CDS and 3'UTR sequence features, (iii) the quality of alignment of the TEs' consensus sequence on the transcripts where they exist, (iv) the frequencies and distributions of the TEs on the cDNAs, (v) descriptions of the types and roles of transcripts containing TEs, (vi) collation of the data thus accumulated into a relational database and finally, (vii) the construction of a web-based interface to facilitate access to the information.

2.0 Literature Review.

2.1 Definition and description.

They are discreet units of DNA that move between and within DNA molecules, inserting themselves at random. They are excised or copied from one site and inserted on another site either on the same or on a different DNA molecule. Both their ends are normally inverted repetitive sequences, with the sequence of base pairs on one end being reversed on the other. Their hallmark is non-reliance on another independent form of vector (such as a phage or plasmid DNA), and hence direct movement from one site in the genome to another. [10]

Table 2. Major types of TEs

Type	Replication	Main Families
Interspersed Repeats		
SINE short interspersed	Rely on LINEs	Alu, B1, B2, MIR
LINE long interspersed	Reverse transcription	L!, L2
LTR retrotransposon	Reverse transcription	ERVs, MaLRs
DNA transposon	DNA transposase	Mariner, MERs
Simple Repeats	DNA replication error	

2.6 Effect and Roles of repetitive sequences on the genome

TEs serve as recombination hotspots, providing a mechanism for genomic shuffling and a source of "readyto-use motifs" for new transcriptional regulatory elements, polyadenylation signals, and protein-coding sequences [17]. Transposons can also be disadvantageous. This is either by being inserted into the coding region of a gene, altering some of the gene in the process, or by being inserted upstream of the coding region of a gene in an area important in determining the expression of the gene, for example in an area where a transcription factor would bind to the DNA. The effects of these activities vary, some resulting in genetic disorders like Ornithine aminotransferase deficiency, Haemophilia B, Neurofibromatosis, B-cell lymphoma among others [18]. Some effects of TEs are however neutral or advantageous, some eventually leading to evolutionary novelties like the human glycophorin gene family which evolved through several duplication steps that involved recombination between Alu elements [19] [20] [21] [22]. Genomes evolve by acquiring new sequences and by rearranging existing sequences, and TEs, given their mobility contribute to this process significantly [10]. Transposons also increase the size of the genome, because they leave multiple copies of themselves in the genome and thus occupy upto 38% of the mouse genome [1]. Also, transposons are used in genetic studies, as in the case of being allowed to insert themselves in specific areas so as to "knock out" genes, a technique that turns genes off so that their function can be determined [16].

2.7 Mouse Genome

Genomic resources for the mouse are increasing at an astounding pace and the ability to manipulate the mouse genome and its sequences make the mouse a unique and effective research tool. [4]

2.7.1 Size

All Mus musculus subspecies have the same "standard karyotype" of 20 chromosomes, 19 of them being autosomes and the X and Y sex chromosomes. The 21 chromosomes together are made of a total of 2.751 billion base pairs of nucleotides.

2.7.2 Functional part of the genome.

Genomes show evolutionary conservation over stretches of sequence that have coding potential or any obvious function [23]. However, sequences can only be conserved when selective forces act to maintain their integrity for the benefit of the organism. Thus, conservation implies functionality, even though we may be too ignorant at the present time to understand exactly what that functionality might be in this case [5]. However, looking at functionality in terms of coding potential, the fraction of the mouse genome that is functional is likely to lie somewhere between 5% and 10% of the total DNA present.[5]

2.7.3 Number of genes

The number of genes in the mouse genome has been estimated using using three tiers of input [1]. First, known protein coding cDNAs were mapped onto the genome. Secondly, additional protein-coding genes are predicted using the GenWise [16]. Thirdly, de novo gene predictions from GENSCAN program [17] that are supported by experimental evidence (such as ESTs) are considered. These three strands of evidence are reconciled into a single gene catalogue by using heuristics to merge overlapping predictions, detect pseudogenes and discard misassemblies. These results are then augmented using conservative predictions from the Genie system, which predicts gene structures in the genomic regions delimited by paired 5' and 3'ESTs on the basis of cDNA and EST information from the region. The predicted transcripts are then aggregated into predicted genes based on sequence overlaps. This procedure, has estimated the number of genes in the mouse genome at about 30,000 [1].

2.7.4 Number of transcripts

The number of known mouse transcripts has been determined at about 60,770 represented as full-length mouse complementary DNA sequences [27]. These are clustered into 33,409 'transcriptional units'. Transcriptional unit (TU) refers to a segment of the genome from which transcripts are generated.

3.0 Data Processing

3.1 Background

The number of known full-length mRNA transcripts in the mouse has been greatly expanded by the RIKEN Mouse Gene Encyclopedia project and is currently estimated at about 60,770 [26] [28] clustered into 33,409 'transcriptional units' [27]. Of these transcriptional units, 4,258 are new protein-coding and 11,665 are new non-coding messages, indicating that non-coding RNA is a major component of the transcriptome. 41% of all transcriptional units showed evidence of alternative splicing [26]. In protein-coding transcripts, 79% of splice variations altered the protein product [27]. The Riken Mouse ESTs and cDNAs are deposited in the public databases DDBJ, GenBank, EMBL.

3.2 GenBank

The GenBank database at NCBI [28] provided source of the dataset of the cDNAs of the known mouse transcriptome.

3.2.1 Sequence and data retrieval from GenBank

Table 3. Queries used in data and sequence retrieval

Search	Query	Results
#1	Search Mus musculus[Organism] AND "biomol mrna"* [Properties] AND complete cds [Title]	55664
#2	Search Mus musculus[Organism] AND "biomol mrna"[Properties] AND "srcdb refseq"[Properties]	26562
#3	Search #1 OR #2	82226

The queries above yielded 82226 sequences, downloaded in GenBank and fasta formats (4). These constituted the starting dataset of cDNAs corresponding to the currently known1 mouse transcriptome.

3.3 Eliminating redundancy

3.3.1 Background

GenBank is a highly redundant database. It is thus pertinent that redundancies are expunged from the cDNA dataset.

3.3.2 Patdb

Patdb is software that removes redundancies by merging all identical strings and sub- strings and removing all sequences that are perfect substrings of other sequences. It then concatenates the identifiers of the affected sequences [29]. For example the sequences MEPVQ and MEPVQWT are merged, and if there is another MEPVQWT sequence elsewhere, it is discarded [29]. For our purposes, this not only deals away with redundancies, but also helps assemble the various incomplete cDNA fragments into full-length transcripts. Of the 82226 sequences subjected to patdb, 72697 sequences satisfied the minimum length requirement of patdb (100 nucleotides), meaning that the difference of 9530 sequences were too short to be relevant for the purposes of this research as they are far shorter than the average transcript (usually $>$ 1200 bp). Patdb found 3585 sequences to be either substrings or perfect replicas of other sequences, resulting in a total of 69112 sequences (102037991 bp) as the unique cDNA dataset.

3.4 Mapping Transposable Elements

3.4.1 TE features (identifying characteristics)

As mentioned in chapter 1, TEs have some characteristic distinguishing features, particularly the universal existence of inverted repetitive sequences on either ends of all TEs. This feature and other characteristics like possession of a transposase ORF (Open Reading Frame) and their existence as multiple copies (middle repetitive DNA) within the genome, were the attributes used in computational identification.

3.4.2 TE libraries

Identification of TEs based on their features has enabled the construction of libraries of consensus sequences of various types of TEs. RepBase Update (RU) [32] [33] which is a service of the Genetic Information Research Institute (GIRI) [31] is a comprehensive database of repetitive element consensus sequences. Most prototypic sequences from RU are consensus sequences of large families and subfamilies of repetitive sequences. Smaller families are represented by sequence examples [32].

3.4.3 RepeatMasker

RepeatMasker [3], developed by Arian Smit and Phil Green, is software that screens DNA sequences for low complexity sequences, repetitive/TEs including small RNA pseudogenes, Alus, LINEs, SINEs, LTR * Though the property "biomol mrna" is what is used for the searches, GenBank actually returns cDNAs elements, and others, producing a detailed annotation that identifies all of the repetitive elements in a query sequence [29] [30]. RepeatMasker makes use of RepBase libraries [33], which act as reference points for the identification of repetitive elements in a query sequence. RepeatMasker employs a scoring system to ensure that only statistically significant alignments are shown. It uses statistically optimal scoring matrices derived from the alignments of DNA transposon fossils to their consensus sequences [7]. However, it does not locate all possibly polymorphic simple repetitive sequences. Only di–pentameric and some hexameric repetitive sequences are scanned for, and simple repetitive sequences shorter than 20 bp are ignored.

4.0 Results And Computational Analysis.

4.1 Specialised Object oriented Tools:

Bioperl and Perl: BioPerl [9], is an object oriented form of the Perl programming language [35] which relies mainly on open source Perl modules for bioinformatics, genomics and life science research. It provides reusable Perl modules that facilitate parsing of large quantities of sequence data from various molecular biology programs.

4.2 General analysis:

Within the non-redundant 69112 sequence dataset, command line computational analysis revealed that RepeatMasker identified 47204 repetitive sequences. 20023 of these were simple-repeats, 9583 lowcomplexity repeats and 17598 complex-repeats/TEs. (Table 4)

Table 4. Relative abundances of types of Repetitive sequences

	number of elements*	length occupied	percentage of sequence
SINEs:	9277	1118994 bp	1.10 %
B1s	4260	472916 bp	0.46 %
B2-B4	3852	549367 bp	0.54 %
IDs	654	45780 bp	0.04 %
MIRs	511	50931 bp	0.05 %
LINEs:	1943	572023	bp 0.56 %
LINE1	1567	532555 bp	0.52 %
LINE2	320	33623 bp	0.03 %
L3/CR1	56	5845 bp	0.01 %
LTR elements:	4192	1341750 bp	1.31 %
MaLRs	1220	257647 bp	0.25 %
ERVL	530	160390 bp	0.16 %
ERV_classI	327	95357 bp	0.09 %
ERV_classII	1914	774665 bp	0.76 %
DNA elements:	603	89481 bp	0.09 %
MER1_type	468	68204 bp	0.07 %
MER2_type	74	13225 bp	0.01 %
Unclassified:	47	10171 bp	0.01 %
Total interspersed repeats:		3132419 bp	3.07 %
Small RNA:	133	9704 bp	0.01 %
Satellites:	15	1750 bp	0.00 %
Simple repeats:	19897	891616 bp	0.87 %
Low complexity:	9530	415965 bp	0.41 %

4.3 Filtering out simple and low complexity regions

The emphasis of this research being on TEs, the simple-repeats and low-complexity regions were filtered out by UNIX command line computation to retain a dataset consisting of only TEs/complex repeats. This dataset contains 17598 records representing an equal number of TEs.

4.4 Getting CDS (Coding Sequences) coordinates.

While RepeatMasker output avails the coordinates of the repetitive sequences on a transcript, it does not show the coordinates of the CDS on the transcript. This necessitated the obtaining of CDS coordinates for each transcript identified as possessing a TE by RepeatMasker. This was effected in two different stages.

Stage 1: Involved computing the GI identifiers of all transcripts with TEs. Because some transcripts contain more than one TE, some GI identifiers feature more than once. Thus this stage also involved removing the resultant redundancy. The result was a list of 10213 GI identifiers each with a tab delimited number on its left showing the number of TEs on that particular transcript.

Stage 2: Involved a BioPerl/Perl script which uses the GI list from above to extract the corresponding CDS coordinates for each transcript from the GenBank dataset that was first downloaded (section 3.2.1). The script then stores each GI with its start and end coordinates separated by tabs in a file.

4.5 Computing location and length of each transposable element

Using a Perl language script, the CDS coordinates were used to; implicitly determine the 5'UTR and 3'UTR of each transcript where these exist, establish the total number of TEs in each of the so determined regions with respect to the entire dataset, determine the length of each single TE.

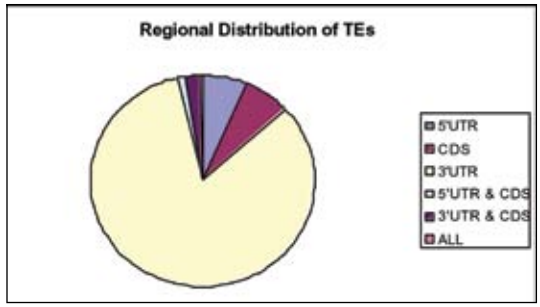
4.5.1 PDB screening of transcripts with TEs in CDS region

The current PDB dataset was downloaded, subjected to patdb to remove redundancies, and then screened against transcripts with TEs in the coding region using blastx. All hits with $>80\%$ identity, alignment length >50, and e -value <0.001 were analysed. None was found to code for a protein of known 3D.

Table 5. Computed TE distribution in sequences and regions.

Initial number of Sequences	82226
Number of nonredundant sequences	69112
Total length (bp)	102037991
GC level (%)	50.01
Sequences with a TE-cassette at all	10213
Sequences with a TE-cassette in CDS	700
TEs lying exclusively in 5'UTR	1179
TEs overlapping 5'UTR and CDS	211
TEs lying exclusively in CDS	1147
TEs lying exclusively in 3'UTR	14618
TEs overlaping3'UTR and CDS	387
TEs overlapping 5'UTR, CDS,& 3'UTR.	59

Fig 2. Pie chart showing relative distribution of TEs in regions



4.6 Average transcript and TE lengths

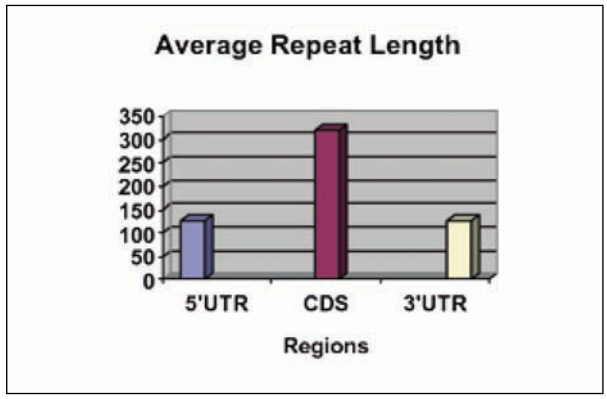
These were calculated from the 69112 non-redundant sequence dataset (see 3.3.2) and the TE lengths files from section 4.5, using a BioPerl/Perl script.

Table 6. Shows average transcript length and average lengths of repetitive sequences in the different regions.

Average transcript length (includes repeatless mRNAs)	1476
Average TE length in 5'UTR	127
Average TE length in CDS	319
Average TE length in 3'UTR	128

averages are calculated to the nearest whole number.

Fig 3. Average regional repetitive sequence lengths.



4.7 Average sequence regional lengths

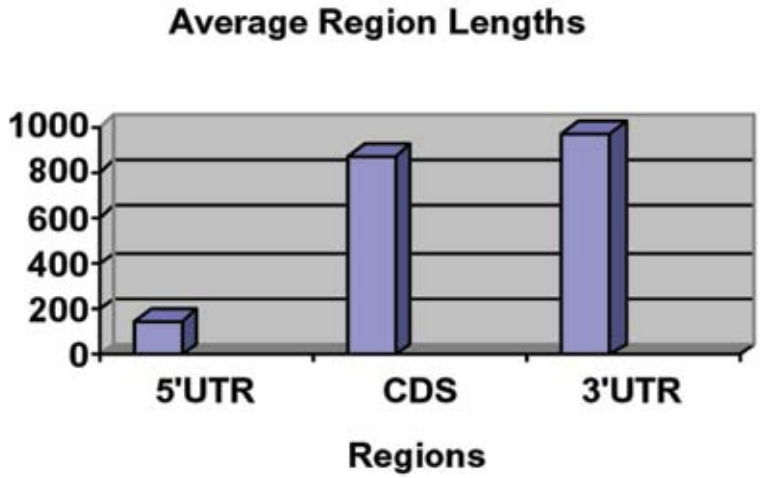
In the context of this research, the primary sequence features are the 5'UTR, CDS and 3'UTR. While these were indirectly alluded to in section s (4.4) and (4.5) for purposes of determining the repetitive sequences in them, this step goes into an

outright and direct calculation of their respective lengths using a Perl/BioPerl script.

Table 7. Average sequence and feature length.

Feature	Average length
5'UTR	150
CDS	877
3'UTR	972
Transcripts (with repetitive sequences)	1998

Fig 4. Average region lengths.



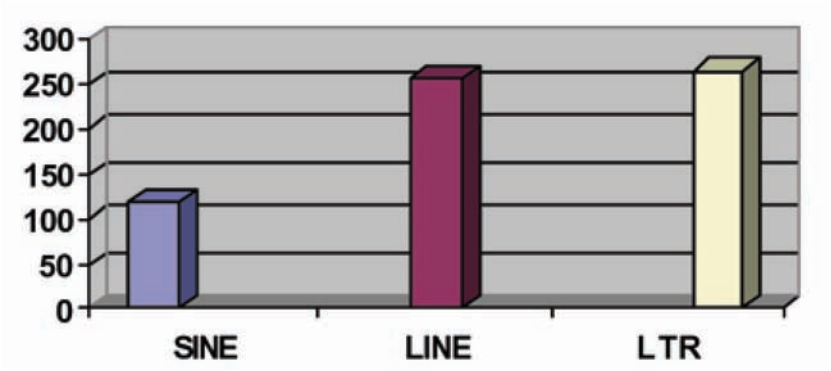
4.8 Average length of major IR families

These were computed using the dataset from the filtering procedure in section 4.3

Table 8. Average lengths of Interspersed repetitive sequence families

IR class	Average length
SINE	119
LINE	257
LTR	265

Fig 5. Average lengths of interspersed repetitive sequences.



4.9 Computing frequencies and occurrence of transposable elements.

TEs tend to move randomly and different transcripts will have varied instances of them. This computational analysis determined the number of transcripts with TEs at all. The analysis was executed in two stages, one involving command line and the other scripted computation. In the first instance, a list of non redundant GI identifiers each tab delimited from the number of TEs it contains (already generated in section 4.4) was analysed to determine occurrence of TEs . It was found that 10213 transcripts or 15% of the 69112 sequences posses TEs. In the second instance the GI list from above was subjected to a Perl script to compute the frequencies of TEs on the transcripts, results of which are shown in the table below.

Table 9. Number of transposable elements per transcript

TEs per mRNA	1	2	3	4	5	6	7	8	9	>=10
Counts	6225	2409	789	366	178	104	58	37	17	30

Fig 6. Showing Distribution of TEs in mouse transcripts

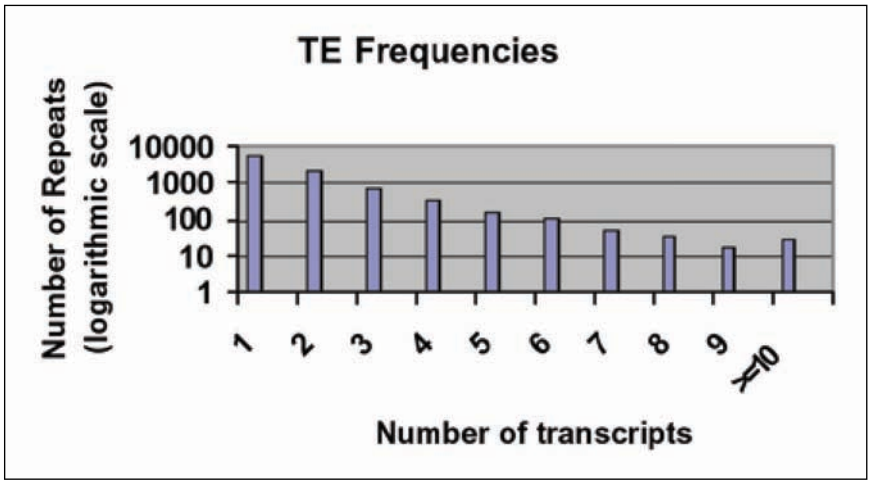
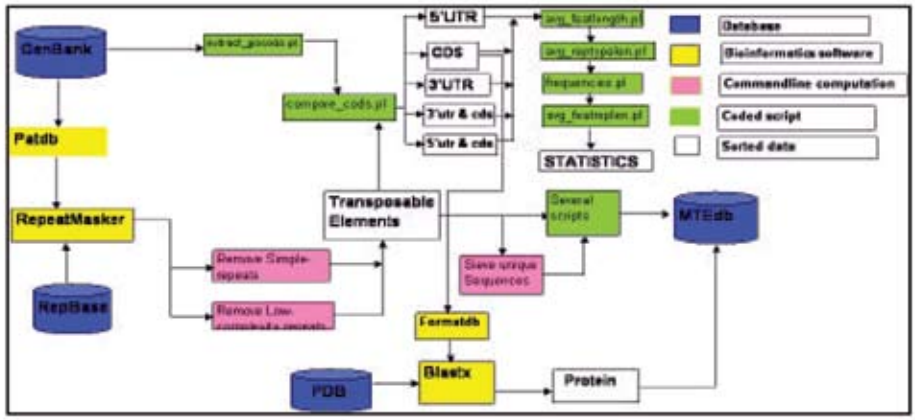


Fig 7. Processing framework used in data processing and computational calculations.



5.0 Discussion

5.1 Quantity versus coverage

Of the 69112 sequences queried, only 10213 or 15% were found to contain repetitive sequences. The actual base pair coverage of these repetitive sequences was 4451205 bp, representing 4.36% of the entire nucleotide set of 102037991 bp. Therefore, within transcripts with CDS information, repetitive sequences cover a much smaller percentage (4.36%) compared to 38% bp coverage in the complete mouse genome. Results in section 4.2 reveal that of the 47204 RepeatMasker identified repetitiv e sequences, low complexity repeats account for 9583 which represents 20% of the total. 20023 simple repeats account for 42%, while complex repeats/TEs represent 37% with a quantity of 17598. However, the coverage in

terms of base pairs is a different picture, with complex repeats (37% by quantity), covering 3.49 times more nucleotides than simple repeats (42% by quantity) and 7.49 times more nucleotides than low complexity repeats (20% by quantity). This is primarily because IRs are much longer in length than their simple and low complexity counterparts, in a measure that more than compensates for their relative inferior quantity.

5.2 Interspersed repetitive sequence occurrence

This project's computational analysis (section 4.3), it must be emphasised, concentrated on IRs/TEs - (excluded simple, satellite and low complexity repeats). The analysis indicated that the mouse genome is dominated by three major classes of IRs (Table 4) namely SINES, LINES and LTR elements. Between themselves, these three classes accounted for 95% of IR (interspersed repeat) sequence coverage. Further still, even within these three, our results revealed a disproportionate presence of some subclasses. The L1 family dominated the LINE class, accounting for 93% of all LINE sequence coverage, leaving L2 and L3 families to share the remaining 6% in a ratio of 3:1 respectively. The SINE class was predominantly shown to be composed of B elements; B1 elements dominated with 44%, followed by B2 elements with 18%, while B3 and B4 elements together contribute 26%. This leaves a mere 12% coverage for the other SINE classes; IDs and MIRs. Within the LTR class, ERV_class II and MaLRs represented 58% and 19% sequence coverage respectively, leaving a dismal 12.07% to be shared among the ERVL and ERV_ class I families.

5.3 Distribution

The position information from the computational analysis in section 4.4 and the statistical results in section 4.5, indicate that IRs can occur in any given region (5'UTR, CDS or 3'UTR) of a transcript. However it appears there is a fairly significant bias towards insertion in the 3'UTR region. The higher numbers of IRs in 3'UTR (see Table 10 below) could be attributed to the fact that this region is longer than both the other two. While that might be part of the reason, a closer examination of the relative lengths of the three regions vis a vis the relative numbers of IRs in each region seems to lend support to the idea of a preference of insertions into the 3'UTR. For instance while the 3'UTR is on average only 1.11 times longer than the CDS region, computational analysis reveals that it contains 15 times as many IRs (Tables 5 and 10). The disparity is not as high in comparison with the 5'UTR but holds nevertheless. 3'UTR is 6 times as long as 5'UTR, but contains 14 time s as many IRs. Perhaps the most conspicuous observation is the statistical revelation of the fact that IR insertions are disproportionately much lower in the CDS region when contrasted with insertion in the UTRs. We have already illustrated this fact with respect to the 3'UTR region. A similar trend can be observed when the CDS region is compared to the 5'UTR region. While the absolute number of IRs within the two regions is approximately equal, with the

ratios being roughly 1:1, it becomes clear that their density within the CDS is much less given the fact that the average CDS region length is 5 times as long as the average 5'UTR length. This observation leads to interesting questions about how IRs or TEs relate to the translated part of the mouse genome. This would imply that while insertions have an equal chance of falling into any region, most IR insertions within the CDS lead to negative consequences for the organism. As a result, these are selected against during genomic evolution. Even at the higher level of transcripts, further evidence of the general undesirability of IR insertion can be adduced, supported by the fact that of the 69112 transcripts that were screened only 10213 or 15% contained any repetitive sequence of any type at all. (section 4.4 and Table 5). Moreover, even within sequences that possessed TEs, the occurrence of sequences containing a given number of TEs dropped logarithmically as the number of TEs per mRNA increased (diagram 6); yet another pointer to the general undesirability of TE insertion. However the mere presence of IRs in the CDS, even though they are relatively fewer in comparison to other regions, bears evidence to the fact that some few insertions into the CDS result in positive or neutral consequences for the host organism and are able to be retained during the course of genomic evolution. One notable and quite interesting observation though, is the fact that the mean length of TEs within the CDS region is significantly higher than those of the other two regions (Table 6). Whether this is of any functional or evolutionary significance is an issue that cannot be appreciably resolved by the scope of this research. In all cases, IRs that overlap or exist in more than one region do not seem to be favoured by genomic evolution. In our case, only 59 or ~0.0025% of the 17598 TEs overlap with all three regions of a transcript, and only 211 or ~0.01% overlap both the 5'UTR and the CDS regions. A mere 387 or 0.02% overlap the CDS and 3'UTR regions.

Table 10. Distribution of repetitive sequences in different mRNA regions

Region	Average Regional Length	Number of IRs
5'UTR	150	1179
CDS	877	1147
3'UTR	972	14618
5'UTR/CDS/3'UT R	1998	59
5'UTR/CDS	N/A	211
3'UTR/CDS	N/A	387

5.4 Possible sources of error

It's pertinent to point out at this point of the discussion that all the results used in the preceding discussion constitute a marginal error arising out of RepeatMasker's identification method. It detects only major simple (di-hexameric repeats) having more than 20 nucleotides, and therefore misses a small number of simple repeats.

Another possible source of error is that a small fraction of potential new mouse repetitive sequences that lack a consensus sequence in RepeatMasker's library against which the query sequences are searched, may either be missed, or incorrectly identified and positioned.

8.0 Conclusions And Further Directions

8.1 Conclusions

While the spread of TEs within a genome is very much a random process, some types of repetitive sequences have been more successful than others. The mouse genome is dominated by three major classes of IRs, namely SINES, LINES and LTRs which between themselves, represent upto 95% of Interspersed repeat coverage in the genome (Table 4). However, even within these three types, some subfamilies are more dominant than others such as the L1 family for the LINEs, the B elements in the SINEs and the ERV_class II in LTRs. The presence of TEs is however much more pronounced in the UTRs than in CDS. This pattern of TE distribution within the transcripts offers two interesting phenomena. On the one hand, the replication and movement of repetitive sequences within the genome as a whole has been a big success, to the extent that they occupy upto 39% of the mouse genome [1]. On the other hand however, their insertion within the CDS, the coding part of the genome, has been relatively minimal. This pattern seems to lend credence to the idea that though insertion of TEs in the genome may lead to desirable evolutionary novelties, for the most part its effects are negative, sometimes fatal and it's therefore selected against by genomic evolutionary pressures and thus their relatively diminished presence in the CDSs. This observation not withstanding, the mere presence of some TEs within the CDSs, attests to the fact that a few of the insertions lead to positive effects and are thus conserved within the region. Their very successful presence and highly conserved status within the non-coding parts of the genome, is further proof to previous observations that even out of the coding region, they serve useful purposes like acting as recombination hotspots [8]. The mouse transposable element database (MTEDB) represents a major resource for the study of functional genomics in the mouse, particularly the complex and intricate phenomena of repetitive sequences in the organism. The database is certainly not perfect, partly because of the continued discovery of novel types of repetitive sequences, shortfalls in the software and methods used among other reasons. It's our belief that though the database may not be perfect, or even complete, it still provides a strong foundation for studying and building of new mouse repetitive sequence data sources, mainly because of the multi dimensional data analysis tools that it offers. A web-based interface for the database, MTEDB can be found at http://warta.bio.psu.edu/htt_doc/MTEDB/homepage.htm.

8.2 Further directions

Taking into account the high abundance and redundancy of repetitive sequences in the mouse genome, it is evidently clear that though the contribution of this and other research efforts is important, there is still an awful lot that science is yet to discover. This is especially so with respect to our knowledge of the biological roles of repetitive sequences and their function in the generation and evolution of the various intricate genomic networks.

Though this research included screening of the data set against the currently known protein structures from PDB, no repetitive sequences were found to have any direct structural information. This was mainly because the PDB is itself still a very limited database, representing only a very small percentage of proteins because the number of known protein structures is still very limited. Further research therefore, could include structural modeling of transcripts with TEs to boost our understanding of the effects of TEs on protein structures.

Another important area of further research could include the study of repetitive sequences in the untranscribed part of the mouse genome and their possible influence on the resultant proteome, say as regulatory regions for gene expression.

Comparison of this project's database with other emerging or existing databases on the mouse genome in general or mouse repetitive sequences in particular is another area of research that would serve to enrich the data set and act as a point of cross reference.

9.0 References

Waterston, R.H, et al. (2002) Initial sequencing and comp arative analysis of the mouse genome. *Nature*, 420. 520-562

Schonbach, C. (2004) From masking repeats to identifying functional repeats in the mouse transcriptome. *Briefings in Bioinformatics*. Vol 5, No.2 107-117.

Smit, A.F.A. and Green, P. RepeatMasker at http://repeatmasker.org

http://www.ncbi.nlm.nih.gov/genome/guide/mouse/

Silver, L.M. (1995) Mouse Genetics Concepts and Applications.Oxford University Press, Oxford.

Kurtz, S. and Schleiermacher, C. (1999) Fast Computation of Maximal Repeats in Complete Genomes, Bioinformatics, 15(5), pp. 426-427.

Nagashima, T. et al. (2004) FREP: A Database of Functional Repeats in Mouse cDNAs. *Nucleic Acid Research, Vol.32*, D471-D475.

Lorenc, A. and Makalowski, W. (2003) Transposable elements and vertebrate protein diversity. Genetica 118: 183-191.

http://www.bioperl.org/

Lewin, B. (1997) Genes. Oxford University Press, Oxford, Newyork, Tokyo, Ibadan.

Makalowski, W. (2001) The human genome structure and organization. Acta Biochimica Polonica 48: 587-598.

Pietrokovski, S. and Henikoff, S. (1997) A helix-turn-helix DNA-binding motif predicted for DNA - mediated transposases. Molecular & General Genetics 254, 689-695.

Dawkins, R. (1982) The Extended Phenotype. Oxford University Press, Newyork.

Russel, P. (2002) iGenetics. Benjamin Cummings publishers, Newyork.

Makalowski, W. (2003) Not Junk After All. Science 300: 1246-1247.

Birney, E. and Durbin, R. (2000) Using GeneWise in Drosophila annotation experiment. Genome Res.10, 547-548.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

Maria, R. (1995) The Impact of Short Interspersed Elements (SINEs) on the Host Genome. R.G. Landes company, Austin, Texas USA.

Fukuda, M. (1993) Molecular genetics of the glycophorin A gene cluster. Semin Hematol 30: 138- 151.

Onda, M., Kudo, A., Rearden, M., and Fukuda, M. (1993) Identification of a precursor genomic segment that provided a sequence unique to glycophorin B and E genes. Proc Natl Acad Sci U S A 90: 7220-7224.

Rearden, A., Magnet, A., Kudo, S. and Fukuda, M. (1993) Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution. J Biol Chem 268: 2260-2267.

Rearden, A., Phan, H., Kudo, S. and Fukuda. M. (1990) Evolution of the glycophorin gene family in the hominoid primates. Biochem Genet 28: 209-222.

Kevles, D. and Hood, L. (1992) The Code of Codes: Scientific and Social Issues in the Human Genome Project (Harvard University Press, Boston)

Milner, C. M. and Campbell, R. D. (1992) Genes, genes and more genes in the human major histocompatibility complex. BioEssays 14: 565-571.

Hasties, N.D. and Bishop, J.O. (1976) The expression of three abundance classes of messenger RNA in mouse tissues. Cell 9: 761-774.

Hidemasa, B., et al. (2003) Systematic Expression Profiling of the Mouse Transcriptome Using RIKEN cDNA Microarrays: Genome Res.2003 June;13 (6b): 13181323.

Okazaki, Y., et al. (2003) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563 – 573.

http://www.ncbi.nlm.nih.gov/

Korf, I., Yandell, M. and Bedell, J. (2003) Blast. O'Reilly & Associates, Inc., California, Cambridge, Tokyo.

Lander, E.S., Linton, L.M., et al. (2001) "Initial sequencing and analysis of the human genome." Nature 409:860-921.

http://www.girinst.org/

5
Properties of Preconditioners for Robust Linear Regression

V. Baryamureeba and T. Steihaug

In this paper, we consider solving the robust linear regression problem $y = Ax + \in$ by an inexact Newton method and an iteratively reweighted least squares method. We show that each of these methods can be combined with the preconditioned conjugate gradient least square algorithm to solve large, sparse systems of linear equations efficiently. We consider the constant preconditioner A^TA and preconditioners based on low-rank updates and downdates of existing matrix factorizations. Numerical results are given to demonstrate the effectiveness of these preconditioners.

1. Introduction

Consider the standard linear regression model

$$y = Ax + \in, \qquad (1)$$

Where $y \in \Re^m$ is a vector of observations, $A \in \Re^{mxn}$ $(m > n)$ is the data or design matrix of rank n, $x \in \Re^m$ is the vector of unknown parameters, $\varepsilon \in \Re^m$ and is the unknown vector of measurement errors. The residual vector r is given by

$$r(x) = y - Ax$$
$$r_j(x) = y_j - A_j. x, j = 1, ..., m,$$

where $A_j.$ denotes the j th row of A. In the least squares method we minimize $\frac{1}{2}\|r(x)\|_2^2$ In order to reduce the influence of outliers (observations that do not follow the pattern of the majority of the data and show extremeness relative to some basic model) we will consider the problem

$$\min_x f(x) \equiv \min_x \sum_{j=1}^{m} ? (r_j(x) /?) \qquad 2$$

where ρ is a given function, and σ is a scale factor (connected to the data). The least squares problem is recovered by choosing $\rho(z) = z^2/2$ for $z \in \Re$. The functions ρ we are interested in are those that are less sensitive to large residuals. The variance of the observation errors is assumed to be known, and thus the scale σ is fixed. This corresponds to the situation where the variability of the process is under (statistical) control. The statistical properties have been thoroughly studied in the literature, see, for example (Huber, 1981). The linear regression model (1) has wide

applications in inverse theory (see, for example Farquharson and Oldenburg (1998) and Scales et al. (1988), and parameter estimation (see, for example Chatterjee and M¨achler (1997)).

If $\rho(z)$ is a convex function (2) will have a unique minimum. We will consider functions that are twice continuously differentiable almost everywhere, with nonnegative second derivatives wherever they are defined. For the case when the variance is unknown, it is possible to obtain simultaneous estimates of the scale factors and the solution (Ekblom, 1988; Shanno and Rocke, 1986) but we will not pursue this case here. However, the results in this paper can be extended to the case when the variance is unknown.

The idea of using preconditioned iterative methods like preconditioned conjugate gradients has not been used much in solving linear regression models (O'Leary, 1990; Scales et al., 1988) as compared to other approaches (e.g. direct methods). The subject of this paper is to suggest preconditioners with a theoretical backing combining the good properties of both direct and iterative methods for use in solving the robust linear regression problem. The paper explores the possibility to use a fixed preconditioner, corresponding to setting all weights to one. In addition we give numerical results to illustrate various properties of the preconditioners.

In Section 2 we formulate the problem as a sequence of weighted linear systems. We suggest preconditioners in Section 3. Section 4 is on implementation details. We discuss numerical results in Section 5 and concluding remarks are given in Section 6.

Throughout this paper we use the following notation: The symbol min_i or max_i is for all i for which the argument is defined. For any matrix A, A_{ij} is the element in the i^{th} row and j^{th} column, and $A_{j\cdot}$ is the j^{th} row. For any vector x, x_j, is the j^{th} component, $|x|$ is the vector of absolute values $|x_j|$ and $x>0$ means all the components of x are positive. For any matrix, vector, or scalar the superscript (k) means the corresponding variable at the k^{th} iteration. The symbol I is used to denote the (square) identity matrix; its size will always be apparent from the context. For any square matrix X with real eigenvalues, $\lambda i(X)$are the eigenvalues of X arranged in nondecreasing order, λ_{min} and λ_{max} denote the smallest and largest eigenvalues of X respectively; i.e.

$$\lambda_{min}(X) \equiv \lambda_1(X) \leq \ldots \leq \lambda_n(X) \equiv \lambda_{max}(X).$$

If X is a symmetric positive definite matrix, then

$$\lambda_i(X^{-1}) = 1/\lambda_{n-i+1}(X) \tag{3}$$

and the spectral condition number of X is $\kappa(X) \equiv \lambda_{max}(X)/\lambda_{min}(X)$ The letter L represents a lower Cholesky factor of a symmetric positive definite matrix. For any set $\mathfrak{I}$, $|\mathfrak{I}|$ is the number of elements in $\mathfrak{I}$.

2 Problem Formulation

For a known scale the first order conditions for (2) are given by

$$F(x) \equiv \nabla f(x) \sum_{j=1}^{m} \nabla \rho(r_j(x)/\sigma) = -\frac{1}{\sigma} A^T v = 0 \text{ or } \frac{1}{\sigma} A^T v = 0 \quad (4)$$

where v is an m – vector with elements $\rho'(r_j(x)/\sigma)$.

2.1 Newton's Method Approach

The inexact Newton method for solving the nonlinear system (4) with line-search (Dembo et al., 1982; Dembo and Steihaug, 1983) is given in Algorithm 2.1.

Algorithm 2.1 *Inexact Newton Method with Line-Search Given initial* $x^{(0)}$ *and* $\eta^{(0)} < 1$.

For $k=0$ step 1 **until** convergence **do**
Let $\eta^k \le \eta^{(0)}$ and find some $\Delta x^{(k)}$ *that satisfy*

$$|| F(x^k) + F'(x^k) \Delta x^{(k)} || \le \eta^{(k)} || F(x^k) || \quad (5)$$

Find $\alpha^{(k)}$ so that $(x^{(k)} + \alpha^{(k)} \Delta x^{(k)})$ is sufficiently less than $f(x^{(k)})$
Set $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$.

To guarantee convergence a line-search is performed along the search direction $\Delta x^{(k)}$. The search direction will be a direction of descent (Dembo and Steihaug, 1983) when the system is solved approximately using a preconditioned conjugate gradient method. The approximate solution $\Delta x^{(k)}$ will have a residual $q^{(k)}$.

$$(1/\sigma) A^T G^{(k)} A \Delta x^{(k)} = A^T v^{(k)} + q^{(k)} \quad (6)$$

where $|| q(x^k) || \le \eta^{(k)} || A^T v^{(k)} ||$ and $G^{(k)} \in \Re^{mxm}$ is a diagonal matrix with diagonal elements

$$G_{jj}^{(k)} = \rho''(r_j(x^{(k)})/\sigma) \ge 0 \; j = 1, \ldots, m. \quad (7)$$

For Newton's method choose $\eta^{(0)} = 0$.

2.2 Iteratively Reweighted Least Squares (IRLS) Method

The IRLS linear system corresponding to (4) differs from the Newton equation (6) in the weight matrix . For the IRLS method (Antoch and Ekblom, 1995; Ekblom, 1988).

$$G_{jj}^{(k)} = \frac{\rho'(r_j(x^{(k)})/\sigma)}{r_j(x^{(k)})/\sigma} \; j = 1, \ldots, m, \quad (8)$$

2.2.1 Approach

On the other hand, Scales et al. (1988) suggest another formulation of the IRLS algorithm based on the optimization problem:

$$\min_x \sum_j |r_j(x)|^p, \ 1 \le p$$

where $r_j(x)$ is defined in (1). The first order condition for (9) is the generalized normal equation

$$A^T G A x = A^T G y \tag{10}$$

for $r_j(x) \neq 0$ and G is the diagonal matrix with diagonal elements $|r_j(x)|p - 2$. For $p = 2$ (least squares) the usual normal equations are recovered. On the other hand, for $p < 2$ the weighting matrix G tends to eliminate the influence of outliers in the data by diminishing the influence of large residuals.

For $k \ge 0$ we solve (10) for $x^{(k+1)}$ by solving a sequence of problems of the form

$$A^T G^{(k)} A x^{(k+1)} = A^T G^{(k)} y, \ k \ge 0 \tag{11}$$

Choosing the weights$\{G_{jj}^{(k)}\}$ to be the inverses of residuals may cause the algorithm to be unstable for very small residuals. Huber (1981) suggests replacing $r_j(x^{(k)})$ with some lower cut–off value μ whenever $|r_j(x^{(k)}) < \mu$. This, in addition to eliminating zero residuals, is sufficient to guarantee convergence of the IRLS algorithm (Scales et al., 1988). The cut–off value μ implies that beyond a certain point, all high-confidence observations (small residuals) are weighted the same.

2.3 General Iterative Scheme

Most of the algorithms for robust regression can be described by the following general iterative scheme:

Algorithm 2.2 *General Iterative Scheme*
Let the initial guess $x^{(0)}$ be given.

For $k = 0$ **step 1 until** convergence **do**
Compute right hand side $h^{(k)}$ and diagonal weight $G^{(k)}$
Find approximate solution $\Delta x^{(k)}$ of

$$A^T G^{(k)} A \Delta x^{(k)} = h^{(k)} \tag{12}$$

Find $\alpha^{(k)}$ so that $f(x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$ is sufficiently less than $f(x^{(k)})$.
Set $x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$.

We remark that the difference between the methods is essentially in the diagonal matrix $G^{(k)}$. For the IRLS method $\alpha^{(k)} = 1$ will give convergence (Huber, 1981).

In Table 1 we state some functions and their derivatives (Coleman et al., 1980; Holland and Welsch, 1977). Except for the Talwar function the functions in Table 1 are convex functions.

Table 1: The parameter β is problem dependent. ρ' and ρ'' stand for the first and second derivatives of ρ

Function	$\rho(z)$	$\rho'(z)$	$\rho''(z)$
Huber	$\begin{cases} \frac{z^2}{2} \text{if} \lvert z \rvert \le \beta \\ \beta \lvert z \rvert - \frac{\beta^2}{2} \text{ if} \lvert z \rvert > \beta \end{cases}$	$\begin{cases} 1 \text{ if } \lvert z \rvert \le \beta \\ \frac{\beta}{\lvert z \rvert} \text{ if} \lvert z \rvert > \beta \end{cases}$	$\begin{cases} 1 \text{ if } \lvert z \rvert \le \beta \\ 0 \text{ if} \lvert z \rvert > \beta \end{cases}$
Talwar	$\begin{cases} \frac{z^2}{2} \text{ if } \lvert z \rvert \le \beta \\ \frac{\beta^2}{2} \text{if } \lvert z \rvert > \beta \end{cases}$	$\begin{cases} 1 \text{ if } \lvert z \rvert \le \beta \\ 0 \text{ if } \lvert z \rvert > \beta \end{cases}$	$\begin{cases} 1 \text{ if } \lvert z \rvert \le \beta \\ 0 \text{ if } \lvert z \rvert > \beta \end{cases}$
Logistic	$\beta^2 \log(\cosh(\frac{z}{\beta}))$	$(\frac{z}{\beta})^{-1} \tanh(\frac{z}{\beta})$	$1 - \tanh^2(\frac{z}{\beta})$
Fair	$\beta^2 \left(\frac{\lvert z \rvert}{\beta} - log\left(1 + \frac{\lvert z \rvert}{\beta}\right) \right)$	$\beta(\beta + \lvert z \rvert)^{-1}$	$\beta^2(\beta + \lvert z \rvert)^{-2}$

3. Preconditioner

Let $G, H \in \Re^{mxm}$ be semi-definite diagonal matrices. Here, A^THA is a nonsingular coefficient matrix (12) from a previous iteration with a known Cholesky factorization. We construct a preconditioner for the current coefficient matrix A^TGA by carrying out low-rank corrections to A^THA (Baryamureeba et al., 1999; Wang and O'Leary, 1995). We want to solve the linear system with coefficient matrix A^TGA and employ a preconditioned conjugate gradient least squares (PCGLS) method with a preconditioner A^TKA, where K is an m x m diagonal matrix constructed from H and G so that the difference between A^TKA and A^THA is a low-rank matrix. In this case, the Cholesky factorization of A^THA can be effectively used to solve linear systems with the preconditioner A^TKA as coefficient matrix.

For a given index set

$$Q \subseteq \{j: 1 \le j \le m \text{ and } G_{jj} H_{jj}\},$$

let the m x m diagonal matrices D and K be given, respectively, by

$$D = K - H$$

$$\text{where } K_{jj} \begin{cases} G_{jj}, & \text{if } j \in Q \\ H_{jj} & \text{otherwise} \end{cases} \tag{13}$$

Let $\bar{A} \in \Re^{q \times n}$, where $q = |Q|$ consist of all rows A_j. such that $j \in Q$ and let $\bar{D} \in \Re^{q \times q}$ be the diagonal matrix corresponding to the nonzero diagonal elements of D. In this notation,

$$A^T KA = A^T(H + D)A = A^T HA + A^T \bar{D} \bar{A},$$

namely, $A^T KA$ is a rank q-correction of $A^T HA$. The elements in the preconditioner are determined by the choice of the index set Q.

Theorem 3.1 (Baryamureeba et al., 1999): Let $K, G \in \Re^{m \times m}$ *be positive definite diagonal matrices. Then*

$$\min_j \left\{ \frac{G_{jj}}{K_{jj}} \right\} \leq \lambda_i (A^T KA)^{-1} A^T GA) \leq \max_j \left\{ \frac{G_{jj}}{K_{jj}} \right\} \quad \blacksquare$$

Let $H \in \Re^{m \times m}$ be a positive definite diagonal matrix. For $Q \neq \varnothing$ let K in Theorem 3.1 be defined as in (13). Then the bounds in Theorem 3.1 simplify to

$$\min \left\{ 1, \min_{j \in Q} \{ \frac{G_{jj}}{H_{jj}} \} \right\} \leq \lambda_i \ (A^T KA)^{-1} \ A^T GA) \leq \max \left\{ 1, \max_{j \in Q} \{ \frac{G_{jj}}{H_{jj}} \} \right\} \tag{14}$$

since $Kjj = Gjj$ for $j \in Q$ and $Kjj = Gjj$ for $j \in Q$.

The preconditioned conjugate gradients least squares (PCGLS) method attain rapid convergence when the spectral condition number of the preconditioned matrix is close to one. With this in mind, (14) suggests that we should choose the index set such that the spectral condition number $k(K^{-1}G)$ is minimized. The preconditioner $A^T KA$ in (14) is directly applicable to weighted linear systems with positive definite diagonal weight matrices. Needless to mention, we can also apply this preconditioner directly to problem (11) with a cut–off value $\mu > 0$.

The normal equations arising from the application of an interior point method to a linear programming problem (see for example Baryamureeba et al. (1999)) are also of the form (12). The idea of using low-rank corrections is not new. In his seminal paper on the polynomial-time interior-point method for linear programming, Karmarkar (1984) proposed to use low-rank updates to decrease the theoretical complexity bounds. His idea has been pursued by many later papers for example Goldfarb and Liu (1991) for the similar reason. Moreover, a combination of a direct method and an iterative method was reported by Karmarkar and Ramakrishnan

(1991). Wang and O'Leary (1995) propose a strategy of preconditioning the normal equation system (with positive definite diagonal weight matrices $G^{(k)}$) based on low-rank corrections, and in their implementation the index set Q consists of the indices corresponding to the largest absolute changes of $G^{(j)} (=H)$ and $G^{(k)} (= G)$, $j<k$. Baryamureeba et al. (1999) also construct preconditioners for the normal equation system based on low-rank corrections and they choose the index set Q to consist of indices corresponding to the largest relative changes of $G^{(j)}$ and $G^{(k)}$.

In the area of robust regression, the idea of updates and downdates has been used, for example, by Antoch and Ekblom (1995), Ekblom (1988), O'Leary (1990), and Wolke (1992). Complete downdates based on the factorization of A^TA have been suggested (Wolke, 1992). In this paper we will suggest partial updates and downdates based on the factorization of A^TA and A^THA. The purpose is not to replace the factorizations, but to reduce the number of factorizations needed.

3.1 Newton Method Approach

3.1.1 Huber and Talwar Functions

Newton's method and the IRLS method lead to the same sequence of problems when the weighting function is the Talwar function. To see this, note that $\rho''(z)/z$ for the Talwar function. So we will only consider the Talwar function under the Newton method approach.

Theorem 3.2 *Let $G, H \in \Re^{m \times m}$ be diagonal matrices with 0 or 1 on the diagonal. Define*

$$\Im_1 = \{j: 1 \le m, H_{jj} = 1 \text{ and } G_{jj} = 1\}$$

$$\Im_2 = \{j: 1 \le m, H_{jj} = 1 \text{ and } G_{jj} = 1\}$$

$$\Im_3 = \{j: 1 \le m, H_{jj} = 1 \text{ and } G_{jj} = 1\}$$

Let $A^T = \left[A_1^T, A_2^T, A_3^T\right]$ be the matrix whose columns have been permuted so that A_1, A_2, and A_3 are block row partitions corresponding to the index sets $\Im_1, \Im_2$ and $\Im_3$ respectively. Let $Q_2 \subseteq \Im_2$, $Q_3 \subseteq \Im_3$ and $Q = Q_2 \cup Q_3$ and $K \in \Re^{q \times n}$ be defined in (13). Assume that A^TKA is nonsingular. Let $M = (A^TKA)^{-1} A^TGA$. Then

$$\lambda(M) \le 1 + \sum_{j \in \Im_2 \setminus Q_2} \left\| A_{j.} \right\|_2^2 / \lambda_{\min} \left(A_1^T A_1 + \sum_{j \in Q_2} A_{j.}^T A_{j.} + \sum_{j \in \Im_s \setminus Q_s} A_{j.}^T A_{j.} \right)$$

If A^TGA is nonsingular, then

$$\lambda(M) \le \left(1 + \sum_{j \in \Im_s \setminus Q_s} \left\| A_{j.} \right\|_2^2 / \lambda_{\min} (A_1^T A_1 + A_2^T A_2) \right)^{-1}$$

Proof: Observe that

$$A^TKA = A_1^T A_1 + \sum_{j \in Q_2} A_{J.}^T A_{j.} + \sum_{j \in \Im_s \setminus Q_s} A_{J.}^T A_{j.}$$

and

$$A^TGA = A_1^T A_1 + A_2^T A_2 = A_1^T A_1 + \sum_{j \in \Im_s \setminus Q_s} A_{J.}^T A_{j.} - \sum_{j \in \Im_s \setminus Q_3} A_{J.}^T A_{j.}$$

Thus

$$A^TGA = A^TKA + \sum_{j \in \Im_s \setminus Q_s} A_{J.}^T A_{j.} - \sum_{j \in \Im_s \setminus Q_3} A_{J.}^T A_{j.} \tag{15}$$

If $(A^TKA)^{-1}$ exists then

$$\lambda_i((A^TKA)^{-1}A^TGA) \leq 1 + \sum_{j \in \Im_s \setminus Q_s} \| A_{j.} \|_2^2 \, \| (A^TKA)^{-1} \|_2 \tag{16}$$

and if $(A^TGA)^{-1}$ exists then

$$\lambda_i((A^TGA)^{-1}A^TKA) \leq 1 + \sum_{j \in \Im_s \setminus Q_s} \| A_{j.} \|_2^2 \, \| (A^TGA)^{-1} \|_2 \tag{17}$$

Observe that $\lambda_{min}(X) = 1/\| X^{-1} \|_2$ for any symmetric positive definite matrix X. Then rest of the proof follows from (3).

The following observations based on Theorem 3.2 are in order:

(i) If $Q = \Im_2$ then

$$\left(1 + \sum_{j \in \Im_s \setminus Q} \| A_{j.} \|_2^2 / \lambda_{\min}(A_1^T A_1 + A_2^T A_2)\right)^{-1} \leq \lambda_i((A^TKA)^{-1}A^TGA) \leq 1.$$

(ii) If $Q = \Im_3$ then

$$\leq \lambda_i((A^TKA)^{-1}A^TGA) \leq 1 + \sum_{j \in \Im_s} \left\| A_{j.} \right\|_2^2 / \lambda_{\min}\left(A_1^T A_1 + \sum_{j \in Q_s} A_{J.}^T A_{j.}\right)$$

(iii) If $\Im_2 = \varnothing$ then

$$\left(1 + \sum_{j \in \Im_s \setminus Q} \left\| A_{j.} \right\|_2^2 / \lambda_{\min}(A_1^T A_1)\right)^{-1} \leq \lambda_i((A^T K A)^{-1} A^T G A) \leq 1. \qquad (18)$$

Equation (18) corresponds to the case when we downdate $A^T A$.

For the case of Newton method with the Huber and Talwar functions, the diagonal matrix G in (6) has either 0's or 1's as diagonal elements. Thus we can use Theorem 3.2 to construct a preconditioner $A^T K A$, where $K \in \Re^{m \times m}$ is a diagonal matrix with 0 or 1 on the diagonal. The index set $Q = Q_2 \cup Q_3$. The sets Q_2 and Q_3 correspond to largest $||A_{j.}||_2$ for $j \in \Im_2$ and $j \in \Im_3$ respectively.

The Huber and Talwar functions may lead to singular linear systems when Newton method is used. Thus instead of solving a linear system of the form (12) we solve

$$(A^T G A + \delta I)\Delta x = h,$$

where $\delta \geq 0$ is the stabilizing factor. When $A^T G A$ is singular $A^T K A$ is likely to be singular too. Thus let $(A^T K A + \delta I)$ be the preconditioner. Using (15) we can write

$$(A^T G A + \delta I) = (A^T K A + \delta I) + \sum_{j \in \Im_2 \setminus Q_2} A_{j.}^T A_{j.} - \sum_{j \in \Im_s \setminus Q_s} A_{j.}^T A_{j.}$$

Let M denote the preconditioned matrix $(A^T K A + \delta I)^{-1}(A^T G A + \delta I)$. $(A^T K A + \delta I)^{-1}(A^T G A + \delta I)$. Then using (16) we have

$$\lambda_i(M) \leq 1 + \sum_{j \in \Im_2 \setminus Q_2} ||A_{j.}||_2^2 / \delta .$$

Using (17) followed by (3) we get

$$\lambda_i(M) \geq \left(1 + \sum_{j \in \Im_s \setminus Q} ||A_{j.}||_2^2 / \delta\right)^{-1} .$$

In what follows we will write $r(x)$ as r to simplify the notation.

3.1.2 Fair and Logistic Functions

Theorem 3.3 *Let* $r \in \Re^m$. *Let* $G \in \Re^{m\times m}$ *be a positive definite diagonal matrix defined by*

$$G_{jj} = \begin{cases} (1 + (|r_j|/\delta\beta))^{-2} & \text{Fair} \\ 1 - \tanh^2(r_j/\delta\beta) & \text{Logistic} \end{cases}.$$

Let $H = I, Q \subseteq \{j: 1 \le j \le m \text{ and } H_{jj} \ne G_{jj}\}$, *and* $\mathrm{K} \in \Re^{m\times m}$ *be defined as in (13). Then*

$$\text{Fair: } \left(1 + (1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right)^{-2} \le \lambda_i((A^TKA)^{-1}A^TGA) \le 1$$

and

$$\text{Logistic: } 1 + \tanh^2\left((1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right) \le \lambda_i((A^TKA)^{-1}A^TGA) \le 1$$

$$\text{Logistic: } 1 + \tanh^2\left((1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right) \le \lambda_i((A^TKA)^{-1}A^TGA) \le 1.$$

Proof: For the Fair function we have $G_{jj} = (1 + |r_j|/(\sigma\beta))^{-2}$. $G_{jj} = (1 + |r_j|/(\sigma\beta))^{-2}$. Thus

$$\min_j\{G_{jj}\} \ge (1 + (||r||_\infty/\sigma\beta))^{-2} \text{and } \max_j\{G_{jj}\} \le 1$$

$\min_j\{G_{jj}\} \ge (1 + (||r||_\infty/\sigma\beta))^{-2} \text{and } \max_j\{G_{jj}\} \le 1.$

Applying Theorem 3.1 we have

$$(1 + (||r||_\infty/\sigma\beta))^{-2} \le \lambda_i((A^TA)^{-1}A^TGA) \le 1\,. \tag{19}$$

Now note that $G_{jj}/K_{jj} = 1$ for $j \in Q$ and $G_{jj}/K_{jj} = G_{jj}/H_{jj} = G_{jj}$ for $j \notin Q$. Then repeating the same technique as for (19) the bounds on $\lambda_i((A^TKA)^{-1}A^TGA)$ are

$$\min_j\left\{G_{jj}/K_{jj}\right\} = \left(1 + (1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right)^{-2} \text{and } \max_j\{G_{jj}/K_j\} \le 1.$$

For the Logistic function $G_{ii} = 1 - \tanh^2(r_i/\sigma\beta)$. Hence $\max_i\{G_{ii}\} \le 1$ and

$$\min_j\{G_{jj}\} \ge 1 - \tanh^2\left(\max_j\{|r_j|\}/\sigma\beta\right) = 1 - \tanh^2(||r||_\infty/\sigma\beta).$$

Applying Theorem 3.1 we get

$$1 - \tanh^2(||r||_\infty/\sigma\beta) \le \lambda_i((A^THA)^{-1}A^TGA) \le 1 \tag{20}$$

Again notice that $G_{jj}/K_{jj} = 1$ for $j \in Q$ and $G_{jj}/K_{jj} = G_{jj}/H_{jj} = G_{jj}$ for $j \notin Q$.

Repeating the arguments for (20) the upper bound on $\lambda_i((A^TKA)^{-1}A^TGA)$ is $\max_i \{G_{ii}/K_{ii}\} \leq 1$and the lower bound is

$$\min_j\{G_{jj}/K_{jj}\} \geq 1 - \tanh^2\left((1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right). \quad \blacksquare$$

Theorem 3.3 suggests that we should choose the index set Q to consist of indices corresponding to the largest components of $|r|$.

3.2 IRLS Method Approach

3.2.1 L_p Approach

Let $r, \hat{r} \in \Re^m$. Let $G, H \in \Re^{m\times m}$ be positive definite diagonal matrices defined by $H_{jj} = |\hat{r}_j|^{p-2}$ and $G_{jj} = |r_j|^{p-2}(1 \leq p < 2)$. For a given Q, let $K \in \Re^{m\times m}$ be defined as in (13). Then

$$\lambda_i((A^TKA)^{-1}A^TGA) \geq \min\left\{1, \min_{j\notin Q}\left\{\frac{|\hat{r}_j|}{|r_j|}\right\}^{2-p}\right\} \tag{21}$$

and

$$\lambda_i((A^TKA)^{-1}A^TGA) \leq \max\left\{1, \max_{j\notin Q}\left\{\frac{|\hat{r}_j|}{|r_j|}\right\}^{2-p}\right\}. \tag{22}$$

Thus, to construct the preconditioner for the coefficient matrix in linear system (11), the results in (21) and (22) imply that we should consider indices corresponding to largest $|\hat{r}_j|/|r_j| > 1$ and/ or smallest$|\hat{r}_j|/|r_j| < 1$ such that $k(K^{-1}G)$ is minimized.

In the rest of this section we construct preconditioners based on modification of an existing factorization $A^TA = LL^T$

3.2.2 General Preconditioner Based on Huber, Fair, and Logistic Weighting Functions

Theorem 3.4 *Let $r \in \Re^m$ and $G \in \Re^{m\times m}$ be a positive definite diagonal matrix defined by*

$$G_{jj} = \begin{cases} 1 \quad if\ |r_j| \le \sigma\beta \\ \sigma\beta/|r_j| if\ |r_j| > \sigma\beta \\ (1+(|r_j|/\sigma\beta))^{-1} & \text{Fair function} \\ (r_j/\sigma\beta)^{-1}\tanh(r_j/\sigma\beta) & \text{Logistic function} \end{cases} \tag{23}$$

Then

$$(1+(||r||_\infty/\sigma\beta))^{-1} \le \lambda_i((A^TA)^{-1}A^TGA) \le 1$$
$$(1+(||r||_\infty/\sigma\beta))^{-1} \le \lambda_i((A^TA)^{-1}A^TGA) \le 1. \tag{24}$$

Furthermore, let $H = I$, $Q \subseteq \{j : 1 \le j \le m \text{ and } H_{jj} \ne G_{jj}\}$ and $K \in \Re^{m\times m}$ $K \in \Re^{m\times m}$*be defined as in (13). Then*

$$\left(1+(1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right)^{-1} \le \lambda_i((A^TKA)^{-1}A^TGA) \le 1. \tag{25}$$

Proof: The proof is in two parts. In the first part we give the proof of (24) based on each weighting function. In the second part we prove (25) based on (24) using the same technique as in the proof of Theorem 3.3.

Consider the Huber weighting function (23).

Then $\max_j\{G_{jj}\} \le 1$ and

$$\min_j\{G_{jj}\} \ge \min\{1, \sigma\beta/||r||_\infty\}$$
$$= \frac{1}{\max\{1, ||r||_\infty/\sigma\beta\}} \ge (1+(||r||_\infty/\sigma\beta))^{-1}.$$

Using Theorem 3.1 we have

$$(1+(||r||_\infty/\sigma\beta))^{-1} \le \lambda_i((A^TA)^{-1}A^TGA) \le 1.$$

Secondly, consider the Fair weighting function $G_{jj} = (1+|r_j|/\sigma\beta)^{-1}$.

Then $\max_i\{G_{ii}\} \le 1$ and
$\min_j\{G_{jj}\} = (1+(||r||_\infty/\sigma\beta))^{-1}$

and (24) follows for the Fair weighting from Theorem 3.1.

For the Logistic function, we first note that for any $x \in \Re$, (see Rottmann, 1991: pp. 86)

$$1/(1+|x|) \le \tanh(x)/x \le 1. \tag{26}$$

From (23) we have $G_{ii} = (r_i/\sigma\beta)^{-1}\tanh(r_i/\sigma\beta)$. Thus using (26) we get

$$\min_j\{G_{jj}\} \ge 1/\left(1+\max_j|r_j|/\sigma\beta\right) = (1+(||r||_\infty/\sigma\beta))^{-1}$$

and $\max_j\{G_{jj}\} \leq 1$. Hence (24) follows from Theorem 3.1.
Theorem 3.1 states that
$\min_j\{G_{jj}/K_{jj}\} \leq \lambda_i((A^TKA)^{-1}A^TGA) \leq \max_j\{G_{jj}/K_{jj}\}$.

Next observe that $G_{jj}/K_{jj} = 1$ for $j \in Q$ and $G_{jj}/K_{jj} = G_{jj}/H_{jj} = G_{jj}$ $G_{jj}/K_{jj} = G_{jj}/H_{jj} = G_{jj}$ for $j \notin Q$. Hence (24) implies that $\max_j\{G_{jj}/K_{jj}\} \leq 1$ $\max_i\{G_{ii}/K_{ii}\} \leq 1$ and

$$\min_j\{G_{jj}/K_{jj}\} \geq \left(1 + (1/\sigma\beta)\max_{j\notin Q}\{|r_j|\}\right)^{-1}. \quad \blacksquare$$

Theorem 3.4 suggests that we should choose the index set Q to consist of indices j corresponding to the largest components $|r_j|$.

3.3 Computing with Preconditioners

We consider three cases:

1. Constant preconditioner A^TA: We compute the factorization $LL^T = A^TA$ once at the beginning and store the factor L for use throughout the whole of the iterative process.

   We note that the theoretical upper bounds for $k((A^TKA)^{-1}A^TGA)$ $k((A^TKA)^{-1}A^TGA)$ can be deduced from equation (18), Theorem 3.3, and Theorem 3.4.

2. Downdating A^TA: The results of equation (18), Theorem 3.3, and Theorem 3.4 suggests modifying A^TA at every iteration. Thus the linear systems involving the preconditioner as coefficient matrix (at every iteration) are of the form

   $$(A^TA + \bar{A}^T\bar{D}\bar{A})s = (LL^T + \bar{A}^T\bar{D}\bar{A})s = v, \tag{27}$$

   Where $\bar{A} \in \Re^{qxn}$ consists of rows of A corresponding to the index set Q. Thus at every iteration we carry out low-rank downdates to A^TA. Similarly, we compute the factorization $LL^T = A^TA$ once at the beginning and store the factor L for use throughout the whole of the iterative process. We suggest solving (27) by the Sherman–Morrison– Woodbury formula approach (Baryamureeba and Steihaug, 2000; Baryamureeba *et al.*, 1999).

3. Low-rank corrections to A^THA: Here we consider the case where $LL^T = A^TA$ is a matrix from the previous iteration with a known factorization. The low-rank correction is determined by the choice of the

index set Q. We choose Q based on (14) when H and G are both positive definite diagonal matrices or based on Theorem 3.2 for the case when H and G are semi-definite diagonal matrices. The low rank correction matrix corresponding to Q is $\bar{A}^T \bar{D} \bar{A}$. Thus the linear systems involving the preconditioner as coefficient matrix are of the form

$$(A^T H A + \bar{A}^T \bar{D} \bar{A})s = (LL^T + \bar{A}^T \bar{D} \bar{A})s = v, \qquad (28)$$

We suggest solving (28) by the Sherman–Morrison–Woodbury formula approach (Baryamureeba and Steihaug, 2000; Baryamureeba *et al.,* 1999) when we alternate between a direct solver and an iterative solver at every iteration, and by the approach based on modifying the factors (Baryamureeba and Steihaug, 2000) when we choose the solver adaptively (Baryamureeba, 2000; Wang and O'Leary,1995).

The factorization can be obtained directly by computing the Cholesky factorization $LL^T = A^T H A$ or from the QR factorization of $H^{1/2}A$ for ill–conditioned problems.

We remark that using $A^T A$ as preconditioner in the early iterative stage and $A^T K A$ (a low-rank correction of $A^T H A$) during the late iterative stage may be computationally efficient.

4. Implementation Details

Full Inexact Newton steps (i.e. the choice $\propto^{(k)} = 1$), leads to superlinear convergence if $\eta^{(k)} \to 0$ close to the solution. However, the Inexact Newton steps may not decrease the value of the objective function (2) away from the solution. Thus there is need to decrease the step length gradually until f is sufficiently reduced. The step-length $\propto^{(k)} = (\frac{1}{2})^i$ where $i \geq 0$ is the smallest power so that $f(x^{(k)} + \propto^{(k)} \Delta x^{(k)}) \leq f(x^{(k)}) + 0.001 \propto^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)}$.

The implementation is done in MATLAB[1] and we use the Matlab function symmmd to compute the symmetric multiple minimum degree ordering for matrix $A^T A$ for the sparse Cholesky solver chol. The iterative method is preconditioned conjugate gradient least squares (PCGLS) method as implemented in Baryamureeba and Steihaug (1999) using the Sherman–Morrison–Woodbury formula. The termination criteria in PCGLS routine are when either the number of PCGLS iterations has reached a prescribed integer t or when $||q^{(k)}||_2 / h^{(k)}||_2 \leq \delta_{pcgls}$, where $q^{(k)} = h^{(k)} - A^T G^{(k)} A \Delta x^{(k)}$ for $h^{(k)}$ defined in (12). The weight matrix $G^{(k)}$ for Newton's method is (7) and (8) for the IRLS method.

We terminate the algorithm when $||\nabla f(x^{(k+1)})||_2 \leq \epsilon_f$ and $||x^{(k+1)} - x^{(k)}||_2 \leq \epsilon_x$ or when we exceed the maximum number of iterations allowed. For the starting value we use the least squares solution (Ekblom, 1988; Scales *et al.,* 1988)

$x^{(0)} = (A^T A)^{-1} A^T y.$

In numerical experiments we set $\epsilon_f = 10^{-3}$, $\epsilon_x = 10^{-3}$ and $\delta_{pcgls} = 10^{-5}$. The maximum number of iterations allowed is set to 100 iterations and the maximum number of PCGLS iterations is $t = 40$ in the numerical experiments.

5. Numerical Experiments

5.1 Distribution of Eigenvalues of Preconditioned Matrix

We extract the matrix A from the netlib set (Gay, 1985) of linear programming problems. The true solution vector x is taken to be a vector of all ones, and the right hand side is chosen to be $y = Ax + \sigma N(0,1)$ except the outliers are obtained by adding $100\sigma N(0,1)$ to $m_1 m_1$ (is number of outliers) randomly chosen elements of y. In all experiments we set $m_1 = 10$, $\beta = 1$, and $\sigma = 0.1$. The percentage of high leverage points or "wild points" is below 1% in all test problems.

The observed behavior seems to be typical of many experiments we conducted on different test problems at different stages of Newton (or the IRLS) algorithm. In the figures H is the weight matrix at the previous iteration and G is the weight matrix at the current iteration. For positive definite H and G we choose the index set Q such that the spectral condition number $k(K^1 G)$ is minimized. In all the figures $|Q| = 20$.

Fig 1: **The IRLS method with the Huber weighting function. The index set Q is chosen according to Theorem 3.4. Test problem is blend $(m = 114, n = 74)$**

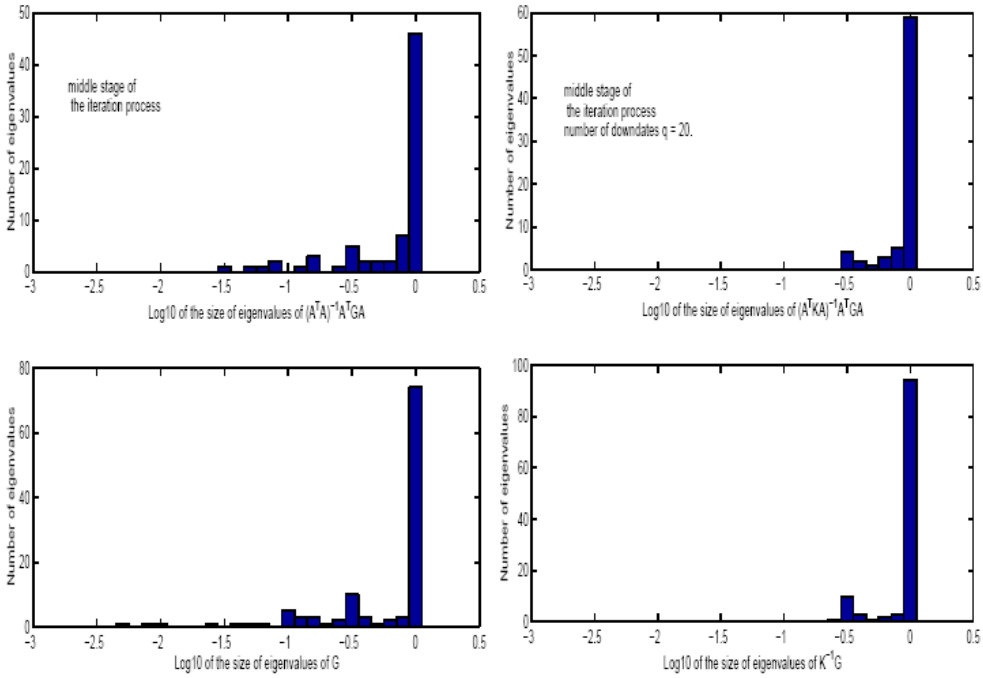


Fig 2: **Newton's method with the Fair weighting function. The index set Q is similar to Theorem 3.3.**

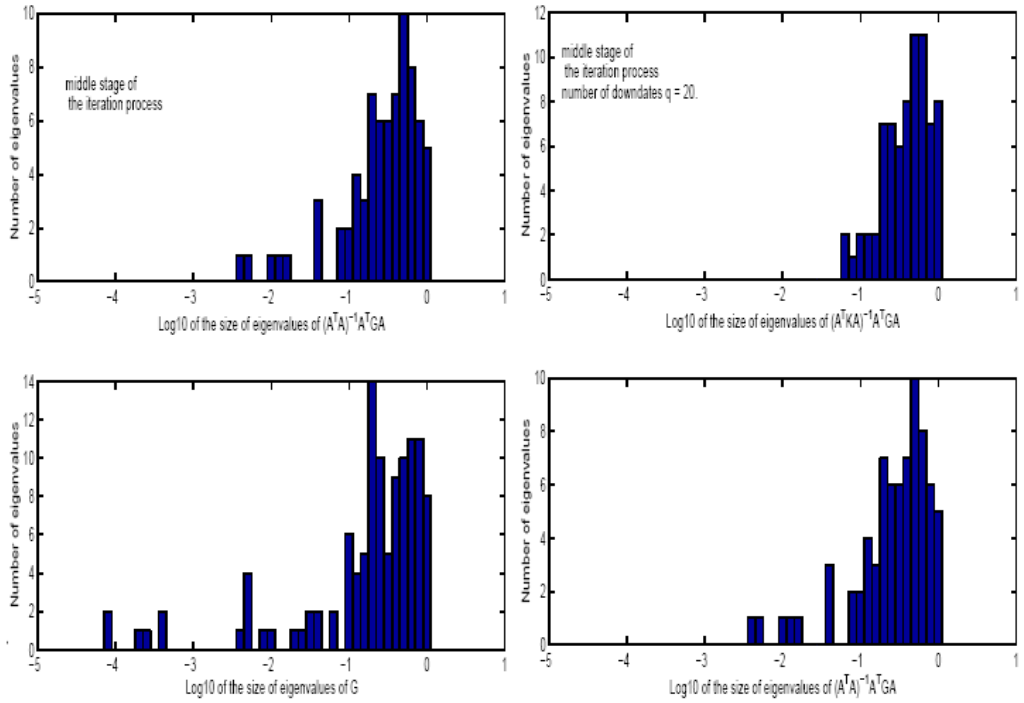


Figure 1 and 2 show the distribution of the eigenvalues of the preconditioned matrix $(A^TA)^{-1}A^TAGA$ are well described by the distribution of $\{G_{jj}: j = 1, \ldots, m\}$; $\min_j\{G_{jj}\}$ is a lower bound of $\lambda_{min}((A^TA)^{-1}(A^TAGA)$; and that introducing low-rank downdates may decrease the spectral condition number $k((A^TKA)^{-1}(A^TAGA)$. Figure 1 and 2 support the choice of the index set based on the largest $|r_j|$ (which corresponds to the smallest G_{jj}).

Corresponding results for the Fair and Logistic functions with the IRLS method lead to similar figures as in Figure 1. The Logistic function leads to similar results as Figure 2 if the index set Q is chosen according to Theorem 3.3.

Figure 3: Newton's method with the Fair weighting function. The index set Q is chosen according to (14). We choose Q so that $k(K^{-1}G)$ is minimized.

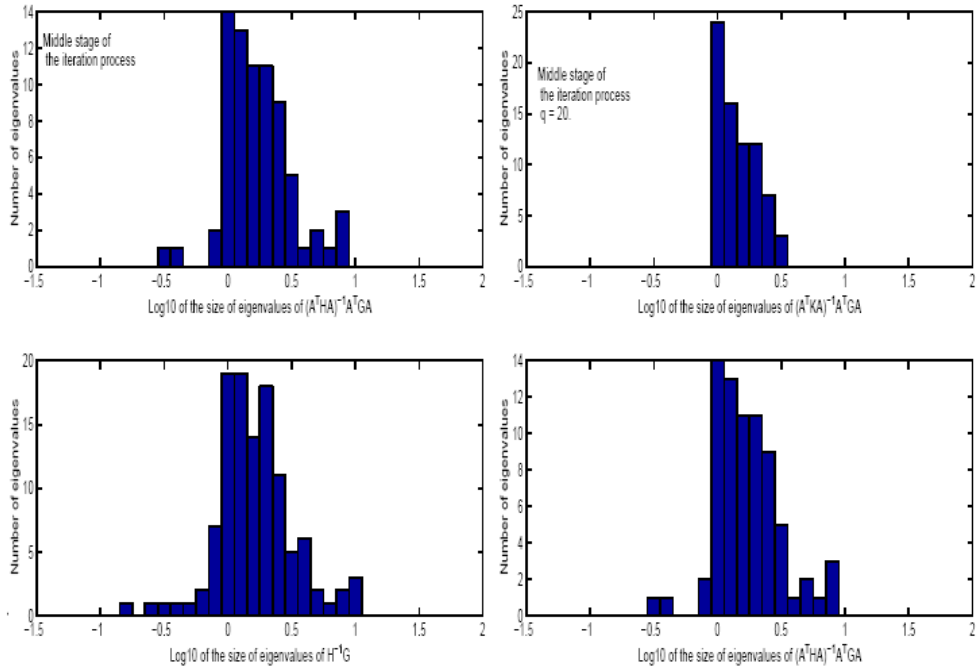


Figure 3 shows that we can form efficient low-rank correction preconditioners based on modifying A^THA (a matrix from the previous iteration with a known factorization).

5.2 Performance of the Preconditioners on Sequences of Linear Systems

In numerical tests we will use the following testmatrices:

stocfor2$(m = 3045, n = 2157)$;

maros $(m = 921, n = 835;$ sccd6 $(m = 1350, n = 147);$ and scsd8 $(m = 2750, n = 397)$ We only give numerical results based on the preconditioners A^TA and its corresponding low-rank downdate A^TKA. In the tables M denotes the preconditioner. Outer refers to inexact Newton or the IRLS iterations and inner to total number of PCGLS iterations. The maximum number of downdates is set to $|Q| \leq 20$ for all the results in Table 2, 3, 4, and 5. The actual number of downdates may be less than the maximum, for instance, for Huber and Talwar functions if $|\mathfrak{T}_2|+|\mathfrak{T}_3|$ is less than the maximum in Theorem 3.2.

We will base our analysis on the values of outer and inner. The results in Table 2 for the Talwar function show that we terminate by the maximum number of PCGLS iterations allowed, $t = 40$, at all iterations, for stocfor2 when A^TA and A^TKA are used as preconditioners, and for maros when A^TA is used as a preconditioner. This suggests that we should increase q or t so that the linear system can be solved to high accuracy. The results in Table 2 and 3 suggest that the Fair function is performing better than other functions on inexact Newton method. The results in Table 4 and 5 show that on average the Fair function is doing better than other functions on the IRLS method. The Logistic function is doing almost as well as the Fair function on the IRLS method. Comparing the results in Table 2 and 3 with the results in Table 4 and 5 we see that Newton's method converges faster than the IRLS method, and the low-rank downdates do not lead to a significant decrease in the inexact Newton (IRLS) iterations carried out (outer). Thus it is worthwhile to use A^TA instead of its low-rank downdate A^TKA as preconditioner.

Table 2: Inexact Newton method. A^TKA is a low-rank downdate of A^TA.

Problem	Huber function				Talwar function			
	M	Outer	Inner	$\|\|r\|\|$	M	Outer	Inner	$\|\|r\|\|$
stocfor2	A^TA	28	1079	6.70	A^TA	100	4000	7.54
	A^TKA	39	1452	6.70	A^TKA	100	4000	36.2
macros	A^TA	28	1031	14.6	A^TA	100	4000	14.7
	A^TKA	28	1019	14.6	A^TKA	98	3960	14.5
scsd6	A^TA	12	258	17.2	A^TA	16	328	16.4
	A^TKA	14	312	17.2	A^TKA	12	328	16.4
scsd8	A^TA	18	454	19.7	A^TA	11	354	18.0
	A^TKA	16	415	19.7	A^TKA	11	355	18.0

Table 3: Inexact Newton method. A^TKA is a low-rank downdate of A^TA.

Problem	Fair function				Logistic function			
	M	Outer	Inner	$\|\|r\|\|$	M	Outer	Inner	$\|\|r\|\|$
stocfor2	A^TA	13	462	6.52	A^TA	24	849	6.67
	A^TKA	11	392	6.52	A^TKA	17	559	6.67
macros	A^TA	9	275	14.6	A^TA	19	667	14.6
	A^TKA	8	251	14.6	A^TKA	21	728	14.6
scsd6	A^TA	8	123	17.2	A^TA	13	199	17.2
	A^TKA	8	115	17.2	A^TKA	10	143	17.2
scsd8	A^TA	9	133	19.7	A^TA	15	333	19.7
	A^TKA	9	129	19.7	A^TKA	10	239	19.7

Table 4: Inexact IRLS method. A^TKA is a low-rank downdate of A^TA. The results for the Talwar function are the same as in Table 2.

Problem	Huber function			
	M	Outer	Inner	$\|\|r\|\|$
stocfor2	A^TA A^TKA	76 75	3064 2982	6.70 6.70
macros	A^TA A^TKA	76 76	845 826	14.6 14.6
scsd6	A^TA A^TKA	11 11	59 54	17.2 17.2
scsd8	A^TA A^TKA	20 20	81 76	19.7 19.7

Table 5: Inexact IRLS method. A^TKA is a low-rank downdate of A^TA

Problem	Fair function					Logistic function			
	M	Outer	Inner	$\|\|r\|\|$		M	Outer	Inner	$\|\|r\|\|$
stocfor2	A^TA A^TKA	49 48	1987 1914	6.52 6.52		A^TA A^TKA	54 53	2178 2097	6.67 6.67
macros	A^TA A^TKA	28 28	394 374	14.6 14.6		A^TA A^TKA	24 24	303 281	14.6 14.6
scsd6	A^TA A^TKA	10 10	57 53	17.2 17.2		A^TA A^TKA	9 9	53 48	17.2 17.2
scsd8	A^TA A^TKA	14 14	72 67	19.7 19.7		A^TA A^TKA	15 15	69 64	19.7 19.7

6. Conclusion

The paper suggests preconditioners for both Newton and IRLS methods. And the theoretical results give bounds on the spectral condition numbers of the preconditioned matrices. The numerical results do demonstrate the theoretical

results. The numerical results show that the suggested preconditioners are promising and merit further study, especially in the presence of high leverage points or a high percentage of "wild points". The paper opens new ways to solve large, sparse regression problems using robust alternatives to least squares criterion.

References

Antoch, J., and Ekblom, H. (1995). Recursive Robust Regression Computation Aspects and Comparison, Computational Statistics and Data Analysis, 19: 115-128.

Baryamureeba, V. (2000). Solution of Large-Scale Weighted Least Squares Problems, Technical Report 186, Department of Informatics, University of Bergen, Bergen, Norway.

Baryamureeba, V., and Steihaug, T. (1999). On a Class of Preconditioners for Interior Point Methods, Proceedings of the 6th Meeting of the Nordic Section of the Mathematical Programming Society, Opuscula 49, ISSN 1400–5468.

Baryamureeba, V. and Steihaug, T. (2000). Computational Issues for a New Class of Preconditioners, Large-Scale Scientific Computations of Engineering and Environmental Problems II, Series Notes on Numerical Fluid Mechanics, VIEWEG, 73: 128–135.

Baryamureeba, V., Steihaug, T., and Zhang, Y.,(1999) Properties of a Class of Preconditioners for Weighted Least Squares Problems, Technical Report 170, Department of Informatics, University of Bergen, Bergen, Norway.

Chatterjee, S. and M¨achler, M. (1997). Robust Regression: A Weighted Least Squares Approach, Communications in Statistics, Theory and Methods, 26: 1381-1394.

Coleman, D., Holland, P., Kaden, N., Klema, V., and Peters, S.C. (1980). A System of Subroutines for Iteratively Reweighted Least Squares Computations, ACM Transactions on Mathematical Software, 6: 327-336.

Dembo, R.S., and Steihaug, T. (1983). Truncated Newton Algorithms for Large-Scale Unconstrained Optimization, Mathematical Programming, 26: 190-212.

Dembo, R.S., Eisenstat, S.C., and Steihaug, T. (1982). Inexact Newton methods, SIAM Journal of Numerical Analysis, 19: 400-407.

Ekblom, H. (1988). A New Algorithm for the Huber Estimator in Linear Models, BIT, 28: 123-132.

Farquharson, C.G., and Oldenburg, D.W. (1998). Non Linear Inversion Using General Measures of Data Misfit and Model Structure, Geophysical Journal International: 213-227.

Gay, D.M. (1985). Electronic Mail Distribution of Linear Programming Test Problems, Mathematical Programming Society, COAL Newsletter, No. 13: 10-12.

Goldfarb, D., and Liu, S. (1991). An o (n3L) Primal Interior-Point Algorithm for Convex Quadratic Programming, Mathematical Programming, 49: 325–340.

Holland, P.W., and Welsch, R.E. (1977). Robust Regression Using Iteratively Reweighted Least Squares, Communications in Statistics: Theory and Methods, A6: 813-827.

Huber, P.J. (1981). Robust Statistics, J. Wiley, New York, NY.

Karmarkar, N. (1984) A New Polynomial Time Algorithm for Linear Programming, Combinatorica, 4: 373-395.

Karmarkar, N.K., and Ramakrishnan, K.G. (1991).Computational Results of an Interior Point Algorithm for Large-Scale Linear Programming, Mathematical Programming, 52: 555-586,

O'Leary, D.P. (1990). Robust Regression Computation Using Iteratively Reweighted Least Squares, SIAM Journal of Matrix Analysis and Applications, 11: 466-480.

Rottmann K. (1991) Mathematische formelsammlung, 4. Auflage, B.I.-Hochschultaschenbuch, Band 13, ISBN 3-411-70134-X (in German).

Scales, J.A., Gersztenkorn, A., and Treitel, S. (1988). Fast lp Solution of Large, Sparse, Linear Systems: Application to Seismic Travel Time Tomography, Journal of Computation Physics, 75: 314-333.

Shanno, D.F., and Rocke, D.M. (1986) .Numerical Methods for Robust Regression: Linear Models, SIAM Journal on Scientific and Statistical Computing, 7: 86-97.

Wang, W., and O'Leary, D.P. (1995). Adaptive Use of Iterative Methods in Interior Point Methods for Linear Programming, Report CS-TR-3560, Computer Science Department, Institute for Advanced Computer Studies, and Report UMIACS-95-111, University of Maryland.

Wolke, R. (1992). Iteratively Reweighted Least Squares: A Comparison of Several Single Step Algorithms for Linear Models, BIT, 32: 506-524.

6
Computational Analysis of Kinyarwanda Morphology: The Morphological Alternations.

Jackson Muhirwe

For more than 30 years, there have been renewed interests in computational morphology resulting in numerous morphological tools. However the interest has always been on the politically and economically interesting languages of the world resulting in a wide language divide between the technologically rich and poor languages. Kinyarwanda language, a Bantu language spoken in East Africa is one of those under-resourced languages without any language technology tools. The two most essential components of most natural language applications are a morphological analyzer and a machine-readable lexicon. These two components are still lacking for Kinyarwanda and so many other under-resourced languages. The task of developing a morphological analyzer involves two problems: the morphotactics (word formation) and the morphological alternations. In this paper we mainly concerned with the morphological alternations.

Introduction

The broad area of computational morphology is concerned with computational analysis and synthesis of words for eventual use in natural language applications. The ultimate goal of computational morphology is incorporating its products, the morphological analyzers into higher level natural language applications. Such applications may include spell checkers and information retrieval systems. Currently there are two broad approaches to computation morphology: rule based and data based approaches. Our main focus in this paper are the rule based methods. Finite state methods have dominated computational morphology research since Johnson's ground breaking 1972 discovery that under certain constraints, phonological rules could be modeled using finite state methods (Johnson, 1972). Although Johnson (1972) was the first to realise that finite state machines could be used to model linguistics structures, the approach was never recognised until the late 1970s when Kay independently discovered that finite state machines can simplify the modelling of phonology and morphology. The invention of the two level phonology approach in 1983 by Koskenniemi was a major breakthrough in Computational linguistics and this led to wide spread use of finite state machines. Since the invention of the two level approach in 1983, finite state methods have successfully been employed to develop morphological analysers for many languages

including English, French, Germany, Arabic, Spanish, Basque, Japanese, Korean and Swahili a Bantu language (Beesley and Karttunen, 2003). Computational morphological tools have mainly been built for the politically and economically interesting languages of Europe and some parts of Asia. The status quo according to Cole (1997) is that morphological analysers should be built for all but the commercially important languages. To emphasize this point, Dhonnchadha et al.(2003) reiterates the need to develop morphological analyser for all languages to avoid creating a language divide. We now have the technologically rich languages and poor languages.

There are two main challenges involved in developing morphological analysers: the morphatactics and the morphological alternations. The morphotactics or sometimes known as the morphosyntax, is concerned with the strict rules that dictate how morphemes are combined together to form words. Morphemes are the smallest indivisible components of a word with a purpose or a meaning. Morphological alternations are concerned with the phonological / orthographical rules that are required to derive the surface representation of a word from its underlying representation. The problem of morphological alternations arises because sometimes a morpheme may have different realisations depending on its phonological environment and other morphemes that make up the word. Our concern here is on this later problem leaving the morphological alternations problem aside.

We focus on Kinyarwanda, a Bantu language, with barely any language tool developed, to be exploited by its 20 million speakers. In this paper we focus on the phonological / orthographical rules that are required to derive the surface form of a word from its underlying /lexical level form. The rest of the paper is organised as follows: The next section presents an overview of Kinyarwanda morphology. This is followed by a review of computational morphology, which is followed by an implementation of the morphological alternations using Xerox finite state tools and finally a discussion of the results and conclusion ensues.

Kinyarwanda Morphology

Kinyarwanda language, the national language of Rwanda is a typical Bantu language classified by Guthrie (1971) as a group D60 Bantu language (D66), together with Kirundi, Giha, Vinza, Hangaza and Shubi. The language is spoken by over 20 Million people, living mainly in the great lakes region of East and Central Africa. Its speakers include, Barundi (from Burundi), Giha, Bafumbira, Banyamulenge and the ethnic Banyarwanda in Masisi and Rutshuro in Northern Kivu of the Democratic Republic of Congo.

Kinyarwanda language major word categories are the nouns and verbs. Kinyarwanda nouns are composed of the preprefix, the prefix or sometimes called the class marker, the stem and the suffix. The noun morphology is mainly influenced by the noun classification system and the ensuing concodia agreement system. Kinyarwanda shares these properties with other Bantu languages. Kinyarwanda

verb morphology is known to be more complex than the nominal. Typically, a verb will have multiple prefixes and suffixes surrounding the stem just like beads on a string. Prefixes have only grammatical information while suffixes have both grammatical and lexical information. The only obligatory morphemes in a verb are the subject agreement prefix, the stem and the final vowel which in most cases is the aspect marker. The optional morphemes include the proclitics *-nti* 'not', *ni* 'if/when'; the tense aspect-modality morphemes; the morphemes *na* 'also'; the object pronouns which can be one or many; lexical verb extensions; grammatical suffixes; the enclitics *-ga;* and locative postclitics *-m´o, -h´o, or -y´o*. Kinyarwanda verbs, like most other Bantu language verbs, can have multiple object pronouns, multiple lexical verbal extensions and multiple grammatical suffixes. Lexical extensions such as *-agur-, -iir-, uur, -aang, iriz-*, etc. add lexical information such inchoativity, iterativity, repetitivity, intensity, frequentativity, reversivity, etc. Grammatical morphemes such as the causative morpheme *-iish-*, the applicative morpheme *-ir-*, the comitative/reciprocal morpheme *-an-*, can be added to any verb stem.

Computational Morphology

Computational morphology has been an active area of research for the last 25 years. Oflazer (1999) defined computational morphology as the study of the computational analysis and synthesis of words to be eventually used in natural language processing applications. Computational morphology is mainly concerned with systems that efficiently analyse and synthesize words. Although the beginning of the field of computational morphology could be traced in the 1950s, the most practically accepted approach was the two-level morphology introduced by Koskenniemi in 1983 (Koskenniemi, 1983a). The two-level morphology approach was immediately accepted by most researchers and since then has been the dominant formalism for dealing with computational morphology (Sproat 1992, 2006; Antworth, 1990). In this research we have chosen the two-level morphology approach, the dominant formalism in practical implementations of morphological analysers.

Two level Morphology

The two-level morphology approach to morphological analysis is a language independent general formalism for analysis and generation of word-forms. Kimmo invented this approach in 1983. The Generative phonology approach creates un-necessary intermediate levels and is also uni-directional. Kimmo decided to eliminate the intermediate levels. This created a new approach, which has only two levels, the lexical level and the surface level, hence the name Two-Level Morphology. This model has also an added advantage of being bi-directional, implying that both analysis and generation could be done using the same system, which was not possible with the earlier approaches which were uni-directional. Two-level morphology depends heavily on finite state methods, which are well known and are often described as elegant (Kartunnen and Bessley, 2003). Several

compilers have been developed to deal with two level rules, but in this paper we shall use Twolc developed by Xerox. The choice for two-level morphology approach was not accidental. The two level approach has already successfully been used to develop a comprehensive morphological analyser for Swahili, a Bantu language.

Two level rules

Two level rules are generally of the form

CP OP LC _ RC

Where

CP stands for Correspondence Part

OP stands for Operator

LC stands for Left Context

RC stands for Right Context

There are four different kinds of rules that may be used to describe morphological alternations of any language.

1. a:b = > LC _ RC. This rule states that lexical //a// can be realized as surface b ONLY in the given context. This rule is a context restriction rule
2. a:b < = LC _ RC This rule states that lexical //a// has to be realized as surface b ALWAYS in the given context. This rule is a surface coercion rule.
3. a:b □ LC _ RC this is a composite rule which states that lexical //a// is realized as surface be ALWAYS and ONLY in the given context.
4. a:b /< = LC _ RC This is an exclusion rule that states that lexical //a// is never realized as surface //b// in the given context.

These rules may be compiled into finite state acceptors either by hand or automatically using one of the many available two level rule compilers. For this paper we used the Xerox two level rule compiler Twolc.

The Xerox Finite State Tools

This is a set of powerful, sophisticated set of algorithms and programming languages for building finite state solutions to a variety of problems in natural language processing. For the purpose of this paper we shall only look at two tools, which were used in this research.

Lexc

This is the Lexicon Compile. Lexc is a highlevel declarative language used for specifying the required lexicon and morphotactic structure of the words in the

lexicon. The compiled lexicon results into a network of well-formed strings. Lexc source files are written using notepad, emacs and any other text editor.

Twolc

This is the two level compiler. This is a highlevel declarative language designed for specifying alternation rules required in morphological descriptions. Gross irregularities and all base forms are also included in the Lexc source file. Twolc source files are typically text files written using notepad, emacs or any other text editor.

The Xerox finite state technology is based on three fundamental insights

1. The morphotacts can be encoded using finite state networks
2. The alternation rules of each morpheme can be implemented as a finite state transducer.
3. The lexicon network and the rule transducer can be combined together by composition to form a single network called a lexical transducer

Lexical transducers constructed using the Xerox finite state technology are mathematically elegant, bi-directional and highly efficient. The applications have a potential for wide lexical coverage, use little memory space, being robust and commercially viable products.

Compiling the Morphological alternations using Twolc

The first thing required in the Twolc source file is the definition of the alphabet which is preceded by the keyword Alphabet.

```
Alphabet
a b c d e f g h i j k l m N:m N:n n o p F:pf q r s t S:ts u v x y w z ;
```

There may also be need for definitions of sets that may be required by the rules

```
Sets
V = a e i o u ;
C = b c d g h j k l m n p q r s t v w x y z ;
C2 = b c d g h j k l m n p q r s t v x y z ; !This subset doesnot have w
C3 = b c d g h j k l m p q s t v w x y z ; !This subset doesnot have n

VR = e o; !ROund vowels
VFR = i e; !Front vowels
VBK = o u; !Back vowels
A = f s; !Affricates
VS = c h k p s t ; !Voiceless Consonants
```

The keyword Rules marks the beginning of all the alternation rules in the source file.

Deletion rules:

Deletion rule concerns the deletion of –vowels, consonants and even syllables

//a// is deleted before any vowel in the initial position of a morpheme. Other vowels are glided as we shall see later.

Example

Aba-aana *abaana*

Ama-ooko *amooko*

Baa-uubatse *buubatse*

!Deletion rules

"(1) a deletion"

a:0 < = > C _ V ;

"(2) y deletion"

y:0 < = > [y|w|z:j|g:z|d:z] _ ;

"(3) w deletion"

w:0 < = > w: _ ;

"(4) r deletion"

r:0 < = C3 _ [y [e|i]] ;

"(5) n deletion"

n:0 < = > _ [n|m] ;

"(6) u deletion" !This rule has been created due to rule 21, so it will be moved to the deletion section

u:0 < = > :o _ C ;

"(7) k deletion"

k:0 < = > _ y:S ;

"(8) t deletion"

t:0 < = > _ y:s ;

Glide formation rule

This rule is about the formation of the glide. For example //i// becomes //y// or //a// becomes //y// when if it is alone with a morpheme (This is about morpheme -a- //u// becomes //w// if is alone on the morpheme.

i-angaaza iyangaaza

u-wu-iishe □uwiiyishe
u-i-ang-a □wiiyaanga

"(9) a becomes y"
a:y < = > .#. _ V ?;

"(10) i becomes y"
i:y < = > _ [a|e|o|u];

"(11) u becomes w" !We need to deal with situations where u is following i e.g u-i-anga
u:w < = > [.#.|C2] _ V;

"(12) o becomes w"
o:w < = > C _ [a|e|i|u] ;

"(13) n becomes m"
n:m < = > _ [b|f|p|v|:p|:f|:v] ;

"(14) r becomes d"
r:d < = n _ ;

"(15) pf becomes f" !pf and ts need to be dealt with in a better way
F:f < = > n _ ;

"(16) ts becomes s"
S:s < = > n _ ;

"(17) voiceless becoming voiced"
Ck:Cg < = > _ [V VS];
where Ck in (k t)
Cg in (g d)
matched;

"(18) h becomes p"
h:p < = > n:m _ ;

"(19) k becomes c"
k:c < = > _ :y ;

!Assimilation rules

```
"20 i becomes e" ! Here a combines with the i to form e
i:e <=> a:0 _ C;

"21 u becomes o" !a combines with u to form o
u:o <=> a:0 _ C ;

"22 Y becomes h" ! S combines with y to form h
y:h <=> s _ ;

"23 e becomes o" !When e of se unites with u they become o
e:o <=> s _ u:0;

"24 y becomes ts" ! Here k combines with y to give birth to ts
y:S <=> k: _ ;

"25 y becomes s"
y:s <=> t:0 _ ;

"26 z becomes j" !In this rule z followed by y becomes j
z:j <=> _ y: ;

"27 d and g become z"
g:z <=> _ y:0 ;

"28 d and g become z"
d:z <=> _ y:0 ;

"29 r becomes z"
r:z <= C3 _ y:0 ;

"30 w and y position change"
w:y <=> C2 _ y:w ;

"31 w and y position change"
y:w <=> w:y _ ;
```

Discussion and Conclusion

In this paper we have presented Kinyarwanda phonological /orthographical rules which form one of the basic two components required to develop a finite state two level morphology based morphological analyser. Some of the rules we have presented in this paper apply also to other Bantu languages, for example Dahl's law is applies to all Bantu languages. Since most rules are language dependant, the

rules presented in this paper may not directly be applied to other Bantu languages. What can be reused is the idea and the formats used. Since the introduction of replace rules which look the traditional rewrite rules, there has been debates on whether to use two level rules or the replace rules. Proponents of the two level rules would argue that the rewrite rules are outdated and should not be used today. On the other hand supporters of the replace rules, prefer them because of the close resemblance to traditional rewrites rules of Chomsky and Halle (1968). Mathematically speaking both rules are the same. Mathematically, a network produced in one way is equally as good as finite state network produced in a different way. So the choice of two level rules or replace rules eventually is a matter of choice and human ease of use (Beesley and Karttunen, 2003). Two level rules where also chosen because they have already successfully been used to develop a comprehensive morphological analyser for Swahili, a Bantu language. Xerox tools are fast, elegant, well documented and modular and have been experimented on other Bantu languages. This is why we decided to use Xerox finite state tools.

The rules presented in this paper have been tested thoroughly well and are now part of a running Kinyarwanda morphological analyser. Over 90% of Bantu languages are known to be tonal languages therefore future work will go into incorporating tone rules into these rules as we apply them to the lexicon for a comprehensive analysis of Kinyarwanda and other Bantu languages.

References

Beesley, K. (2003). Finite-State Morphological Analysis and Generation for Aymara. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics: the Workshop on Finite-State Methods in Natural Language Processing, pages 19-26. Budapest, Hungary.

Beesley, K. (1997). Finite-state descriptions of Arabic morphology. In Proceedings of the Second Cambridge Conference: Bilingual Computing in Arabic and English, Literary and Linguistic computing Center, Cambridge University, UK.

Beesley, K. AND Karttunen Lauri. (2003). Finite State Morphology: CSLI Studies in Computational Linguistics. Stanford University, CA: CSLI Publications.

Bosch, S.E and Pretorius, L. (2002). The significance of computational morphological analysis for Zulu lexicography, in South African Journal of African Languages, 2002, 22.1:11-20.

Bosch, S.E. and Pretorius, L. (2003). Building a computational morphological analyser/ generator for Zulu using the Xerox finite-state tools. Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 1314 2003, Budapest, Hungary. pp. 2734.

Chomsky, N., and Halle M. (1968). The sound Pattern of English. New York: Harper and Row. xiv, 470 pages. Reprinted 1991, Boston: MIT press.

Dale, R., Moisl, H., and Somers, H. (2000). Handbook of Natural Language Processing. New York: Marcel Dekker.

U´ı Dhonnchadha E., Nic Ph´aid´ın C., Genabith, J.V. (2003) Design, implementation and evaluation of an inflectional morphology finite state transducers for Irish. Machine Translation, Springer Vol 18 Number 12pgs 173-193.

Hurskainen, A. (1992). A two-level computer formalism for the analysis of Bantu Morphology an application to Swahili Nordic journal of African studies 1(1): 87-119 (1992)

Hurskainen, A. (1999). SALAMA Swahili Language Manager Nordic journal of African studies 8(2): 139-157

Karttunen, L., (2001) Applications of Finite State Transducers in Natural language Processing. In Implementations and Applications of Automata. Lecture notes in Computer science Vol 2088, Eds S. Yu and A. Paun, 34-46 Heidelberg:Springer Verlag.

Kimenyi, A.(1980)A Relational Grammar of Kinyarwanda. Berkeley and Los Angeles. University of California Press.

Kimenyi, A.(1979) Studies in Kinyarwanda and Bantu Phonology, Edmonton, Alberta: Linguistc Research Inc

Kimenyi, A. (1986) Syntax and semantics of reduplicatoin: A semiotic account La Linguistique, Vol 22 Fasc 2/1986

Kimenyi, A. (2002) A tonal Grammar of Kinyarwanda An Autosegmental and Metrical Analysis.The Edwin Mellen Press Ltd.

Kimenyi, A. (2004) Kinyarwanda morphology, In the International Handbook for inflection and word formation vol2.

Koskenniemi, K. (1983). Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. University of Helsinki: Department of General Linguistics.

Koskenniemi, K. AND Kenneth, C., (1988). Complexity, two-level morphology and Finnish. In Proceedings of the 12th International Conference on Computational Linguistics, pages 335-340. Association for Computational Linguistics

Liddy, E.D. (2003). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Oflazer, K. (1994). Two-level description of Turkish morphology. Literary and Linguistic Computing, 9:2.

Oflazer, K. (1999). Morphological analysis. In: Van Halteren H (ed) Syntactic Wordclass Tagging. Dordrecht: Kluwer Academic Publishers. pp. 175205.

Pretorius, L. & Bosch, S.E., (2003). Computational aids for Zulu natural language processing. Southern African Linguistics and Applied Language Studies 21/4:267-282

Pretorius, L. and Bosch, S. (2003). Towards technologically enabling the indigenous languages of South Africa: the central role of computational morphology. Interactions of the Association for Computing Machinery 10 (2) (Special Issue: HCI in the developing world): pp.56-63.

Roxas, R.S., (2000). Computational Linguistics Research on Philippine Languages. Software Technology Department: De La Salle University, Manila, Philippines. http://acl.ldc. upenn.edu//P/P00/P00-1074.pdf

Sproat, R.(1992). Computation and Morphology MIT press

7
A Methodology for Feature Selection in Named Entity Recognition

Fredrick Edward Kitoogo and Venansius Baryamureeba

In this paper a methodology for feature selection in named entity recognition is proposed. Unlike traditional named entity recognition approaches which mainly consider accuracy improvement as the sole objective, the innovation here is manifested in the use of a multiobjective genetic algorithm which is employed for feature selection basing on various aspects including error rate reduction and time taken for evaluation, and also demonstrating the use of Pareto optimization. The proposed method is evaluated in the context of named entity recognition, using three different data sets and a K-nearest Neighbour machine learning algorithm. Comprehensive experiments demonstrate the feasibility of the methodology.

1. Introduction

The Machine Learning approaches to the named entity recognition (NER) problem follow three major steps namely; (i) feature engineering, where identification of lexical and phrasal characteristics in text which expresses references to named entities (NEs) is done, (ii) algorithm selection, when the decision of which machine learning algorithm/algorithms to use for learning is made and (iii) classification, when the actual learning of the feature list to detect and classify the named entity phrases is done.

NE's are theoretically identified and classified by using features (various abstract entities that combine to specify underlying phonological, morphological, semantic, and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations).Two kinds of features that have been defined by McDonald (1996) [17] are internal and external features; internal features are the ones provided from within the sequence of words that constitute the entity, in contrast, external features are those that can be obtained by the context in which entities appear.

The choice of the best discriminative features to represent NE's affects many aspects of the NER problem such as accuracy, learning time, and the optimal training data set size. In many NER applications, it is not unusual to find problems involving hundreds of features. However, it has been observed that beyond a certain point, the inclusion of additional features leads to a worse rather than better performance (Oliveira et al., 2003) [19].

Feature engineering refers to the task of identifying and selecting an effective subset of features to represent entities from a larger set of often mutually redundant or

even irrelevant features. It encompasses feature design, feature selection, feature induction, and feature impact optimization (Rinnger, 2005) [22]. Feature selection; a sub-task of feature engineering is not a trivial problem since there may be (i) redundancy, where certain features are correlated so that it is not necessary to include all of them in modeling and (ii) interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own.

Many real-world problems like feature selection for named entity recognition involve the optimization of multiple objectives, such as number of features and accuracy. The tendency is that the different objectives to be optimized represent conflicting goals (such as improving the quality of a product and reducing its cost), in multiobjective optimization the optimization of each objective corresponds to an optimal solution. Therefore, in multiobjective optimization one usually wants to discover several optimal solutions, taking all objectives into account, without assigning greater priority to one objective or the other. Most named entity recognition systems as demonstrated in Tjong Kim Sang and De Meulder (2003) [24] tend to consider only one objective of improving accuracy.

There is need for systems which put into consideration other objectives like the cost of the solution on top of improvement of accuracy. The systems should be further able to provide users with different sets of optimal solutions thus giving the end-user the option of being able to choose the solution representing the best trade-off between conflicting objectives a posteriori, after examining a set of high-quality solutions returned by the named entity recognition system.

Intuitively, this is better than forcing the user to choose a trade-off between conflicting goals a priori. This paper proposes the use of a multi-objective genetic algorithm (MOGA) as a means to search for subsets of features (feature selection), which contain discriminatory information to classify named entities. The MOGA will generate a feature set of alternative solutions (from a fixed entire feature population) and use a cross-validation method to indicate the best accuracy/complexity (number of features)/cost of using the feature sub-set (in this case only time for classification was used as a cost) trade-off. The classification accuracy will be supplied by a machine learning algorithm.

The remainder of the paper is structured as follows: in Section 2, previous approaches are reviewed, in Section 3, we outline the proposed method and specify the feature population from which feature selection will be done. We describe the search (optimization) procedure of the method, in Section 4 the experiments and results are presented. Finally, Section 5 closes with a conclusion and an outlook for future work.

2. Previous Approaches

Many researchers have tackled feature selection in various ways and likewise the performance of the different approaches varies substantially. The more closely related approaches are presented in this section.

2.1 Complete Search

Some researchers manually designed features and used all of them without selecting optimal subsets (Carreras et. al (2003) [2]; Zhou et al., 2004 [26]; Shen et al., 2004 [23]) , while others leave the task of ignoring the useless features to the learning algorithm (Mayfield et. al, 2003) [16]. The problem with these approaches, is that they are computationally not feasible in practice.

2.2 Randomized Search

Randomized algorithms make use of randomized or probabilistic steps or sampling processes. Several researchers have explored the use of such algorithms for feature selection (Kira and Rendell, 1992 [12]; Liu and Setiono 1996 [15]) Li and McCallum (2004) [13] use conjunctions of the original features; where they use feature induction which aims to create only those conjunctions which significantly improve performance by starting with no features at all and iteratively choosing new features, from which sets are built and correspondingly evaluated at each iteration.

Hakenberg et. al (2005) [6] tackle feature engineering using what they refer to as recursive feature elimination (RFE); where they study the impact of gradual exclusion of features. Their model starts with a full model containing all features, they iteratively remove a number of features with the lowest weight, retrain the model, and check the resulting performance.

Bender et. al (2003) [1] use count-based feature reduction; where a threshold K is predetermined, and only those features that have been observed in the training data set at least K times are considered for the learning algorithm.

Jiang and Zhai (2006) [8] use what they termed generalizability-based feature ranking; here they target selecting features which are more generalizable (perform well in different domains). They use generalizability to mean the amount of contribution a feature can make to the classification accuracy on any domain, and to identify highly generalizable features, they compare their individual contributions to accuracy among different domains using a predefined scoring function.

The problem with the randomized search techniques is that they cannot properly handle the interdependence and correlation problem often associated with feature selection from large search spaces because they do not explore the whole search space at once.

2.3 Heuristic Search

Several authors have explored the use of heuristics for feature subset selection, often in conjunction with branch and bound search. Forward selection and backward elimination are the most common sequential branch and bound search algorithms used in feature selection (John et al., 1994) [10]. Most of these approaches assume monotonicity of some measure of classification performance. This ensures that adding features does not worsen the performance. However, many practical scenarios do not satisfy the monotonicity assumption. Moreover, this kind of search is not designed to handle multiple selection criteria.

Another branch of heuristic approaches employ genetic algorithms which do not require the restrictive monotonicity assumption. They can also deal with the use of multiple selection criteria, e.g. classification accuracy, feature measurement cost, etc. Due to the ability of genetic algorithms to deal with multiobjective optimization, some authors have explored genetic algorithms for feature selection for handwritten character recognition (Kim and Kim, 2000) [11].

Feature selection using genetic algorithm is often performed by aggregating different objectives into a single and parameterized objective, which is achieved through a linear combination of the objectives. The main drawback of this approach is that it is very difficult to explore different trade-offs between accuracy and different subsets of selected features. In order to overcome this kind of problem, Emmanouilidis et al. (2000) [3] proposed the use of a multi-criteria genetic algorithm to perform feature selection.

Li et al. (2005) [14] presented a gene selection approach based on a hybrid between genetic algorithms and support vector machines. The major goal of there hybridization was to exploit fully their respective merits (e.g., robustness to the size of solution space and capability of handling a very large dimension of feature genes) for identification of key feature genes (or molecular signatures) for a complex biological phenotype.

Jirapech-umpai and Aitken (2006) [9] designed an evolutionary algorithm to identify the near-optimal set of predictive genes that classify micro-array data, for multiple objectives of their problem they used a weighting function to compute the fitness of an individual in the population.

Hong and Cho (2006) [7] noted the problem of conventional feature selection with genetic algorithms in handling huge-scale feature selection. They modified the representation of a chromosome to be suitable for huge-scale feature selection and adopted speciation to enhance the performance of feature selection by obtaining diverse solutions.

3. The Proposed Methodology

A Genetic Algorithm (GA) (Goldberg, 1989) [5] is a search algorithm inspired by the principle of natural selection. The basic idea is to evolve a population of individuals, where each individual is a candidate solution to a given problem. Each individual is evaluated by a fitness function, which measures the quality of its corresponding solution. At each generation (iteration) the fittest (the best) individuals of the current population survive and produce offspring resembling them, so that the population gradually contains fitter and fitter individuals i.e., better and better candidate solutions to the underlying problem (Paapa et al., 2002) [21].

The work proposes a methodology that employs a multi-objective genetic algorithm (MOGA) as a means to select subsets of features from a pool of predesigned features, which contain the most optimal discriminatory information to classify named entities.

The inspiration of employing a MOGA for feature selection, was that: (i) GAs are a robust search method, capable of effectively exploring the large search spaces often associated with feature selection problems; (ii) GAs perform a global search (Paapa et al., 2002) [21], so that they tend to cope better with feature correlation and interdependence ; and (iii) GAs are suitable for multiobjective problem solving (Morita et al., 2003) [18], where the search algorithm is required to consider a set of optimal solutions at each iteration.

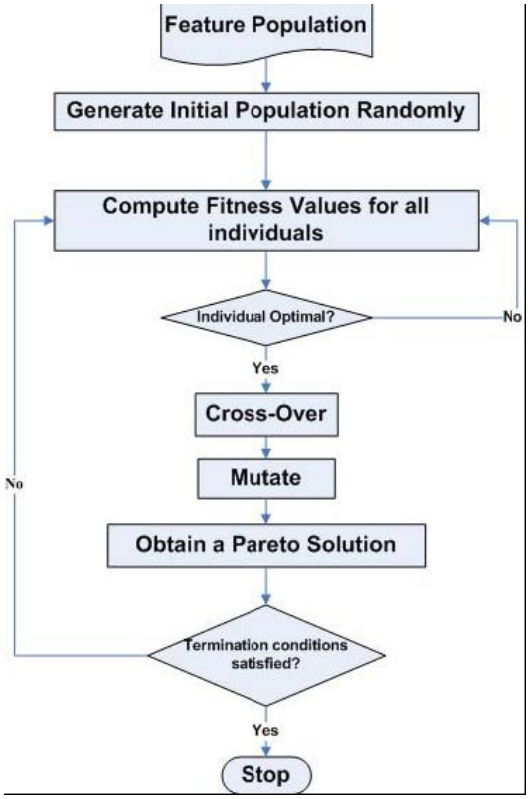
The general purpose Multiobjective Genetic Algorithm is to search for optimal subsets of relevant features the minimizes both the classification error rate, the classifier evaluation time involved, number of learning examples, number of features, and the cost of learning. This work is however, limited to the minimization of classification error rate and the classifier evaluation time involved.

Basically the proposed methodology as shown in Figure 1 works as follows: A MOGA will generate from a feature population, alternative feature set solutions from which the best fitness (error rate/time for evaluation of using the feature subset trade-off) will be determined. The fitness of all the individuals in the population will be ranked and a set of optimal individuals will be passed on for crossing-over and subsequently mutation. Pareto Optimal solutions will be selected for each generation. Consequently if the termination condition of the MOGA is reached the algorithm stops and gives the final solutions.The classification error rate and time will be supplied by the evaluation of a classifier built from a learning algorithm using 10 fold cross-validation on the training data set. The fitness function in this problem will be based on the error rate and time taken to evaluate the classifier which will be the objectives to be minimized using a Pareto optimization approach.

3.1 Individual Representation

Any candidate subset of selected features is represented as an individual which consists of N genes (equating to the total number of all features in the population). Each of the gene values is either turned on or off (0 or 1) for presence or absence in the candidate selected feature subset.

Fig. 1. Flow chart of the proposed methodology



3.2 Fitness Function

The fitness function is for measuring the quality of an individual (a candidate feature subset). There are two quality measures for an individual; (i) the error rate of the classification algorithm and (ii) the cost of using the feature set (in this case only time spent for evaluation was used as a cost). The two quality measures are computed by evaluating the average classifier algorithm error rate using cross-validation over ten folds and the time taken to evaluate a classifier using the specific feature set.

3.3 Pareto Optimization

A promising approach for performing feature selection is the multiobjective GAs aiming at producing solutions with Pareto Optimization (EvoGrid, 2006) [4]. The key concept here is dominance; To illustrate the concept, the multiobjective problem is represented mathematically in Eq. (1)

$$Minimize/Maximize\ F(x) = [f_1(x), f_2(x), f_3(x), ..., f_o(x)] \quad (1)$$

Where; $F(x)$ is a multiobjective function vector, $f_i(x)$ is the ith objective function, x is an input vector, o is the number of objective functions.

A solution is said to be Pareto optimal if it cannot be dominated by any other solution available, ie. a solution $x_i \in X$ is Pareto optimal iff there is no $x_j \in X$ such that $f_p(x_j) \leq f_p(x_i)\ \forall p \in \{1, 2, 3, ...o\}$

The use of a multiobjective algorithm basing on the concept of dominance can maintain population diversity, which allows for the discovery of a range of feature sets using accuracy/cost (time for evaluation) trade-offs.There are two popular *Pareto dominance techniques*:-

- ***Weak Pareto Dominance:*** A vector $F_m(x)$ weakly Pareto-dominates another vector $F_n(x)$ if none of the $F_n(x)$ coordinates is strictly greater than those in $F_m(x)$ and at least one of the coordinates in $F_m(x)$ is strictly greater than its counterpart $F_n(x)$.
- Strong Pareto Dominance: Here, all components of a dominant vector $F_m(x)$ are strictly greater than their counterpart in $F_n(x)$.

4. Experiments and Results

4.1 The Data Set

Three real-world datasets (as shown in Table 1) from the UCI Machine Learning Repository [25] were used for the experiments. Different feature sets for experimentation were built randomly out of the entire feature population of each data set.

Table 1. Details of the datasets used in the experiments

Datasets used in the Experiments			
Datasets	# Instances	# Classes	# Features
Breast Cancer	286	discrete/2	9
Hayes-roth	132	discrete/3	4
zoo	101	discrete/7	16

4.2 The Learning Algorithm

For these experiments, the Orange Data Mining Software, (2006) [20] K-nearest Neighbour machine learning algorithm (with default parameter settings) was used. The algorithm is separately trained using the MOGA on the three different data sets for five runs. For each, the optimal feature set is identified and the results generated (error rate and time for evaluation) are compared to those achieved using the entire feature set.

4.3 The Parameter Settings

In our experiments, the MOGA is based on bit individual representation, mutation, crossover and tournament selection based on weak-Pareto optimization (EvoGrid, 2006) [4]. Below are the parameter settings used in the experiments:

- Population size: 100
- Number of Generations: 10
- Probability of crossover (Two-Point): 0.9
- Probability of mutation (Bit Flip): 0.1
- Tournament Size: 8
- Pareto Optimization: Weak

4.4 Results

The main results of our experiments are reported in Table 2; the Table is divided into three Sub-Tables, each representing a particular data set. For each data set five runs of the MOGA were made; in each run, a comparison of performance (the number of features, error rate, and time spent on evaluation) between the original feature set and the one generated by the MOGA is made. The Left Hand Side of the Tables shows the performance of the original feature set while the Right Hand Side shows that using the feature set selected by the MOGA. In the columns "Original Features" and "Optimal Features" which are a bit representation of the used features, a "1" indicates presence of a feature in order of position, while "0" denotes absence of that position feature. The aim of the MOGA is to minimize both the error rate and time, implying that a value depicts better performance than another (comparable one) when the former value is less than the latter value.

Clearly, as shown in Table 2, the performance (error rate and time for evaluation) for all runs in two data sets [Breast Cancer and Zoo] is better when using the MOGA selected feature sets than with all features. In general the average performance for the MOGA selected feature sets in the two data sets is significantly (a value is significantly better than another when the corresponding upper and lower bounds using the standard deviation do not overlap) better than that with all features; this finding supports the fact of inclusion of only optimal features in a set, which precipitates the assertion of many researchers that interdependence and correlation in target features affect feature selection task. The performance on the Hayes-Roth Data Set is actually identical, except the time to evaluate which is marginally significantly divergent.

Table 3 demonstrates the concept of *Pareto Dominance*, which explains the difficulty in determining trade-offs between several optimization objectives used in other approaches like the simple weighted approach; in the Breast Cancer data set it is seen that the second result has a better error rate than the first result, however the time component results are the opposite. A similar situation is shown in the Zoo Data Set where the error rate for the third result is better than that of the first and second result but the time performance is the opposite. In both situations the feature sets perform are better than all the other possible feature sets (non-dominated) but a better solution between themselves is not easily identifiable a priori. The situation here can only be determined a posteriori by an end-user for a perceived better option (using a choice of their best error rate/ Time trade-off).

Table 2. Performance of the K-Nearest Neighbour using a MOGA on the three different data sets

Hayes-Roth Data Set; Instances = 132								
Run	Original Feature	Features #	Error Rate	Time (Secs)	Selected Features	Feature #	Error Rate	Time (Secs)
Run 1	1111	4	0.3044	0.0160	1111	4	0.3044	0.0149
Run 2	1111	4	0.3044	0.0620	1111	4	0.3044	0.0149
Run 3	1111	4	0.3044	0.0630	1111	4	0.3044	0.0149
Run 4	1111	4	0.3044	0.0620	1111	4	0.3044	0.0149
Run 5	1111	4	0.3044	0.0469	1111	4	0.3044	0.0149
Average			0.3044	0.0500			0.3044	0.0149
Std. Dev.			0.0000	0.0201			0.0000	0.0000

Breast Cancer Data Set; Instances = 286								
Run	Original Feature	Features #	Error Rate	Time (Secs)	Selected Features	Feature #	Error Rate	Time (Secs)
Run 1	111111111	9	0.3005	0.1099	011001110	5	0.2793	0.0780
Run 2	111111111	9	0.3005	0.1089	000010100	2	0.2311	0.0779
Run 3	111111111	9	0.3005	0.1089	010110110	5	0.2797	0.0929
Run 4	111111111	9	0.3005	0.1870	010011001	4	0.2369	0.0779
Run 5	111111111	9	0.3005	0.1100	000010100	2	0.2311	0.0779
Average				0.3005	0.1249		0.2516	0.0809
Std. Dev.				0.0000	0.0347		0.0256	0.0067

Table 3. Pareto Optimal Solutions for the a K-nearest Neighbour machine learning algorithm on the Breast Cancer and Zoo Data Sets

Zoo Data Set; Instances = 101								
Run	Original Feature	Features #	Error Rate	Time (Secs)	Selected Features	Feature #	Error Rate	Time (Secs)
Run 1	1111111111111111	16	0.0291	0.0160	1111010011001010	9	0.0091	0.0001
Run 2	1111111111111111	16	0.0291	0.0469	1111110100111101	12	0.0100	0.0149
Run 3	1111111111111111	16	0.0291	0.0469	1010110010011010	8	0.0190	0.0001
Run 4	1111111111111111	16	0.0291	0.0469	1111110011001111	12	0.0091	0.0149
Run 5	1111111111111111	16	0.0291	0.0320	1111110011001111	12	0.0091	0.0149
Average				0.0291	0.0377		0.0113	0.0090
Std. Dev.				0.0000	0.0138		0.0043	0.0081

Table 4. Comparison of Percentage Improvement in Average Error Rate of all three Data Sets

Cancer Data Set - Run 1			
Selected Feature Set	Features #	Error Rate	Time (Secs)
001101000	4	0.2974	0.0780
011001110	9	0.2793	0.0929
000101000	16	0.2898	0.0779

Zoo Data Set - Run 4			
Selected Feature Set	Features #	Error Rate	Time (Secs)
0110110001001100	7	0.0291	0.0001
0101110000011100	7	0.0291	0.0001
1111110011001111	12	0.0091	0.0149

Table 5

All Data Sets				
Date Set	Original Feature Set Size	Original Average Error Rate [C3]	MOGA Average Error Rate C4	% age Error Rate Improvement
Hayes-Roth	4	0.2974	0.0780	0
Breast Cancer	9	0.2793	0.0929	16.26
Zoo	16	0.2898	0.0779	61.30

Table 5 shows that as the number of features increases in a data set, MOGA performs (error rate) better, with this observation it is possible to conclude that employment of is this technique is best suited for data sets with a large number of original features. All the above results were obtained using a Pentium-M 760 Laptop with clock speed of 2.0 GHz and 1 GB of RAM

5 Conclusion and Future Work

We have introduced a methodology to perform feature selection for the classification task in named entity recognition based on a multiobjective genetic algorithm. We have experimented this approach with the application of a weak Pareto-tournament selection genetic algorithm and a k-Nearest Neighbour machine learning algorithm and demonstrated its efficacy on three real world data sets. We have shown that the multiobjective algorithm is well suited for feature selection and that it has a benefit of yielding different solution options based on error rate/cost trade-offs, leaving end-users with an option of deciding from alternative solutions.

It has also been discovered that multiobjective genetic algorithms are justifiably more suitable for feature selection where the number of features is large enough to make other methods computationally more expensive.

Future work should examine niched approaches to Pareto optimization, and more pragmatic approaches to determining optimal population and generation sizes for experiments. Studies about specific feature impact optimization and interdependence need to be carried out. Deeper comparison of weighted approaches to multiobjective genetic algorithms should be considered. The multiobjective genetic algorithm methodology to feature selection for named entity recognition offers more potential for considering many more objectives in the named entity recognition task other than accuracy improvement only.

References

Bender, O., Och, F. J. and Ney, H. (2003). Maximum Entropy Models for Named Entity Recognition. Proceedings of CoNLL-2003, 148151.

Carreras, X., Marquez, L. and Padro, L. (2003) A simple named entity extractor using AdaBoost. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 4: 152-155.

Emmanouilidis, C., Hunter, A., and MacIntyre, J. (2000). A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. Proc. Congress on Evolutionary Computation,1,309-316.

EvoGrig - Multi Objective Optimization with EvoGrid, http://champiland.homelinux.net/evogrid/doc/multiobjective.html, accessed 15/12/2006.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company

Hakenberg, J., Bickel, S., Plake, C., Brefeld, U., Zahn, H., Faulstich, L., Leser, U., and Scheffer, T. (2005). Systematic feature evaluation for gene name recognition. BMC Bioinformatics, 6(l):S9.

Hong, J. H., and Cho, S. B. (2006). Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognition Letters, 27: 143-150.

Jiang, J. and Zhai, C. X. (2006). Exploiting Domain Structure for Named Entity Recognition. Urbana, 51: 61801.

Jirapech-umpai, T. and Aitken, S. (2006). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6:148.

John, G., Kohavi, R., and Peger, K. (1994). Irrelevant features and the subset selection problems. Proc. 11th Int. Conf. Machine Learning, 121-129.

Kim, G., and Kim, S. (2000). Feature selection using genetic algorithms for handwritten character recognition. Proc. 7th Int. Workshop on Frontiers of Handwriting Recognition (IWFHR), pp. 103-112.

Kira, L., and Rendell, L. (1992). A practical approach to feature selection. Proc. 9th Int. Conf. Machine Learning, pp. 249-256

Ll, W. and McCallum, A. (2004). Rapid development of Hindi named entity recognition using conditional random fields and feature induction. In ACM Transactions on Asian Language Information Processing (TALIP), 2: 290- 294.

Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L.,Wang, Q., Topol, E. J. and Rao, S. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85: 16-23.

Liu, H., and Setiono, R. (1996). A probabilistic approach to feature selection a filter approach. Proc. 13th Int. Conf. Machine Learning, pp. 319-327.

Mayfield , J., McNamee, P. and Piatko, C. (2003). Named entity recognition using hundreds of thousands of features. In Proceedings of CoNLL-2003, pp. 184187.

McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. Corpus Processing for Lexical Acquisition, pp. 21-39.

Morita, M., Sabourin, R., Bortolozzi, F., Suen, C. Y. and De Technologie Superieue, E. (2003). Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition. Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference, pp. 666-670.

Oliveira, L. S., Sabourin, R., Bortolozzi, F., and Suen, C. Y. (2003). A Methodology for Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Digit String Recognition. International Journal of Pattern Recognition and Artificial Intelligence, 17:903-929

Orange Data Mining: Fruitful & Fun, http://magix.fri.uni-lj.si/orange, accessed 04/05/2006.

Pappa, G. L., Freitas, A. A. and Kaestner, C. A. A. (2002). Attribute Selection with a Multiobjective Genetic Algorithm. XVI Brazilian Symposium on Artificial Intelligence. 2057:280290.

Rinnger, E. (2005). The ACL 2005 Workshop on Feature Engineering for Machine Learning in Natural Language.

Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C. (2004). Multi-Criteria-based Active Learning for Named Entity Recognition.

Tjong Kim Sang, E.F. and De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2003 (2003) 142-147

The Machine Learning Repository. http://www.cs.uci.edu/mlearn/ MLRepository. html accessed 04/05/2006 (2006)

Zhou, G.D., Shen, D., Zhang, J., Su, J. and Tan, S.H. (2004). Recognition of Protein/Gene Names from Text using an Ensemble of Classifiers. BMC Bioinformatics.

8

Extraction of Interesting Association Rules Using Genetic Algorithms

Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba

The process of discovering interesting and unexpected rules from large data sets is known as association rule mining. The typical approach is to make strong simplifying assumptions about the form of the rules, and limit the measure of rule quality to simple properties such as support or confidence. Support and confidence limit the level of interestingness of the generated rules. Comprehensibility, interestingness and surprise are metrics that can be used to improve on interestingness. Because these measures have to be used differently as measures of the quality of the rule, they can be considered as different objectives of the association rule mining problem. The association rule mining problem, therefore, can be modelled as multi-objective problem rather than as a single-objective problem. In this paper we present a Pareto–based multi–objective evolutionary algorithm rule mining method based on genetic algorithms. We use confidence, comprehensibility, interestingness, surprise as objectives of the association rule mining problem. Specific mechanisms for mutations and crossover operators together with elitism have been designed to extract interesting rules from a transaction database. Empirical results of experiments carried out indicate high predictive accurracy of the rules generated.

1. Introduction

Association rule mining (ARM) is one of the core data-mining techniques and has attracted tremendous interest among researchers and practitioners. The major aim of ARM is to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. ARM algorithms discover high-level prediction rules in form: *IF the condition of the values of the predicting attributes are true, THEN predict a value for some goal attribute.*

Association rule mining is defined as follows: Let $I = \{ i_1, i_2, ..., i_m \}$ be a set of items. Let T be a set of transactions, where each transaction t is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ holds in the transaction set T with confidence c if c% of transactions in T that support X also support Y. The rule has support s in T if s% of the transactions in T contains $X \cup Y$.

In the following example of a bookstore sales database, the association rule mining task is exemplified. There are five different items (authors of novels that the bookstore deals in), $I = \{A,C,D, T,W\}$. There are six customers in the database who purchased books authored by these authors. The table below shows

all frequent itemsets containing at least three authors (i.e. minimum – support = 50%.) It also shows the set of all transactions

Items	
John Ayo	A
Alfred Chihoma	C
Bernard Dungu	D
Thomas Talire	T
Peter Walwasa	W

Database	
Transaction	Items
1	ACTW
2	CDW
3	ACTW
4	ACDW
5	ACDTW
6	CDT

FI	
Support	Item sets
100%	C
83%	W, CW
67%	A,D,T,AC,A, CD, CT,ACW
50%	AT,DW,TW,ACT,ATW,CDW,CTW,ACTW

Association Rules		
A → C(4/4)	AC →W(4/4)	TW →C(3/3)
A →W(4/4)	AT →C(3/3)	AT →CW(3/3)
A →CW(4/4)	AT →W(3/3)	TW→AC(3/3)
D →C(4/4)	AW →C(4/4)	ACT→W(3/3)
T →C(4/4)	DW →C(3/4)	ATW→C(3/3)
W →C(5/5)	TW →A(3/3)	CTW→A(3/3)

Considering the first association rule A → C [support = 50%, confidence = 67%], which says that 50% of people buy books authored by John Ayo (A) and those by Alfred Chihoma (C) together, and 67% of the people who buy books written by John Ayo (A) also purchase those by Alfred Chihoma (C).

According to Zaki (1999) [11] the mining task involves generating all association rules in the database that have a support greater than the minimum support (the rules are frequent) and that have a confidence greater than minimum confidence (rules are strong). The ARM problem is an NP – Hard problem because finding all frequent itemsets (FI's) having a minimum support results in a search space of 2^m, which is exponential in m, where m is the number of itemsets. The final step involves generating strong rules having a minimum confidence from the frequent itemsets. It also includes generating and testing the confidence of all rules. Since each subset of X as the consequent must be considered, the rule generation step's complexity is $O(r.2^l)$, where r is the number of frequent itemsets, and l is the longest frequent itemset.

Traditionally, ARM was predominantly used in market-basket analysis but it is now widely used in other application domains including customer segmentation, catalog design, store layout, and telecommunication alarm prediction. ARM is computationally and I/O intensive. The number of rules grows exponentially with the number of items. Because data is increasing in terms of both the dimensions (number of items) and size (number of transactions), one of the main attributes needed in an ARM algorithm is scalability: the ability to handle massive data stores. Sequential algorithms cannot provide scalability, in terms of the data dimension, size, or runtime performance, for such large databases.

In this paper, we shall deal with the ARM problem as a multi-objective problem rather than as a single one and try to solve it using multi-objective evolutionary algorithms (MOEA) with emphasis on genetic algorithms (GA). The main motivation for using GAs is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining tasks. Multi-objective optimisation with evolutionary algorithms is well discussed by (Fonseca and Fleming, 1998) [4] and (Freitas, 2003 [7]).

1.1 Organisation and notation

The rest of this paper is organised as follows. In Section 2 we provide an overview of work related to the association rule mining problem. In Section 3 we discuss the proposed MOEA. In Section 4, the proposed algorithm is presented. In Section 5 the analysis of the results are presented. Section 6 is comprised of the conclusion.

Throughout this paper, we use the following notation. Po denotes the initial population and $\bar{}$ Po denotes the external set. | • | denotes the cardinality of a set, corresponds to the Pareto dominance relation. An association is defined as an expression X → Y with confidence c and support s where I = $\{ i_1, i_2, ..., i_m \}$ is a set of items and X = $\{i_{x1}, i_{x2}, ..., i_{xr}\} \subseteq$ I, $1 \leq r \leq (m - 1)$ and Y = $\{i_{y1}, i_{y2}..., i_{yr}\} \subset I$ are subsets of items. $D = \{T_1, T_2, ..., T_n\}$, $T \subseteq I$ denotes a database of transactions T, which are themselves subsets of items. The support of an itemset X, is denoted by σ(X). σ (X) is the number of transactions in which that itemset occurs as a subset. A k – subset is a k – length subset of an itemset. An itemset is *frequent* or large if its support is more than a user-specified minimum – support (min sup)

value. F_k is the set of frequent k – *itemsets*. A frequent itemset is *maximal* if it is not a subset of any other frequent itemset. The rule's support is the joint probability of a transaction containing both X and Y , and is given as $\sigma (X \cup Y)$. The confidence of the rule (also known as the predictive accuracy of a rule) is the conditional probability that a transaction contains B, given that it contains A and is given as $(X \cup Y)/ \sigma (X)$. O denotes the order of complexity of a computation and ! denotes a logical negation.

2. Related Work

The Association Rule mining problem was introduced in 1993 by Agrawal et al. (1993) [1]. Agrawal et al. (1993) [1] developed the Apriori algorithm for solving the association rule mining problem. Most of the existing algorithms are improvements to Apriori algorithm (Ghosh and Nath, 2004) [5], (Zhao and Bhowmick, 2003 [12]). These algorithms work on a binary database, termed as market basket database. On preparing the market basket database, every record of the original database is represented as a binary record where the fields are defined by a unique value of each attribute in the original database. The fields of this binary database are often termed as an item. For a database having a huge number of attributes and each attribute containing a lot of distinct values, the total number of items will be huge. Storage requirements resulting from the binary database is enormous and as such it is considered one of the limitations of the existing algorithms.

The Apriori – based algorithms work in two phases (*Agrawal et al.*, 1993) [1]. The first phase is for frequent item-set generation. Frequent item-sets are detected from all-possible item-sets by using support and minimum – support. If the value of minimum support is too high, number of frequent item sets generated will be less, and thereby resulting in generation of few rules. And, if the value is too small, then almost all possible item sets will become frequent and thus a huge number of rules may be generated. This causes inference basing on these rules to be difficult. After detecting the frequent item-sets in the first phase, the second phase generates the rules using minimum confidence. Confidence factor or predictive accuracy of a rule is defined as

$$Confidence = \sigma (X \cup Y) / \sigma (X) \qquad (1)$$

The second phase is concerned with generating the rules using the minimum confidence. Another limitation of the Apriori – based algorithms is the encoding scheme where separate symbols are used for each possible value of an attribute (Ghosh and Nath, 2004) [5]. This encoding scheme may be suitable for encoding the categorical valued attributes, but not for encoding the numerical valued attributes as they may have different values in every record. To avoid this situation, some ranges of values may be defined. For each range of values an item is defined. This approach is also not suitable for all situations. Defining the ranges will create yet another problem, as the range of different attributes may be different.

Existing Apriori – based algorithms, try to measure the quality of generated rule by considering only one evaluation criterion, i.e., confidence factor or predictive accuracy (Ghosh and Nath, 2004) [5]. This criterion evaluates the rule depending on the number of occurrence of the rule in the entire database.

To improve the quality of the generated rules, some multi-objective algorithms have been developed with more measures considered (Ghosh and Nath, 2004) [5]. Ghosh and Nath used the *comprehensibility, interestingness,* and *confidence* measures of the rules as objectives to model the association rule mining problem as a multi – objective problem. Using only three measures compromises the quality of the generated rules. The developed algorithm, Multi – objective rule mining genetic algorithm (MOGA) make use of J – measure, and surprise, which are measures of the quality of the generated rules.

In this paper we will apply the comprehensibility, surprise, interestingness and confidence measures to form a coherent set of four complementary measures to extract interesting rules. The details of these measures are given in the following Section.

3. Multi-objective optimization and rule mining problems

It is not an easy task to find a single solution for a multi-objective problem. In such situations the best approach is to find a set of solutions depending on nondominance criterion. At the time of taking a decision, the solution that seems to fit better depending on the circumstances can be chosen from the set of these candidate solutions. A solution, say a, is said to be dominated by another solution, say b, if and only if the solution b is better or equal with respect to all the corresponding objectives of the solution a, and b is strictly better in at least one objective. Here the solution b is called a non-dominated solution. So it will be helpful for the decision-maker, if we can find a set of such non-dominated solutions. Vilfredo Pareto suggested this approach of solving the multi objective problem. Optimization techniques based on this approach are termed as Pareto optimization techniques. Based on this idea, several genetic algorithms were designed to solve general multiobjective problems (Ghosh and Nath, 2004) [5]. Some association rule mining multiple – objective algorithms have been developed and MOGA is one of them.

In association rule mining if the number of conditions involved in the antecedent part is less, the rule is more comprehensible. We therefore require an expression where the number of attributes involved in both the parts of the rule has some effect. The following expression can be used to quantify the comprehensibility of an association rule

$$\text{Comprehensibility} = \log(1+ \mid C \mid) + \log(1+ \mid A \cup C \mid) \qquad (2)$$

Here, $\mid C \mid$ and $\mid A \cup C \mid$ are the number of attributes involved in the consequent part and the total rule, respectively.

It is important that we extract only those rules that have a comparatively less occurrence in the entire database. Such a surprising rule may be more interesting to the users; which again is difficult to quantify. According to Liu et al. (2000), the interestingness issue has long been identified as an important problem in data mining. It refers to finding rules that are interesting/useful to the user, not just any possible rule. The reason for its importance is that, in practice, it is all too easy for a data mining algorithm to discover hige range of rules most which are of no interest to the user.

For finding interestingness the data set is to be divided based on each attribute present in the consequent part. Since a number of attributes can appear in the consequent part and they are not predefined, this approach may not be feasible for association rule mining. So a new expression is defined which uses only the support count of the antecedent and the consequent parts of the rules, and is defined as

$$I = [\sigma (A \cup C)/(A)] \times [\sigma (| A \cup C |)/ \sigma (C)] \times [1 - \sigma (A \cup C)/ | D |)] \quad (3)$$

where I is Interestingness and $| D |$ is the total number of records in the database, $(A \cup C)/ \sigma (A)$ gives the probability of generating the rule depending on the antecedent part, $\sigma (| A \cup C |)/ \sigma (C)$ gives the probability of generating the rule depending on the consequent part, and $\sigma (A \cup C)/| D |)$ gives the probability of generating the rule depending on the whole data-set. This means that the complement of this probability will be the probability of not generating the rule. Thus, a rule having a very high support count will be measured as less interesting. The formulae for comprehensibility, interestingness and support as shown in equations (1), (2) and (3) respectively, were adopted from (Ghosh and Nath 2004) [5].

4. Genetic Algorithms with Modifications

In this paper we propose to solve the association rule-mining problem with a Pareto based multipleobjective genetic algorithm. The possible rules are represented as chromosomes and a suitable encoding/decoding scheme has been defined. A modified Michigan encoding/decoding scheme is used in this paper which associates two bits to each attribute. If these two bits are 00 then the attribute next to these two bits appears in the antecedent part and if it is 11 then the attribute appears in the consequent part. And the other two combinations, 01 and 10 will indicate the absence of the attribute in either of these parts. So the rule ACF $\Rightarrow$ BE will look like 00A 11B 00C 01D 11E 00F. In this way we can handle variable length rules with more storage efficiency, adding only an overhead of $2k$ bits, where k is the number of attributes in the database. The decoding should be performed as follows:

$$DV = \text{minval} + (maxval - minval) \times ((\Sigma^{(2i-1 \times \text{ithbitvalue})/(2n - 1)} \quad (4)$$

where DV is the decoded Value, 1 i n and n is the number of bits used for encoding; and Minval and maxval are minimum and maximum values of the attribute. For

brevity, this encoding scheme will not deal with relational operators and as such the rules generated from this formula will not include relational operators. Due to the fact that there may be a large number of attributes in the database, in this paper we propose to use multi-point crossover operator. There are some difficulties to use the standard multi-objective GAs for association rule mining problems. In case of rule mining problems, we need to store a set of better rules found from the database. If we follow the standard genetic operations only, then the final population may not contain some rules that are better and were generated at some intermediate generations. These rules should be kept. For this task, an external population is used. In this population no genetic operation is performed. It will simply contain only the nondominated chromosomes of the previous generation. At the end of first generation, it will contain the nondominated chromosomes of the first generation. After the next generation, it will contain those chromosomes, which are non-dominated among the current population as well as among the non-dominated solutions till the previous generation.

The scheme applied here for encoding/decoding the rules to/from binary chromosomes is that the different values of the attributes are encoded and the attribute names are not. For encoding a categorical valued attribute, the market basket encoding scheme is used. For a real valued attribute their binary representation can be used as the encoded value. The range of values of that attribute will control the number of bits used for it.

The archive size is fixed, i.e., whenever the number of nondominated individuals is less than the predefined archive size, the archive is filled up by dominated individuals. Additionally, the clustering technique used does not loose boundary points.

4.1 Fitness Assignment

In this paper, to avoid the situation that individuals dominated by the same archive members have identical fitness values, for each individual both dominating and dominated solutions are taken into account. In detail, each individual i in the archive $\bar{P}t$ and the population Pt is assigned a strength value S(i), representing the number of solutions it dominates:

$$S(i) = |\{j \mid \in Pt + \bar{P}t \ \wedge i \succ j\}| \tag{5}$$

On the basis of S the values, the raw fitness R(i) of an individual i is calculated:

$$R(i) = \sum_{j \in P_t + \bar{P}t \; j \succ i} S(j) \tag{6}$$

This implies that the raw fitness is determined by the strengths of its dominators in both archive and population. With the fitness value $R(i) = 0$ to be minimized this means that this is a nondominated individual, while a high $R(i)$ value means that i is dominated by many individuals. Although the raw fitness assignment provides a sort of niching mechanism based on the concept of Pareto dominance, it may fail when most individuals do not dominate each other. Therefore, additional density information is incorporated to discriminate between individuals having identical raw fitness values (Zitzler, et al. 2001) [15]

4.2 Environmental Selection

The number of individuals contained in the archive is constant over time, and the truncation method prevents boundary solutions being removed. During environmental selection, the first step is to copy all nondominated individuals, i.e., those which have a fitness lower than one, from archive and population to the archive of the next generation. If the nondominated front fits exactly into the archive the environmental selection step is complete. In case the archive is too small, the best dominated individuals in the previous and population are copied to the new archive. Otherwise, truncate the archive.

5. Experiments

The proposed algorithm was tested on a dataset drawn from the UCI repository of machine learning databases [10]. For brevity, the data used is of a categorical nature. The datasets contains zoo information. The zoo database contains 101 instances corresponding to animals and 18 attributes. The attribute corresponding to the name of the animal was not considered in the evaluation of the algorithm. This was mainly due to its descriptive nature. The attribute from the datasets that were used for analysis include: hair [H], feathers [F], eggs [E], milk [M], predator [P], toothed [TH], domestic [D], backbone [B], fins [N], legs [L], tail [T], catsize [C], airborne [A], aquatic [Q], breathes [BR], venomous [V], and type [Y].

Default values of the parameters are: Population size = 40, Mutation rate = 0.5, Crossover rate = 0.8, Selection in Pareto Archive (elitism) = 0.5. The stopping criterion used is the non evolution of the archive during 10 generations, once the minimal number of generations has been overpassed.

5.1. Results and Discussion

In the following table are the results of the experiments conducted. In the first row is the discovered rule, in the second row is the rule's comprehensibility, in the third is the interestingness of the rule and in the last column is the predicative accuracy of the rule. In these tests, different predictions were made by combining different attributes to determine a result.

The following are results from the given data

Discovered Rule	Compre-hensibility	Interest-ingness	Predictive Accuracy
If (!H and E and !M and B and T and D) Then (!P)	0.97	0.74	0.90
If(!A and Q and B and C) Then (P)	0.95	0.67	0.94
If(E and A and P and !V) Then (!D)	0.96	0.36	0.98
If(!E and !Q and !T) Then (D)	0.97	0.93	0.50
If(!E and !V and !D) Then (Y = 1)	0.95	0.77	0.98
If(F and !V and !D) Then (Y = 2)	0.94	0.90	0.97
If(E and !Q and P and TH and !N and !D and !C) Then(Y = 3)	0.94	0.97	0.83
If(Q and !BR and !V and T) Then(Y = 4)	0.93	0.93	0.95
If(!A and Q and T and BR and !C) Then(Y = 5)	0.94	0.98	0.80
If(A = 1)and(!N)and(!T) Then(Y = 6)	0.93	0.96	0.90
If(P)and(BR)and(!T)and (!D) Then(Y = 7)	0.95	0.95	0.92

As it is indicated in the results table, overall the discovered rules have a high predictive accuracy ande are quite interesting. Four rules have a predictive accuracy of over 90% while seven rules have a confidence of over 90%. Three rules have a high predictive accuracy of over 80%. Only two rules have a predictive accuracy of less than 50%.

6. Conclusion and Future Works

In this paper we have dealt with a challenging NP-Hard association rule mining problem of finding interesting association rules. The results of this paper are good since the discovered rules are of a high predictive accuracy and of a high intrestingness value.

It is worth noting, however, that the test data set was on categorical data and of a small sample size. Subjecting this algorithm to larger sample sizes and different data types is paramount.

References

Agrawal R., Imielinski T., Swami A., (1993). Mining Association Rules Between Sets of Items in Large Databases. Proc. of the 1993 ACM SIGMOD Conf. on Management of Data

Agrawal R., Srikant R.(1994). Fast Algorithms for Mining Association Rules. Proc. of the 20th Int'l Conf. on Very Large Data Bases

Liu B., Hsu W., Chen S. and Ma Y. (2000). Analyzing the Subjective Interestingness of Association Rules. IEEE Intellgent Systems

Fonseca M. C. and Fleming J. P. (1998) Multi–objective Optimization and Multiple Constraint Handling with Evolutionary Algorithms-Part I: A Unified Formulation. IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans, 28(1):26-37

Ghosh A. and Nath B.,(2004). Multi – objective rule mining using genetic algorithms. Information Sciences 163 pp 123133

http://www.lifl.fr/OPAC/guimoo. Accessed April 2006

Freitas A. A., (2003) A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. Advances in evolutionary computing: theory and applications, Pp 819 845

Khabzaoui M., Dhaenes C., and Talbi E. (2005). Parallel Genetic Algorithms for Multi – Objective rule mining. MIC2005. The 6th Metahueristics International Conference, Vienna, Austria.

Liu B., Hsu W., Chen S. and Ma Y. (2000). Analyzing the Subjective Interestingness of Association Rules. IEEE Intellgent Systems.

http://www.ics.uci.edu/˜ mlearn/MLRepository.html. Accessed April 2006

Zaki, M.J. (2001). Generating non-redundant association rules. In Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY: ACM, 34-43.

Zhao Q. and Bhowmick S. S. (2003). Association Rule Mining: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

Zitzler E., Deb K., Thiele L. (1998). An evolutionary Algorithm for Multi – objective Optimization: The Strength Pareto Approach

Zitzler E., Deb K., Thiele L. (2000). Comparison of Multi – objective Evolutionary Algorithms: Empirical Results. Evolutionary Computation Vol. 8 Number 2 pp 173 -195

Zitzler E., Luamanns M., Thiele L. (2001). SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Computer Engineering and Networks Laboratory (TIK), TIK-Report 103

9
Efficient IP Lookup Algorithm

K.J.Poornaselvan1,S.Suresh, C.Divya Preya and C.G.Gayathri

The rapid growth of traffic in the Internet, backbone links of several gigabits per second are commonly deployed. To handle gigabit-per-second traffic rates, the backbone routers must be able to forward millions of datagrams per second on each of their ports. Fast IP address lookup in the routers, which uses the datagram's destination address to determine for each datagram the next hop, is therefore crucial to achieve the datagram forwarding rates required. Also the packet may encounter many routers before it reaches its destination. Hence decrease in delay by micro seconds results in immense cut down in the time to reach the destination .IP address lookup is difficult because it requires a Longest Matching Prefix search . Many lookup algorithms are available to find the Longest Prefix Matching; one such is the Elevator-Stairs Algorithm. It provides a total search time of O (w/k + k) by indexing hash table to Practical Algorithm to Retrieve Information Coded in Alphanumeric (PATRICIA), where w is the length of the IP address and k is the level of Trie. Elevator Stairs Algorithm uses linear search at the k-level is modified to binary search at the k-level of Trie. At the kth-level, non branching nodes are added to jump k levels of Trie which reduces the time for searching in the Trie. It provides a better search time over the existing Elevator- Stairs Algorithm, by accomplishing a two-way search in the trie.

1.0 Introduction

1.1 Address Lookup

The primary role of routers is to forward datagram toward their final destinations. For this purpose, a router must decide for each incoming datagram where to send it next. More exactly, the forwarding decision consists of finding the address of the next-hop router as well as the egress port through which the datagram should be sent. This forwarding information is stored in a forwarding table that the router computes based on the information gathered by routing protocols. To consult the forwarding table, the router uses the datagram's destination address as a key; this operation is called ***address lookup***.

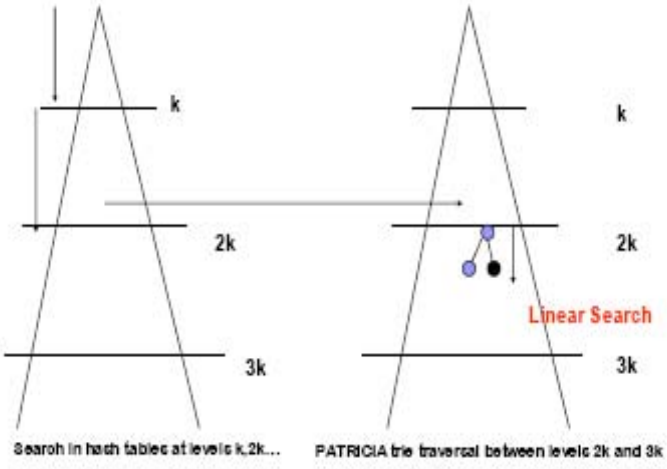
Once the forwarding information is retrieved, the router can transfer the datagram from the incoming link to the appropriate outgoing link, in a process called ***switching***. The key issue in router performance is the IP routing lookup mechanism used for the ferrying of the large number of datagrams which are transmitted from source to intermediate/destination router (see [5]). A serious time delay in datagram transmission occurs over the internet as a delayed search process occurs when finding the best matching IP address in the routing table.

These limitations probed to implement an efficient IP lookup Algorithm which reduces the searching time of an IP Address in Routing Table. The feature of insertion of a new IP address and deletion of an existing IP address is encapsulated within this project.

2.0 Elevator Stairs Algorithm

Elevator-Stairs Algorithm ameliorates the IP lookup efficiency in the large IP Routing tables (see [4]). It uses a path compressed trie which is called as Trie (refer [2]). In this trie, traversed path of any two nodes are compressed through the removal of non-branching nodes between them. IP lookup is initialized by searching the longest matching prefix in the k-level tree, which is build from Trie by adding non-branching nodes. As a next step, multiple bit key corresponding to the matching prefix is determined. Then a search of the hash table at the levels k, 2k, 3k,.. is performed, where the value of k is a constant between zero and maximum length of the IP address. If it matches with the level it searches for the destination node in the unmodified Trie using linear search.

Fig. 1.1 Searching in elevator-stairs algorithm



Lookup time in the k-level of the PATRICIA is of the order O (w/k).

- Total search time of the algorithm is of the order O (w/k+k)
- Some of the problems in the Elevator-Stairs Algorithm are

• Increased Lookup time due to Linear search at the kth level

- Lookup time in the k-level of the PATRICIA is of the order O (w/k).
- Total search time of the algorithm is of the order O (w/k+k)

3.0 Modified Elevator Stairs Algorithm

General Description

Traditional technique involves the usage of Elevator-Stairs Algorithm for searching the IP Address in the IP Lookup Table. The proposed system is a modified version of Elevator-Stairs Algorithm by altering the searching technique. Binary search is replaced by the bi-directional linear search at the kth-level Trie instead of linear search in the Elevator- Stairs Algorithm. So it drastically reduces the lookup time of an IP address in the IP Lookup table. The searching is initiated by matching the longest matching prefix in the kth-level Trie, which is build from the Trie by adding non-branching nodes.

In Elevator-Stairs Algorithm Searching in the k^{th}-level Trie is done using Linear Search. So the Lookup time in the K-level of the PATRICIA is in the order $O(w/k)$ and the total search time of the Algorithm is of the order $O(w/k+k)$, where w is the maximum length of the IP Address and k is a constant value between 1 and w. In the Case of Modified Elevator-Stairs Algorithm searching for the Best Matching prefix is done using Binary Search at the kth-level of the Trie. So, this proposed technique algorithm provides a better lookup time when compared to the existing algorithm of order $O(w/k+k)$.

Two cases of LPM search is shown in the Fig. 1.2, one at level 3k and the other at level $2k+2$ from the root node.

In the former case, the search finds successive matches in the hash tables at levels k,$2k$, and $3k$, but fails to find a match at level $4k$ and also fails to traverse the Trie. Thus, an inference is drawn that the LPM for the IP address is in level 3k and subsequently the NHP information is retrieved from the corresponding node in the hash table.

In the latter case, the search finds a match in hash tables at levels k and $2k$, but fails to find a match in level $3k$. Subsequently the search traverses the Trie starting from the corresponding node in level $2k$, and finds a match at a node in level $2k+2$. Thus, in the worst case, the search goes through W/k levels of the k^{th}-level-tree and $k-1$ node traversals in the Trie.

This System consists of three modules namely Search, Insertion and Deletion. The search module finds exact match for the IP address. Insertion module, searches the location where to be inserted and IP address is inserted. Similarly, Deletion calls the Search modules and then sweeps out the IP address.

Search

The search module uses the Trie and kth-level Trie to accomplish the search operation

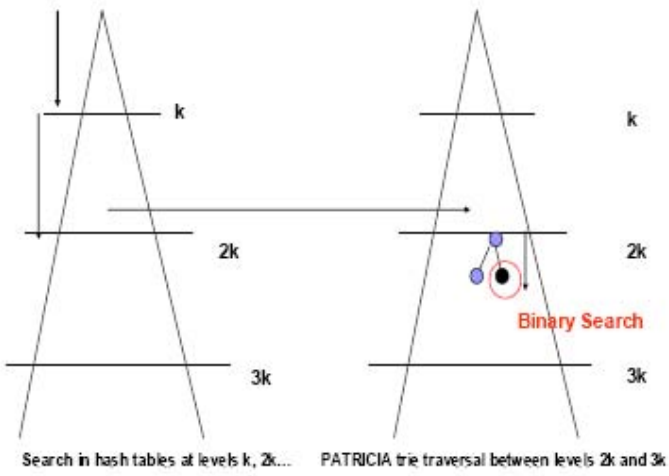
- Trie
  a) Each edge is labeled with exactly one bit.
  b) For any node, edges connected to its child nodes have distinct labels.

c) Every string is mapped to some node in the tree such that the concatenation of the bits on the path form toot to exactly spell out, and every leaf of the tree are mapped to by some string.

- **Kth-Level Trie**

Procedure Build kth-level-tree (v)
Create an empty hash table H (v)
For each edge that crosses level k form root
Add a non-branching node at level k
For each node u at level k
Path =path label between v and u
p=build kth-level tree (u)
Insert p into H (v) with key=path

Fig. 1.2 Searching in Modified Elevator Stairs Algorithm



In this function, a non-branching node is added to each level that crosses from the root node in Trie. The non-branching node at various levels indirectly represents the length of the IP address at each level. These non-branching nodes are recorded in a hash table at each of the root node. Non-branching node acts index for jumping various levels of the Trie. A data structure is created so that the search algorithm can jump *k* levels of Trie, where *k* is an integer between *1* and *w*. The algorithm is initially called with the root node as input. Let klevel of a Trie denotes a level (string depth) *ik* for some integer i such that $0 \leq i \leq w/k$.

- Searching in Kth Level Trie

Procedure FindIP(node,p,pos,port)
port=copy of NHP at node
if node represents a leaf in Trie
return (port)
key = p[pos+1....pos+k]

```
if key is in H(node)
v=node corresponding to key in H(node)
if v is no more than k level away from node
return(FindIP(v,p,pos+k,port)
else
e=edge between node and v
if p matches the edge label of e
child = node at the end of e
l =length of e
return(FindIP(v,p,pos+l,k,port)
else
return(port)
else
pnode=node in PATRICIA tree corresponding to node
return(FindIP_PAT(pnode,p,pos,port))
```

The search for longest mactching prefix on the IP address p using kth -level-tree starts at root and set the variable current_port_number to the default port number. The variable current_port_number stores the port number assigned to the longest matching prefix of length less than the string depth of the current node. At string depth of ik in the kth-level-tree, the search algorithm updates the current_port_number to the port number stored in the current node in the kth-level-tree, if the node has a copy of the port number. The search algorithm checks for the node with key *p[ik+1...(i+1)k]* exits. If such a node exists, the lookup mechanism follows the pointer to the node at level *(i+1)k* and continues the search. If such a node does not exits, it depicts the search for the LPM must end in the Trie between levels ik and *(i+1)k*.

- **Searching in Trie**

```
Procedure FindIP_PAT(node,p,pos,port,length)
if a NHP is assigned to the node
port=NHP assigned to the node
if the node is a leaf
return(port)
e = edge of node that starts with p[pos+1]
mid =length/2
while length !=0
if pos+1 >mid
root =p[mid+1]
else
root =p[pos+1]
end while
if no such node exists
return(port)
```

```
if edge label of e matches p
child = node at the end of e
l = length of e return(FindIP_PAT(child,IP,pos+l.port))
else
return(port)
```

The search operation in the Trie starts with the root nodes with some arguments which describe the node and its position. This recursive algorithm searches for the required IP address with the length of the address. It searches for the edge label of the trie with IP address, after finding the match it assigns it as the child node and calls again the same function still it finds the leaf node. the linear search used in the traditional algorithm is replaced by the binary search here. This decreases the search time drastically.

Insertion

```
procedure Insert(p)
Search prefix p in the same way as FindIP
If the search finishes inside an edge
Create a node there
pnode = node where the search ends
If p = path-label (pnode)
pn = NHP assigned to p
Copy pn to appropriate nodes in
kth-level-tree
Else
Attach a leaf for p at pnode
E =new edge between p and pnode
If e crosses a k-level
Create anode at the k-level on e
parent =pnode's parent in
kth-level-tree
knode= node in kth-level-tree
to represent pnode
Insert a pointer to knode in
H (parent)
If H (parent) becomes full
Double the size of H (parent)
Rehash the elements inH (parent)
```

Insertion of an IP address in the routing table starts with the search for the longest match to the IP address being inserted. It is akin to searching the longest matching prefix of IP address except that the prefix may be shorter than bits. When the search is completed, the nodes of kth-level tree and PATRICIA tree that need to be modified are known.

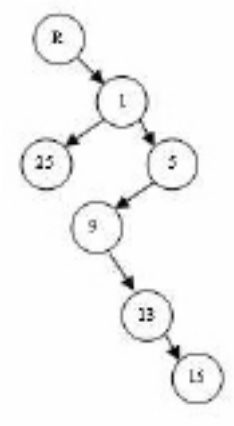
There are two cases of insertion:

a) The prefix that needs to be inserted forms a new leaf in the Trie, or

b) The inserted prefix does not form a new leaf and in that case the structures of the PATRICIA tree and the kth-level-tree do not change.

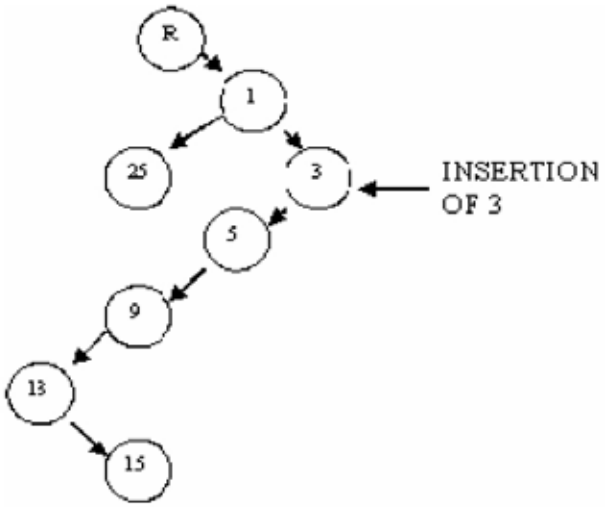
Tree Representation

Fig. 1.3: Tree Before Insertion



The tree shown in Fig. 1.3 is created from the Root position with null string as the key value. The node is inserted into the tree based upon the length of the IP address. Insertion of node is performed based on the binary tree criteria. Fig. 3.3 shows tree updation after insertion.

Fig. 1.4: Tree Updation After Insertion



In case (a), a new leaf node is attached at the point where the search ends. If the edge to the leaf crosses a k-level, a new node is created on this edge at the level in the Trie. A pointer to the new node is added to the hash table at *k(i-1)*level with key *p[ik+1...(i+1)k]*.

In case (b), there are no changes in the structure of the Trie, but the new port number must be copied to the nearest kth-level below the updated node in kth-level-tree.

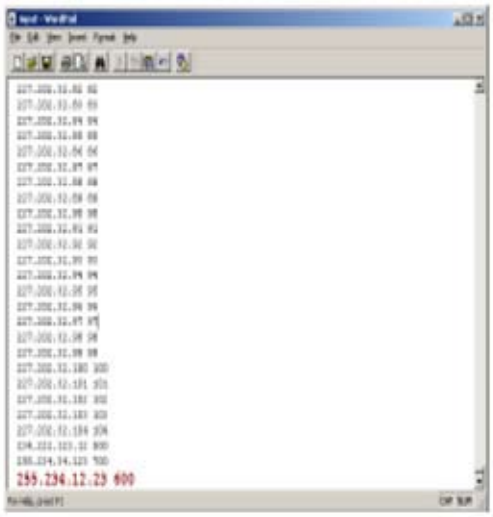
Deletion

```
procedure Delete(p)
Search prefix p in the same way as FindIP
pnode = node matching p
if p = path-label (pnode)
pn = NHP assigned to p's parent
copy pn to appropriate nodes in kth-
level-tree
else
e= the edge to p
delete the leaf for p
if e crosses a k-level
knode = the node on e at the k-level
klnode = node in kth-level-tree
representing knode
parent = pnode's parent in kth-leveltree
Delete knode
Delete the pointer to kpnode in
H(parent)
Delete klnode
Delete pnode
If H(parent) has few entries
Half the size of H(parent)
Rehash the elements in H(parent)
```

Deletion of an IP address in the routing table starts with the search for the longest match to the IP address being inserted. It is the same as searching the longest matching prefix of IP address except that the prefix may be shorter than bits. When the search is completed, the nodes of kth-level-tree and PATRICIA tree that need to be modified are removed.

Database File

Fig 1.5 Database File – input.txt

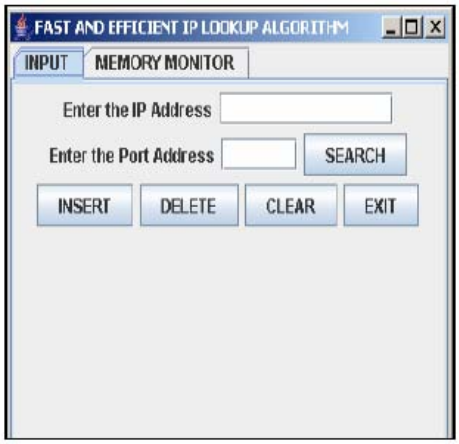


The Fig. 1.5 displays the database window for the system, which contains the IP address and port address.

Fig 1.6 Input Insertions from Database

The window is displayed after the construction of Trie with data form input.txt.

Fig. 1.7: Main Interface



This is the main interface of the system which contains the input panel and the memory monitor panel.

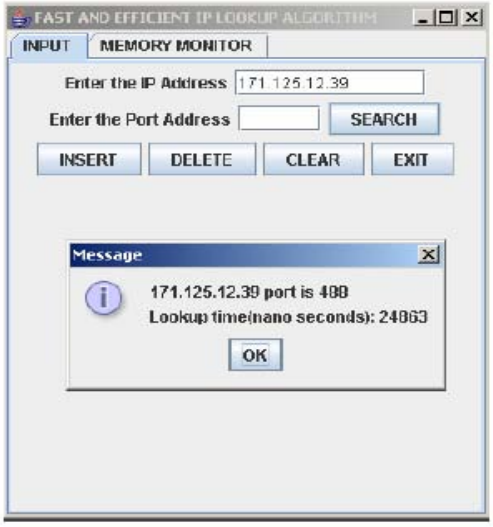
Input

Fig.1.8: Search Module – Input Panel

This shows the input window for the search module, the IP address is entered in the appropriate textbox.

Output

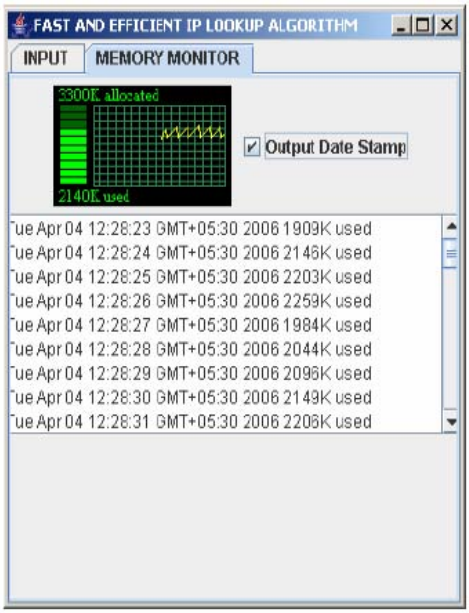
Fig. 1.9 Search Module - Lookup Time



The output of the search module is shown with IP address entered in the apposite textbox. It displays the Port number and the Lookup time respectively.

Memory Monitor

Fig. 5.7: Search Module - Memory



Monitor

This shows the Memory used by the system to search for the IP address specified to the system.

Conclusion and Future Work

The rapid growth of Internet Traffic causes increase in size of routing table. The current day routers are expected to perform longest prefix matching algorithm to forward millions of datagram each second, and this demand on router is increasing even while the prefix search database is expanding in both the dimensions, i.e., IP address length (128 bits for IPv6) and number of prefixes. When there is migration from IPv4 to IPv6, the routing table size increases exponentially. So this system has modified the Elevator-stairs algorithm to reduce lookup latency. This provides an efficient way of searching through Trie and klevel- tree.

The modified Elevator-Stairs algorithm currently applicable for IPv4 address scheme. IPv4 address scheme consists of address space of 32 bits. So, the size of routing table is upto the maximum of 232 entries. This is itself constituted to increase in size of the routing table. The algorithm can be further modified to support to IPv6, which is 128 bit size address space. This Algorithm can be used to enhance the searching technique in Mobile IP Lookups, which builds an efficient and effective Wireless LAN.

References

D. Morrison, 'PATRICIA—Practical Algorithm To Retrieve Information Coded In Alphanumeric (Oct. 1968)' J. ACM, vol. 15, no. 4, pp. 514–534.

S. Nilsson and G. Karlsson (Jun. 1999) 'IP address lookup using LC-Tries,' IEEE J. Sel. Areas Commun., vol. 17, no. 6, pp. 1083–1092,.

Rama sangireddy, Natsuhiko Futamura, Srinivas Aluru and Arun K.Somani (Aug 2005).'Scalable, Memory Efficient, High-Speed IP Lookup Algorithms', IEEE/ACM Transactions on Networking, Vol 13, NO.4,.

M. A. Ruiz-Sanchez, E.W. Biersack, and W. Dabbous (Mar.–Apr. 2001) 'Survey and taxonomy of IP address lookup algorithms,' IEEE Network, vol. 15, no. 2, pp. 8–23.

V. Srinivasan and G. Varghese (Jun. 1998) 'Fast address lookups using controlled prefix expansion,' in Proc. ACM SIGMETRICS, , pp. 1–11.

M. Waldvogel, G. Varghese, J. Turner, and B. Plattner (Oct. 1997) 'Scalable high speed IP routing lookups,' in Proc. ACMSIGCOMM, vol. 27, pp. 25–36.

Chuang Lin, Weidong Liu & Jinpeng Jia , 'A Fast Two-Way Algorithm for IP Lookup' in the Proc. of ICCNMC 2003

10
Towards Human Language Technologies for Under-resourced languages

Jackson Muhirwe

Of the over 6000 languages in the world, only a few have the resources for developing human language technologies. Human language technologies are readily available for most languages of the developed nations. Under-resourced languages, which are the majority of the world languages, have not attracted much attention from researchers and donors due to economical and political reasons. We present strategies for improving the human language technologies for under-resourced languages.

1. Introduction

1.1. Background

There are over 6000 languages in the world (Gordon, 2005). Few of these languages have human language technologies. Berment (2004) presented a metric for measuring the availability of language technologies for different languages. The majority of the languages with low scores are languages from the least developed countries of the world. Different researchers have given different names to these languages depending on their scope of coverage. The term minority languages has been used to languages in a region where there is a dominant language. In this case you find that most research and funding goes towards the dominant language, leaving the minority languages to suffer. A minority language is not necessary a minority everywhere. This is the case for Somali, a minority language in England but a dominant language in Somalia. Another commonly used term is "less documented languages". Here the focus is on the availability of written resources. There are very many languages around the world without any written resources. The only documentation you may be able to find is the Holy Bible translated in that language. There are also other languages without any single written resource, for example the languages for the pygmies who live in the impenetrable Forests of central Africa. The term "Under-resourced languages" is quite often found in literature referring to availability of computational resources for the languages. Under-resourced languages are characterized by: little or no information technology available, no substantial presence on the Internet, existing software has not been adapted for their use (Berment, 2004). This denition suits most or all the languages of the less developed countries of the world. These languages are also referred to as pi-languages or poorly equipped languages.

Languages are birthed, languages grow and languages die. The ethnologue of world languages (Gordon, 2005) lists living languages and gives a number of dead languages by country and by continent. This means that if there are no efforts put in place to preserve languages, so many less documented will disappear slowly slowly by the end of this century. There are many factors trigger this, but that is beyond the scope of this paper. Our focus in this is discussing strategies for improving on language technology for under-resourced languages.

1.2. The cause for under-resourced languages

Doing Human language technology involves lots funds. Due to this, most of the research which has been sponsored around the world has focused on the languages of the developed nations, English, French, Germany, Japanese and other languages that are economically or and politically important. This has left out languages of the less developed nations. The poor governments of the less developed nations usually have many priorities of national interests like fighting famine and diseases. There is lack of human experts in the target language and language engineering in general To add an insult to injury, there's also lack of written text for some languages and where it is available, it is not in electronic form. Due to all this, a huge language technology divide has been created between the languages of the developed countries and those of the less developed countries. According to Dhonnchadha et al.,(2003) languages must endeavor to keep up with and avail of language technology advances if they are to prosper in the modern world.

Our main goal in this paper is to present strategies for improving on human language technologies for the under-resourced languages. The remainder of this paper is as follows. We will discuss some of the success stories around Africa and then later we look at strategies for improving language technology resources.

2. Success stories

However much we have mentioned that there has not been any language technology for most languages in the less developed countries, there are some few exceptions. Here below I discuss some of the success stories in sub-Saharan Africa.

2.1. Swahili.

Swahili or Kiswahili as called by the native speakers in Eastern Africa is so far the most successfully researched indigenous African language. Why? Most of the research that has been done on Swahili was done at the University of Dar Es-Salaam and later in collaboration with the University of Helsinki. There was a political will spearheaded by the then President of Tanzania, Mwalimu Julius Nyerere to unite the republic of Tanzania by developing Swahili as the national and official language. To a great extent Voluntarism. Most of the human language language technology that has been done so far for Swahili was either done by Arvi Hurskainen or has its roots on his work. We need more experts who are willing to put in their time and effort to develop and avail resources for other under-resourced languages

2.2. South African languages

After the decline of Apartheid in South Africa, the indigenous South African languages got some attention from the government and the researchers. It is obvious why there has been an increase in the study of indigenous South African languages, which had been neglected like all the other Bantu languages. A South African Bantu language project was commissioned in 2001 to study and develop tools for all the indigenous national languages and several tools have been developed so far (Bosch, 2006). This project has concentrated on developing morphological analysers and machine-readable text resources that are needed for higher applications. See --- for the status of the project. All the tools, which have been developed, are using the existing approaches, which are not necessarily adequate for Bantu languages. All the tools that have been developed are based on the finite state technology. This approach has successfully been applied to a wide a range of languages around the world.

2.3. Why the Interest in under-resourced languages?

1. *Culture Breakdown*. We know that languages represent the culture and diversity of different people around the world. Unavailability of language resources eventually leads to extinction. Failing to salvage the language will lead to extinction of the culture and consequently the people.
2. *Testing*. Due to the morphological complexity of under-resourced languages, they provide a good ground for testing of new approaches and technologies.
3. *Economical development*. The governments of the most least developed countries have identified computer technologies as the main engine for their development. In order for this to take place, there's need to develop software in indigenous languages.
4. *Reducing the language technology divide*. As we mentioned above, a very huge language technology divide has been created between the languages of the developed nations and those of the less developed. According to Dhonnchadha et al.,(2003) languages must endeavor to keep up with and avail of language technology advances if they are to prosper in the modern world"

3. Strategies for Improving Human Language technologies for Under-resourced languages

Here present strategies that could be taken by any group or individuals who are interested in developing language technologies for under-resourced languages

1. *Cooperation* – How African languages in general can succeed.

   There's need for cooperation for all researchers and institutes involved in doing research on African languages. This will help avoid reduplication and wastage of efforts and resources, resources will be directed towards the most needful areas. There's need for African governments to wake up and realize, that they to savage their languages, otherwise their culture be history.

2. ***Building machine-readable lexicon.*** The most time consuming task in language engineering is developing and maintaining a machine readable lexicon (Trosterud, 2006). The machine-readable lexicon also happens to be one of two most indispensable parts of any language tool the other being a morphological parser. Due to the importance of the lexicon to any language project and its future extensions it is imperative that the project should start with building of a machine-readable lexicon. The lexicon could be stored using XML and scripts can be written to be used to extract all the needed information from the lexicon. This information could be used in the construction of paper and electronic dictionaries, developing of high end user language tools like spell checkers, information retrieval systems, speech recognition systems etc.
3. ***Continuity focused.*** In this I mean the philosophy of the project should to build for the future. Here you don't focus on just getting results, but you also look will other be able to continue from what I have done. In many cases you find that in the whole world this project is unique and because of this uniqueness you don't want your efforts to end with you. If it is software it should be made in such a way that others can easily come and extend your work to another level.
4. ***Open source.*** Open source can easily turn a failing project to a successful project. One of the problems we cite above is the lack of finances due to poverty and priority areas. Make sure that when ever possible you use only free and open source software and cut down on the financial issues.
5. ***Documentation.*** Documentation will also greatly help in the continuation of the project when new people take up the project. The documentation could be in terms of manuals both soft and hard copies. Some of the task of the project could be documented on website since other people who may be interested in the work may access your project online and start working with you on some of the issues where they can of help.

4. Conclusion

In this paper we have presented strategies for developing language technologies for under-resourced languages. We have seen that there's a substantial need to salvage all languages of the world. In this paper we have identified 5 key indicators to language technology growth: Cooperation; building machine-readable lexicon; open source; documentation. This list is not conclusive, there is need to do more research in the area in identifying other strategies for language technologies. Some of the strategies could be specific to region or could across all regions irrespective of the economic status.

References

Berment V., (2004). "Méthodes pour informatiser des langues et des groupes de langues peu dotées" PhD Thesis, J. Fourier University – Grenoble I, May 2004.

Besacier, L., Le, V-B., Boitet, C., Berment, V. (2006) ASR AND TRANSLATION FOR UNDER-RESOURCED LANGUAGES . Proceedings IEEE ICASSP 2006. Toulouse,France. May 2006.

Bosch, S.E (2006) Computational Morphological Analysis Project

Bosch, S.E and Pretorius, L. (2002). The significance of computational morphological analysis for Zulu lexicography, in South African Journal of African Languages, 2002, 22.1:11-20.

Dhonnchadha U. E., Cahilfhionn N.P, Genabith, J.V. (2003) Design, implementation and evaluation of an inflectional morphology finite state transducers for Irish. Machine Translation, Springer Vol 18 Number 3 pgs 173-193.

Gordon, R.G.J., (ed.),(2005). Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International.

Hurskainen, A. (1992). A two level computer formalism for the analysis of Bantu Morphology an application to Swahili Nordic journal of African studies 1(1): 87-119 (1992)

Hurskainen, A. (1999). SALAMA Swahili Language Manager Nordic journal of African studies 8(2): 139-157

Pretorius, L. and Bosch, S. (2003). Towards technologically enabling the indigenous languages of South Africa: the central role of computational morphology. Interactions of the Association for Computing Machinery 10 (2) (Special Issue: HCI in the developing world): pp.56-63.

Trosterud T., (2006) Grammar-based Language Technology for the Sámi Languages

Part 3

Information Technology

21
A Framework for Adaptive Educational Modeling: A Generic Process

P. Mahanti S. Chaudhury and S. Joardar

The present paper deals with the education system with the approach of facing challenges of globalization through e-education, and discusses possibilities of using existing IT developments in the field of education so as to enable it to evolve new paradigms of developmental education. Our aim is to briefly describe an adaptive and systematic method so that the educational requirements can be translated into a system for generation, monitoring and re-adaptation of teaching and student's model.

1. Introduction

The education system generally has a form of mass education, is institutionalized and decided by academia. It is therefore teacher centric. The education is localized in the sense that it is available at the local or nearby institution with all its advantage and disadvantages. Students have to move to other institutions, and even go abroad for further and higher education.

The single and dual mode universities as well as conventional universities are now using ICT (Information Communication Technologies) for various purposes. This has created a new scenario of modes of education which can be classified as follows:

1. Formal education: Classroom/ Campus based education imparted by traditional universities.
2. Non-formal-Open and Distance Education: Offered by single mode open universities.
3. Mixed Mode Education: offered by Distance Education Institutions of traditional universities by using both formal and non formal components of the two modes.
4. ICT Based Convergent Mode: Uses Web Based Education, Computer based education, center/classroom Based education.
5. . Entirely WBE E-education: use Internet and the Web extensively so that teaching and learning is almost distributed

It should be noted that learning process always take place in the cognition of an individual; and is dependent on the psychomotor and effective development of an individual being educated. Education is therefore very personal process of learning.

One of the key issues is, therefore, to identify the processes and methods of education that are mode and technology independent. They could then be followed in the emerging and unknown scenario. In the fast changing scenario covering all aspects of socio economic and cultural changes, attempt should be made to identify invariant educational processes and practices that will support education and relate it to support developments in living and working places and people in a self-sustainable way.

Our aim is to create educational programs that could

- Make teaching and learning possible from
  Anywhere, Anytime.
- Link education- learning with life and work
  related processes and places
- Create a network of educational content and services, which can flow in the network and support the processes of education-learning, teaching and evaluating- anywhere anytime and;
- Enable educators and educational institutions to create new paradigms of education dependent on various developmental processes and models.

There are mainly two mega paradigm shifts in education. The first is from traditional university to open and distance education (ODE) and the second is from ODE to E-Education. The new paradigm of e-education is however of a non-industrial form and should offer personalized education on a mass scale (Mass Personalization). Paradigm shift in education is essentially a learning teaching evaluation as shown in Table.1

Table:1 Learning Teaching Evaluation

Learning- Teaching-Evaluation		
From Teaching From Classroom		Distributed and group Teaching Distributed classroom
From learning from a teacher		Learning from Resources, group of teachers/ experts and through Interactivities
From Content Learning		Objectives and Outcome Oriented Learning
From Course Content		Granulated Object Based Content forming Meta Database
From examinations		Continuous Formative and Summative Evaluation
Educational Management		

From education From whole time Education		Development Education Just-In-Time Education
From Campus Education From Campus Environment		Distributed Education Virtual Educational Environment
From a Single Institution		Consortia of Institutions/Distributed Institutions/Virtual Organizations
From Mass Education		Personalized Mass Education

2. E- Learning

E-learning is becoming the de-facto standard for education these days. It makes the learning environment open for implementing pedagogical innovations. The target in majority of cases is to make education independent of time and place, tools and technology constraints so as to optimize the performance of participants through individualized education. To realize the potentials of E-learning, there is a need for a systematic software development approach, because lack of a systematic approach can result in poor e-learning quality [12]. In addition the very basis of E-learning is a pedagogical foundation based on learning theories [6,7,10,11,15]. That is to say, progress in E-learning will come from a better understanding of the learning process and not automatically from improved technology [21]. Therefore learning theories must be one of the driving forces behind E-learning development. While a number of existing approaches to E-learning incorporate learning theories, few of them are grounded in software engineering principles. As a result much of the construction of E-learning is still carried out without a true understanding of how learning theories can be translated into pedagogical requirements [8].

E-Education is essentially the same education with the same basic processes of educating, creating, developing and managing which are carried out by individuals, institutions and communities for achieving the goals of education. In the information age it is supported by IT enabled and IT driven processes and made accessible through IT tools and techniques to make education globalized, localized and personalized.

E-Education system requires the following framework and infrastructure:

- Network with latest hardware and technologies and grid architecture giving network access to anyone, anywhere, anytime.
- Software tools and techniques that enable creation of databases and information flows, offer facilities to learners, teachers and institutions to receive/give personalized education on mass scale.
- Content in e-formats on a knowledge grid that enables teachers and students to get personalized curriculum of high quality, relevance and utility.
- Educational delivery system that ensures quality and developmental relevance of educational offerings for individual, institutions and community.

- Quality assurance and Certification Mechanism to maintain competitively high and acceptable standards at international level.

If these principles could be incorporated in the design and development of IT enabled and IT driven process of social mobilization and organizations, the nature and character of the emerging trends could be different. Such a system can promote culture of participatory democratic decentralization, accountability and local relevance and help in efforts for overall development of a community.

3. The need For Agile Models

Adaptivity is of paramount importance for educational applications on WWW as they are expected to be used by very different classes of users without any assistance by a real teacher [17]. Usually, we distinguish two basic categories of educational applications: those that complement and support the tasks of the traditional teacher based classroom, and those that function independently, providing self learning environment [20]. In both these cases the role of internet is important as it provides easy flow of communication and contains a large amount of information.

In the context of this paper Agile Models are particularly attractive for attempts to model the changing scenario of education. It is seen that any educational model has to take into account class colour, customs, ethnicity, gender, language, moral codes, power-distance, religion, social conventions and many more apart from the teaching content. How these and other elements impact on educational practice-curricula, the nature and role of the teacher and learner etc. – is ill defined as there is no one model that can be used to compare and relate such issues to educational practice [16].

Moreover, the linear sequential model known as the waterfall model is not flexible enough to be applied to e-Learning, because it does not deal with evolution, change, and feedback to previous steps. But, the waterfall model is interesting from the management point of view, since it can help developers plan everything from the very beginning [18,19]. The spiral model, which also addresses the entire development process, is relatively complicated and difficult to manage in order to be applied to e-Learning, but it can help developers, in particular during the analysis phase, to minimize risks by focusing on what really matters.

As a result, previously developed system development process models do not fit the specifics of e-Learning. Hence, it may be necessary to combine the advantages of pre-existing models, and eventually modify them, to meet the specific requirements of e-Learning.

Given the evolutionary character of e-Learning, it seems that the most suitable process model for e-Learning is the Agile model, which modifies an early prototype until it provides all required features. The model is suitable for e-Learning, because the process often involves feedback to earlier phases. This model is however not without problems since e-Learning systems are evolving constantly. It is thus difficult to determine when they are going to end. Finally, the model must rely

on reusability since the reuse of learning objects and components is a necessary option for e-Learning.

As a result, it seems that the most suitable process model for e-Learning is a Agile model that includes some important aspects of other process models. Thus, the key issues of developing e-Learning are:

- First, e-Learning needs a structured evolutionary process model with feedback to previous phases in order to deal with change and evolution.
- Second, e-Learning needs to reuse learning objects, so that developers are not forced to start over again when they design e-Learning for new courses.
- Third, particular attention must be placed on the analysis of the teaching and learning environment, that is to say the scope of the system, at an early stage, since it is of crucial importance to understand environmental factors that affect e-Learning.
- Finally, the model must include an evaluation phase to ensure that pedagogical principles and learning issues are kept in mind [8].

4. Agile Modeling

Agile modeling (AM) is proposed by Scott Ambler [1]. It is a method based on values principles and practices that focuses on modeling and documentation software. AM recognizes that modeling is a critical activity for a project success and addresses how to model in an effective and agile manner [2].

The three main goals of AM are [2]

1. To define and show how to put into practice a collection of values, principles practices that lead to effective and lightweight modeling
2. To address the issue on how to apply modeling techniques on agile software development process
3. To address how one can apply effective modeling techniques independently on software process in use.

It is important to note that AM is not a complete software development method. Instead, it focuses only on documentation and modeling and can be used with any software development process. One can start with a base process and tailor it to use AM.

5. Adaptive Model

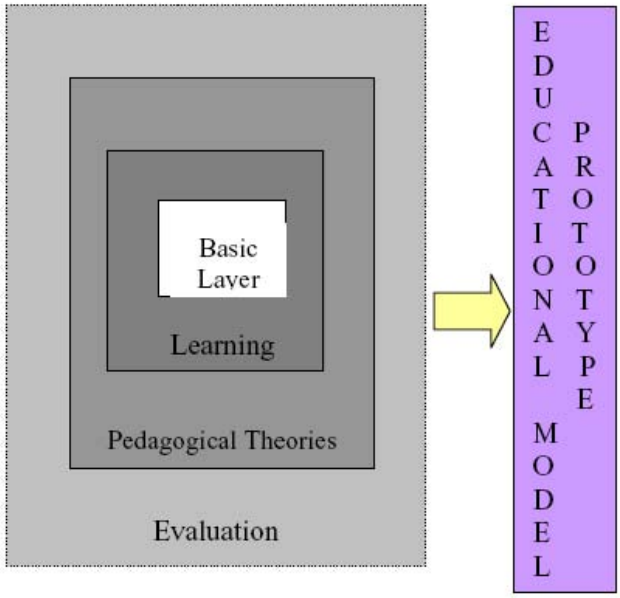
The educational model is the description of theories, principles and processes that aim at standardizing teaching processes and regrouping methods. Our objective in this section has been to describe this process in a nutshell in order to make clear the particular character of this model and the need of specialized modeling using suitable stereotypes in this particular case.

Further, use-case modeling may be adopted in educational modeling. As Arlow

& Neustadt (2002) point out, a use –case is a "description of a sequence of actions that a system performs to yield value to a user" (p.15). Modeling and collecting use-cases has proven useful in other instances of educational modeling. The work here, however, addresses a preceding stage in which a domain model is constructed independently from services provided by a system.

This section outlines an adaptive educational model. It consists of four layers which influence each other by exchanging information. UML does not provide a good visual model to define the scope of the system hence whiteboard, normal diagrams are used here to describe the task.

Fig 1. The processes of definition of an Educational Model.

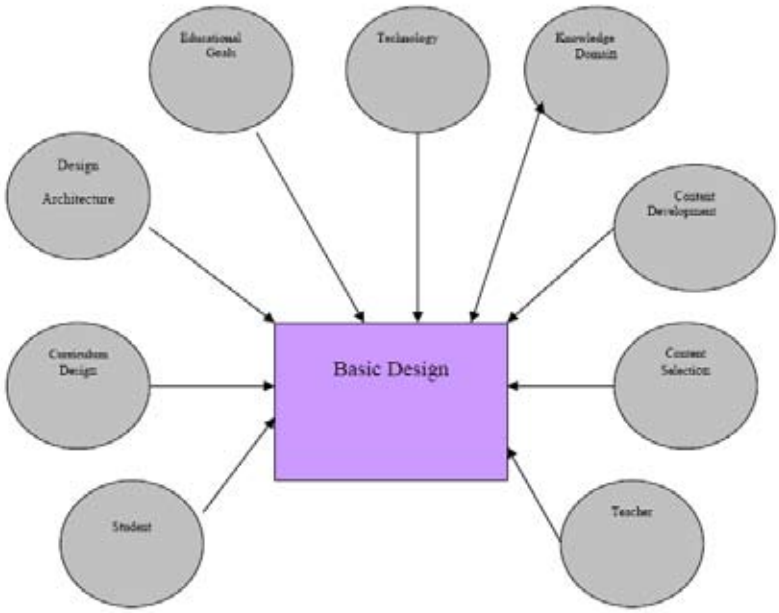


As shown in Fig.1 that each layer represents a different set of design decisions and foci. By functionally separating design decisions into the categories of basic layer, learning theories, pedagogical theories and evaluation, we decouple the ways in which learners interact with the content, teachers and each other, and how they develop to their maximum potential in this environment. In these layers it is necessary to consider issues such as learning theories, pedagogy etc. all of which help to ensure a learner-centered active learning environment on top of basic technology [16].

1. **Basic Layer:** The basic layer focuses on selection and implementation of technology, finding out the design architecture, content development, content selection, curriculum design, student and teacher interface, development of a knowledge domain and prescribing the proper educational goals. We called this a basic layer as this serves as the base model towards agile development.

The knowledge domain may change depending upon the changing scenario and evolution of new concepts but the base model remains the same. The components of basic layer is shown in Figure.2.

Fig 2: Components of the Basic Layer



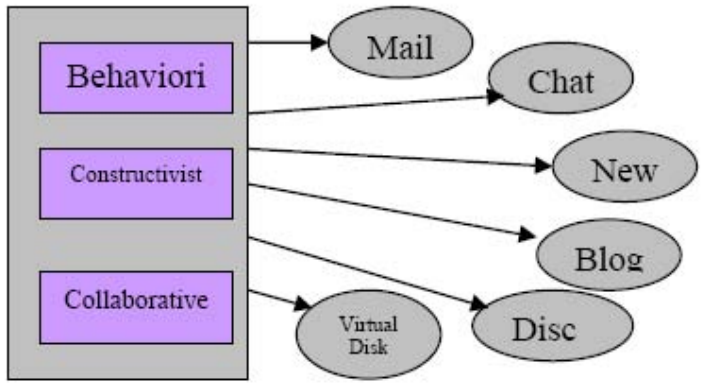
2. **Learning Theories:** The scope of this layer is to identify models that will guide the ways in which users interact with students, staff and component: process facilitation design- mentoring models, use of discussions, e-mails, chat, blog etc.

   Learning theories can be related to three main models [8].

   - Behaviorist- Suitable for novice learners and instructors are central to learning activities. This promotes stability and certainty of knowledge acquisition but the learners are not often free to express their own ideas.
   - Constructivist- It is a constructed entity made by each and every learner through a learning process based upon prior knowledge. The role of the teachers is to serve as guides and facilitators of learning.
   - Collaborative- In this type learning emerges through interaction of learners with other people, instructors and fellow learners.

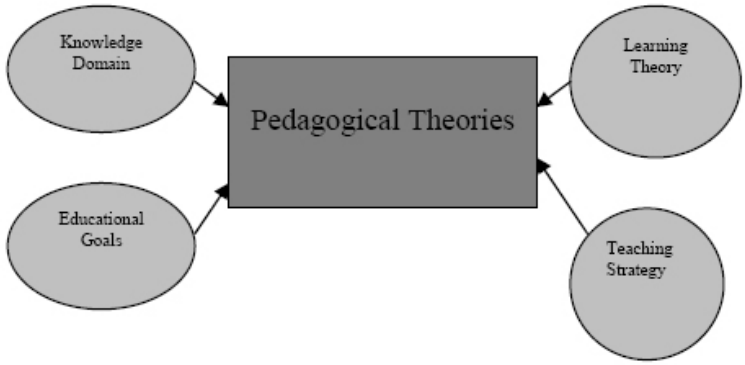
This entire layer contributes to the knowledge domain constructed in the basic layer and it becomes a regenerative process and the components of learning theories as shown in Fig.3 .

Fig 3: Components of Learning Theories



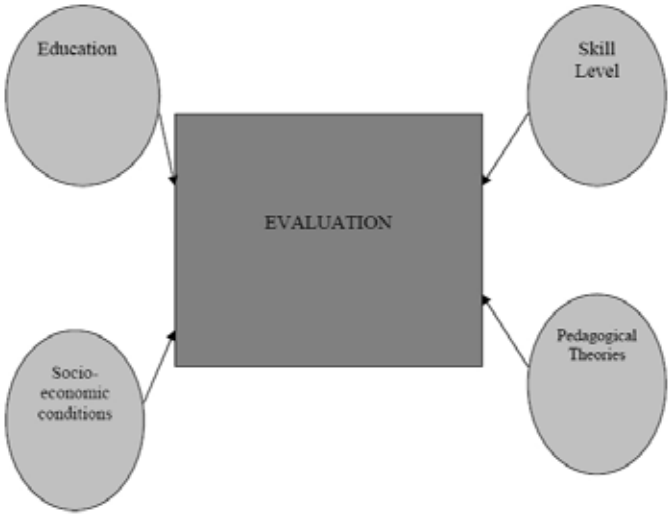
3. **Pedagogical Theories:** As shown in Fig.4, the focus of this layer is on the elements that contribute to the learner's optimal cognitive development. This involves knowledge based pedagogical theories and rule based extraction mechanism. The aim is to combine learning theories with teaching strategies and data extracted from the repository. Specific educational goals as well can be blended in this layer to formulate strategies adapted to a specific student and a specific learning object.

Fig 4: Components of Pedagogical Theories



4. **Evaluation:** The key idea contained in this layer is to develop an assessment protocol that allows for a diverse range of cultural and educational responses. The evaluation process should take into account the knowledge and skill level of the student, his needs and motivation, his personal learning style, customs, ethnicity, gender, language, moral codes, social conventions apart from the educational content. This layer thus ensures that pedagogical principles and learning issues are kept in mind. This itself is an agile process as the environment of a learner, course content, technology used etc are all evolutionary and are constantly changing as shown in Fig.5.

Fig 5: The Evaluation process

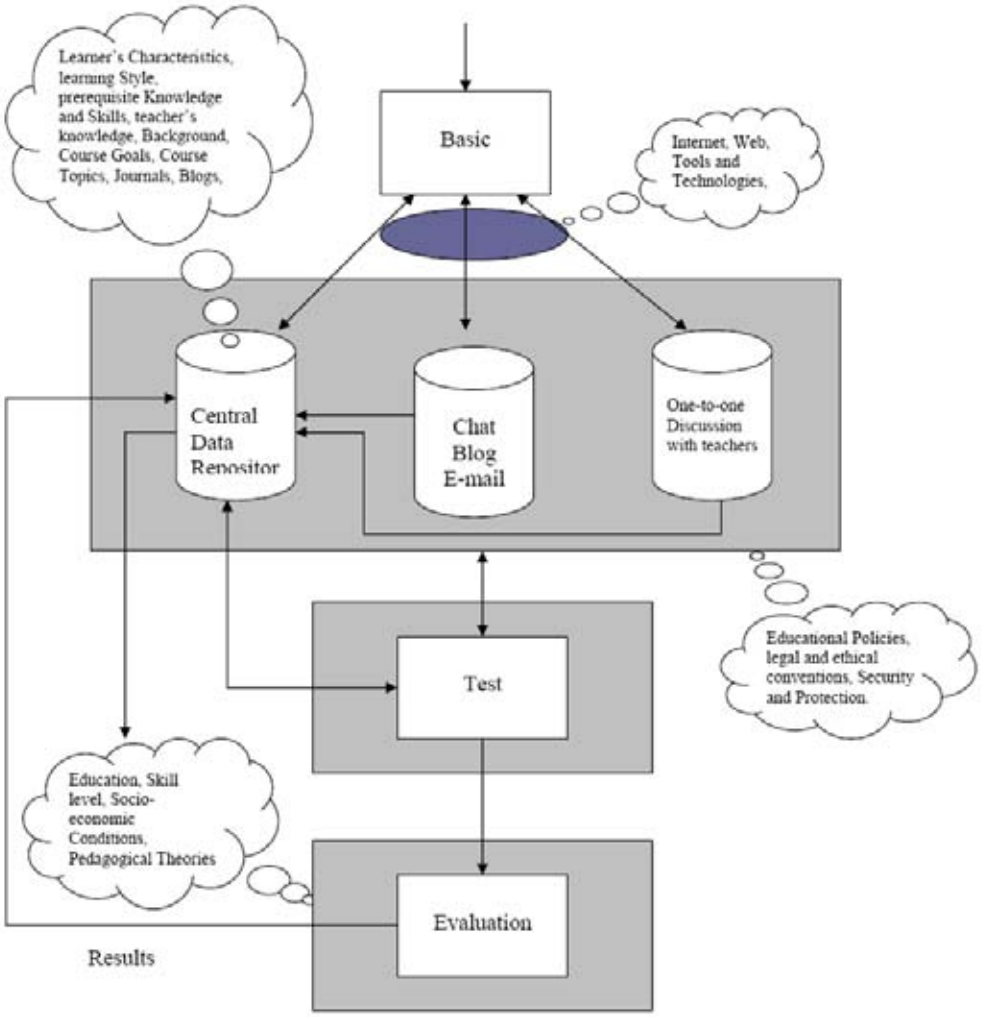


6.0 Development Process Model:

In this approach the pre-existing process models have been modified to integrate the educational context and associated teaching and learning environment at an early stage into the system scope. The approach also incorporates learning theories and pedagogical theories as they are major contributing forces behind designing, developing and evaluating E-learning.

Since agile methods are used in this modeling approach which prescribes the use of a base model. We have also considered a base model which consists of the basic requirements of an E-educational environment like teachers requirement, learners' requirement, pedagogical requirements, technological requirements and institutional requirements. Fig. 6 below explains certain features of the educational model, with arrows showing its adaptive and iterative nature. UML is used to show the requirements specification, though whiteboard, CASE-tools or even DFDs can be used.

Fig 6. System Development Model



6.1 Design Principles and Considerations

The above system is an integrated E-learning solution intended to satisfy the needs of university education. The main goals are:

- To develop a platform for authoring and delivery of web-based courses
- To design and implement an open, platform independent set of tools, automating the whole process of courseware authoring
- To provide web-based courses with a common delivery framework incorporating user and system management, course and curriculum management, different kinds of communications, and student performance assessment tools.

- To make the process iterative and adaptive giving feedback to the earlier processes.

6.2 Functionality:

This section describes the main modules (packages) and their relationships, thus presenting the system architecture.

Student – any person that uses learning resources to gain knowledge or skills, can

- Log into the system by entering their user name and password;
- Modify their personal data;
- Search and browse registered courses and display general information about them;
- Enroll in the courses announced as publicly available and receive a notification for registration;
- Browse course materials for all courses which he/she is enrolled in;
- Receive information and participate in evaluation of his/her performance;
- Upload and download materials to/from the shared workspace associated with the course;
- Communicate with other participants through the communication facilities defined for the course.

Course Author – is responsible for creation and modification of courses according to a pre-defined learning goals. He/she is allowed to:

- Create and modify a course meta-data record according to the existing standards;
- Describe the course structure;
- Provide course materials by uploading resources;

Instructor – supports *Students* during the learning process. Tasks performed by the instructor include creation of a *course instance* (course customization according to specific educational needs of students with a given profile) from a given course, and defining and tuning course instance environment. Instructors are allowed to:

- Define the navigation strategy for the course by using a tool provided by the system;
- Choose a template for a visual course instance representation;
- Describe the set of communication facilities available in the course instance;
- Define the course instance schedule;
- Set up the evaluation procedures - create and add tests, define assignments and other assessment objects, that are not included in the initial course

description, review the assignments results, and form the final mark for the students;
- Exchange messages with other course participants;
- Generate statistics for the students progress;

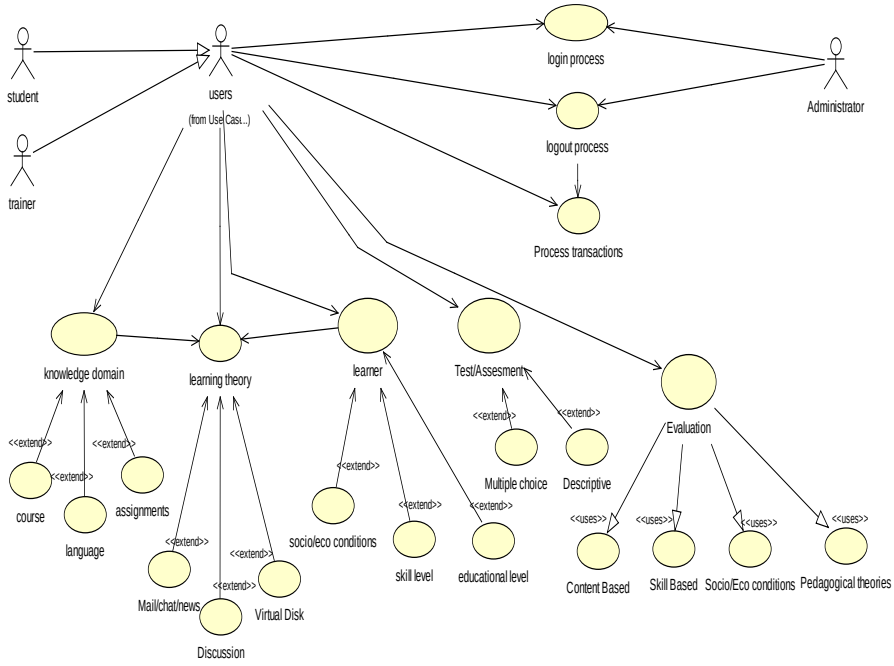
Course Administrator – manages courses, course instances and curriculum. A course contains the main learning materials for the course instances based on it. Different course instances may share one course as a source for learning resources. Curriculum is a cross table between course instances, student groups and instructors. The course administrator is responsible for:
- Registering courses and course instances;
- Deleting courses and course instances from the system;
- Defining access mode for a course instance as open or limited (a fee is required);
- Enrolling students and groups of students in a course instance;
- Removing students and groups from a course instance;
- Creating and managing the curriculum.
- Managing student groups and assigning an instructor to each group.
- Sending messages to other users via the internal mail.

System Administrator – takes full control over the system, manages system resources such as user accounts and groups of users, assigns permissions, and defines settings for the system. He/she also monitors event logs and maintains the system in secure and stable state. He/she has rights to:
- Add and modify personal data and delete users;
- Create, edit and delete groups of users – a user may participate in more than one group;
- Assign permissions to users and groups of users for every object in the system (e.g. course, course instance, test, shared/personal space, chat, discussion board, etc.);
- Manage event log files, determine which events to be logged, view and delete event logs;
- Send messages to other users via the internal mail.

6.3 System Architecture

In order to implement the functions described in the previous section, to be able to add new ones and to improve the existing ones, the system needs to have stable and extensible architecture. Realization of this task was greatly facilitated by use of UML and the chosen methodology of development. Fig-7 describes briefly the UML architecture using a Use-Case diagram.

Fig 7: Functionality for different users



7. Conclusion and Future Work

E-learning thus is an adaptive process and should be constantly evolving so as to deliver correct, competent, relevant and timely information available online. E-learning can be influenced or can obtain information from teachers, learners, learning theories and information technology. The approach is thus iterative, incremental and adaptive which should deal with change and adaptation of various aspects of E-learning at different levels.

The power of this model is not in its ability to prescribe a particular instructional design but rather, to keep instructional design process focusing on designing for learners. The focus on sensitive evaluation is an important element of the model.

A fundamental problem in this type of modeling is that educators often do not have the requisite technical knowledge to bridge the gap between their educational understandings, desires and intents for learning environments, and those of technically savvy courseware developers. There is a urgent need to find out ways to bridge this knowledge and technology divide.

Agile methods are there to stay and probably will not win over traditional methods but live in symbiosis with them; traditional software engineering can be enriched paying attention to new ideas springing up from the field. One important factor when selecting a development method is the number of people involved. The more people involved in the project and the more it grows larger the less

agile the methods become. In many cases, being both agile and stable at the same time will be necessary. A contradictory combination, it seems, and therefore extra challenging, but not impossible.

We firmly believe that agility is necessary but that it should be built on top of an appropriately matured process foundation, not instead of it.

References

Ambler, S. W. (2001a, May 01). *When is a model agile? Part 1: What agile models should accomplish*. Retrieved from http://www-106.ibm.com/developerworks/webservices/library/co-tipam2.html

Ambler, S. W. (2001b, May 01). *When is a model agile? Part 2: Traits agile models should have*. Retrieved from http://www- 106.ibm.com/developerworks/webservices/library/co-tipam3.html

Arlow, J.,& Neustadt, I.(2002). UML and the Unified Process, London: Pearson Education.

Booch,G.,Jacobson, I.,& Rumbaugh, J.(1998) Unified Modeling language user guide, Boston, MA: Addion-Wesley.

Cohen, D., Lindvall, M., Costa, P., agile Software Development (2005).

Conole, G., Dyke, M., Oliver, M., & Seale, J. (2004). Mapping pedagogy and tools for effective learning design. *Computers and Education, 43*(1-2), 17-33. The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

Govindasamy, T. (2002). Successful implementation of e-learning: pedagogical considerations. *The Internet and Higher Education, 4*, 287-299.

Hadjerrouit, S. (2007). Applying a System Development Approach to Translate Educational Requirements into E-Learning. Interdisciplinary Journal of Knowledge and Learning Objects. Vol. 3, pp 107-134

Hadjerrouit, S. (2007). Using an understanding of the learning cycle to build effective e-Learning. In N.A.

Hamid, A. A. (2002). E-Learning: Is it the "e" or the learning that matters? *The Internet and Higher Education,4,* 311-316.

Harasim, L. (2000). Shift happens: Online learning as a new paradigm in learning. *The Internet and Higher Education 3*, 41-61.

Kay, R., & Knaack, L. (2005). Developing learning objects for secondary school students: A multicomponent model. *Interdisciplinary Journal of Knowledge and Learning Objects, 1,* 229-254 http://ijklo.org/Volume1/v1p229-254 Kay_Knaack.pdf

Mahanti, P., Chaudhuri, S. It and its role in Indian education system: An overview , GITMA world Conference, Orlando, Florida. June 2006.

Mayes, J. T., & Fowler, C. J. (1999). Learning technology and usability: A framework for understanding courseware. *Interacting with Computers, 11*(5), 485-497.

Mayes, J. T., & Fowler, C. J. (2005). Mapping theory to practice and practice to tool functionality based on the practitioners' perspective. Retrieved March 28, 2006, from: http://www.jisc.ac.uk/uploaded_documents/Stage%202%20Mapping%20(Version%201).pdf

Nicholson, p., Thalheim, B., (2004) Adapting to changing time and needs.(Culturally adaptive learning objects: Challenges for educators and developers) UNESCO-SEAMEO Conference, Bangkok.

Papasalouros, A, & Retails, S. (2002) Ob-AHEM: A UML-enabled model for adaptive Educational Hypermedia Application. Interactive Educational multimedia, number 4, pp 76-85 http://www.ub.es/multimedia/iem

Powell, T.,A.,(1998). Web Site Engineering: Beyond Web Page Design. Prentice Hall: London.

Pressman, P. (2000). *Software engineering: A practitioner's approach* (5th ed.). New York: McGraw-Hill.

Rokou., F.,P., Rokou, E., Rokos, Y. Modeling Web based educational systems. Process design teaching model (2004). Educational Technology & Society, 7(1),pp 42-50.

Watson, D. M. (2001). Pedagogy before technology. Re-thinking the relationship between ICT and teaching. *Education and Information Technologies, 6*(4), 251-266. Available at http://www.springerlink.com/content/w343nq3375053k23/fulltext.pdf

22

Does Interactive Learning Enhance Education: For Whom, In What Ways and In Which Contexts?

Anthony J. Rodrigues

To get a balanced view we review the body of research during the past three decades that has investigated interactive learning in a variety of forms ranging from the earliest days of mainframe-based computer-assisted instruction to contemporary multimedia learning environments accessible via the World Wide Web. In light of this some researchers believe that we are on the verge of developing a true instructional science whereas others conclude that we simply cannot pile up generalizations fast enough to adapt the interactive designs to the myriad variables inherent in human learning. This paper, summarizes what is know and what is not known about interactive learning, describe the strengths and weaknesses of various approaches to interactive learning research, and concludes by describing a structured evaluation framework for technology mediated learning i.e. design, process and outcome from a technical, human and educational systems perspective respectively based on a development research agenda.

Introduction

The U.K. e-University was launched in 2000. The initial business plans forecast rapid growth to 110,000 students within six years and 250,000 in a decade, with projected profits of more than £110m. An investigation by the Commons education select committee found that studying at the UK e-University, which folded last year six months after the launch of its first courses, cost an average of £44,000 per student - more expensive than going to Oxford or Cambridge.

A government initiative to offer British university degree courses over the internet was condemned by MPs in March 2005 as a "disgraceful waste" of public money after it recruited just 900 students at a cost of £50million. The e-University "blindly" pursued a policy of offering entirely internet-based learning, despite evidence that students preferred to supplement online study with traditional lectures and seminars.

The first lesson one learns is that the development of e-learning needs to be learner-centered than technology driven. We need to learn a lot more about the needs of learners and the form of pedagogy that e-learning involves.

It is necessary to appreciate the importance and complexity of researching interactive learning. More specifically, an outline of what is know about interactive

learning as well as what is not known, and some directions are suggested for further research.

What is meant by interactive learning? Almost every field of inquiry today is beset with dichotomous controversies. In biology it is nature versus nurture and culture. If research on interactive learning can be regarded as a field, then it too has its controversies. Some view it as a branch of science or technology and others regard it as more akin to a type of craft or even art (Clark & Estes, 1998). Nonetheless, any skepticism concerning learning sciences and educational technology does not preclude a strong commitment to development research and evaluation as necessary, but insufficient, methods for collecting information to guide the decisions that must be made when designing (crafting) interactive learning environments.

What is the intrinsic definition of interactive learning? Faced by a history of failed technology-based innovations e.g.,

- programmed instruction,
- teaching machines, and
- computer-assisted instruction), the latest buzzwords for interactive learning e.g.,
- interactive multimedia,
- the World Wide Web (WWW), and
- virtual reality;

attract both enthusiasm (Perelman, 1992) and serious skepticism (Postman, 1995). Ultimately, all learning is interactive in the sense that learners interact with:

- content to process,
- tasks to accomplish, and/or
- problems to solve.

However, in this paper, reference is made,see Reeves 1999, to a specific meaning of interactive learning as involving some sort of technological mediation between a teacher/designer and a learner. From this perspective an interactive learning system requires an electronic device equipped with a microprocessor (e.g., a computer) and at least one human being (a learner). The adult school dropout developing basic literacy skills via a multimedia simulation, the high school student surfing the WWW for archival material about indigenous people to prepare a class presentation, and the three-year old practicing color-matching skills with Big Bird with a Sesame Street CD-ROM program are all engaged in interactive learning.

What is known and what is not known?

There are two major approaches to using interactive learning systems and programs in education, although many of the ideas expressed in this paper may apply within training contexts. First, people can learn "from" interactive learning systems and

programs, and second, they can learn "with" interactive learning tools. Learning "from" interactive learning systems is often referred to in terms such as computer-based instruction (CBI) or integrated learning systems (ILS). Learning "with" interactive software programs, on the other hand, is referred to in terms such as cognitive tools and constructivist learning environments.

The foundation for the use of interactive learning systems as "tutors" (the "from" approach) is "educational communications theory," or the deliberate and intentional act of communicating content to students with the assumption that they will learn something "from" these communications. The instructional processes inherent in the "from" approach to using interactive learning systems can be reduced to four simple steps:

1) exposing learners to messages encoded in media and delivered via an interactive technology,
2) assuming that learners perceive and encode these messages,
3) requiring a response to indicate that messages have been received, and
4) providing feedback as to the adequacy of the response.

The findings concerning the impact of interactive learning systems and programs can be summed up as:

- Computers as tutors
  - have positive effects on learning as measured by standardized achievement tests,
  - are more motivating for students,
  - are accepted by more teachers than other technologies, and
  - are widely supported by administrators, parents, politicians, and the public in general.
- Students are able to complete a given set of educational objectives in less time with CBI than needed in more traditional approaches.
- Limited research and evaluation studies indicate that integrated learning systems (ILS) are effective forms of CBI which are quite likely to play an even larger role in classrooms in the foreseeable future.
- Intelligent tutoring systems have not had significant impact on mainstream education because of technical difficulties inherent in building student models and facilitating human-like communications.
- Overall, the differences found between interactive learning systems as tutors and human teachers have been modest and inconsistent. It appears that the larger value of these systems as tutors rests in their capacity to:
  - motivate students,
  - increase equity of access, and
  - reduce the time needed to accomplish a given set of objectives.

The foundation for the use of interactive learning systems as "cognitive tools" (the "with" approach) is "cognitive psychology." Computer-based cognitive tools have been intentionally adapted or developed to function as intellectual partners to enable and facilitate critical thinking and higher order learning. Examples of cognitive tools include:

- databases,
- spreadsheets,
- semantic networks,
- expert systems,
- communications software such as teleconferencing programs,
- on-line collaborative knowledge construction environments,
- multimedia/ hypermedia construction software, and
- computer programming languages.

In the cognitive tools approach, interactive tools are given directly to learners to use for representing and expressing what they know (Jonassen & Reeves, 1996). Learners themselves function as designers, using software programs as tools for analyzing the world, accessing and interpreting information, organizing their personal knowledge, and representing what they know to others.

The basic principles that guide the use of interactive software programs as cognitive tools for teaching and learning are:

- Cognitive tools will have their greatest effectiveness when they are applied within constructivist learning environments.
- Cognitive tools empower learners to design their own representations of knowledge rather than absorbing representations preconceived by others.
- Cognitive tools can be used to support the deep reflective thinking that is necessary for meaningful learning.
- Cognitive tools have two kinds of important cognitive effects, those which are
  - with the technology in terms of intellectual partnerships and
  - of the technology in terms of the cognitive residue that remains after the tools are used.
- Cognitive tools enable mindful, challenging learning rather than the effortless learning promised but rarely realized by other instructional innovations.
- The source of the tasks or problems to which cognitive tools are applied should be learners, guided by teachers and other resources in the learning environment.
- Ideally, tasks or problems for the application of cognitive tools will be situated in realistic contexts with results that are personally meaningful for learners.

- Using multimedia construction programs as cognitive tools engages many skills in learners such as:
  - project management skills,
  - research skills,
  - organization and representation skills,
  - presentation skills, and
  - reflection skills.
- Research concerning the effectiveness of constructivist learning environments such as microworlds, classroom-based learning environments, and virtual, collaborative environments show positive results across a wide range of indicators.

In summary, over thirty years of educational research indicates that various interactive technologies are effective in education as phenomena to learn both "from" and "with." Historically, the learning "from" or tutorial approaches have received the most attention and funding, but the "with" or cognitive tool approaches are the focus of more interest and investment than ever before. Preliminary findings suggest that in the long run, constructivist approaches to applying media and technology may have more potential to enhance teaching and learning than instructivist models (Jonassen & Reeves, 1996). In other words, the real power of interactive learning to improve achievement and performance may only be realized when people actively use computers as cognitive tools rather than simply interact with them as tutors or data repositories.

Concurrently, there is a paucity of empirical evidence that interactive learning technologies are any more effective than other instructional approaches. This is because most research studies confound media and methods. Sixteen years ago, a debate was ignited with the provocative statement that "media do not influence learning under any conditions" (Clark, 1983, p. 445). It was later clarified by explaining that media and technology are merely vehicles that deliver instructional methods, and that it is instructional methods, the teaching tasks and student activities, that account for learning. It was maintained that as vehicles, interactive technologies such as computer-based instruction do not influence student achievement any more than the truck that deliver groceries changes our nutrition. Clark (1994) concluded that media and technology could be used to make

- learning more efficient (enable students to learn faster),
- more economical (save costs), and/or
- more equitable (increase access for those with special needs).

Robert Kozma challenged Clark in the debate about the impact of media and technology on learning by arguing the separation of media and methods creates "an unnecessary and undesirable schism between the two" (Kozma, 1994, p. 16). It was recommended that we move away from the questions about whether technologies impact learning to questions concerning the ways can we use the capabilities of interactive technology to influence learning:

- for particular students
- with specific tasks and
- in distinct contexts.

Kozma recognized that although interactive technologies may be essentially delivery vehicles for pedagogical dimensions, some vehicles are better at enabling specific instructional designs than others.

Both perspectives are important ideas. It is evident that the instructional methods students experience and the tasks they perform matter most in learning. In addition, the search for unique learning effects from particular interactive technologies appears ultimately futile since fifty years of media and technology comparison studies have indicated no significant differences in most instances. Whatever differences are found can usually be explained by differences in

- instructional design,
- novelty effects, or
- other factors.

However, even though technologies may lack unique instructional effects, some educational objectives are more easily achieved with interactive learning than in other ways. Revealing effective implementations of interactive learning for:

- various types of learners ;
- discrete learning objectives; and
- content

is an important goal for educational researchers and evaluators.

A Development Research Agenda

The fact that educational research is not highly valued by educational practitioners is widely recognized. A large part of the problem can be attributed to the fact that the interests:

- of academics who conduct research and
- those of administrators, teachers, students, parents, and others involved in the educational enterprise are often quite different. Tanner (1998) reminds us that educational research should be focused on the mission of enhancing educational opportunities and outcomes.

As noted in the previous section, research reveals that students learn both from and with interactive learning technology. Computer-based instruction and integrated learning systems have been demonstrated to be effective and efficient tutors, and there is considerable evidence that learners develop:

- critical thinking skills as authors, designers, and constructors of multimedia or as
- active participants in constructivist learning environments.

Unfortunately, the level of knowledge about interactive learning is somewhat analogous to what health practitioners know about the functions of vitamins and herbs in supporting good health. There is general agreement within the healthcare professions that most vitamins and many herbs have health benefits, but there is considerable disagreement about the proper dosages, regimens, and protocols for using various products. Similarly, in education, while there is general agreement that interactive learning is good, very little is known about the most effective ways to implement interactive learning. In fact, the need for long-term, intensive research and evaluation studies focused on the mission of improving teaching and learning through interactive learning technology has never been greater.

Both international and national government and commercial interests are pushing interactive learning in various forms from preschool through lifelong learning, and major decisions are being made about these technologies based upon habit, intuition, prejudice, marketing, politics, greed, and ignorance rather than reliable and valid evidence provided by research and evaluation.

Research and evaluation efforts should be primarily development in nature, i.e., focused on the invention and improvement of creative approaches to enhancing human communication, learning, and performance through the use of interactive learning technologies. The purpose of such inquiry should be to improve, not to prove. Further, development research and evaluation should not be limited to any one methodology. Any approach, quantitative, qualitative, critical, and/or mixed methods, is legitimate as long as the goal is to enhance education.

Policy-makers in the USA the Panel on Educational Technology of the President's Committee of Advisors on Science and Technology (1997) established three priorities for future research:

1. Basic research in various learning-related disciplines and fundamental work on various educationally relevant technologies.
2. Early-stage research aimed at developing new forms of educational software, content, and technology-enabled pedagogy.
3. Empirical studies designed to determine which approaches to the use of technology are in fact most effective. (p. 38)

The second of these priorities reflects the need for development research issued above. However, some researchers feel that the President's Committee of Advisors on Science and Technology (1997) has placed too much faith in the ability of large-scale empirical studies to identify the most effective approaches to using interactive learning in schools. In the final analysis, the esoteric and complex nature of human learning may mean that there may be no generalizable "best" approach to using interactive learning technology in education. The most one may hope for is more creative application and better informed practice.

Evaluation Framework: Technology Mediated Learning

Salomon (1991) describes the contrast between analytic and systemic approaches to research that transcends the "basic versus applied" or "quantitative versus qualitative" arguments that so often dominate debates about the relevancy of educational research. Salomon concludes that the analytic and systemic approaches are complementary, arguing that "the analytic approach capitalizes on precision while the systemic approach capitalizes on authenticity".

One reason for this state of affairs is that there has long been great disagreement about the purpose and value of educational research. Should researchers and evaluators seek to establish immutable laws akin to those found in the harder sciences? Or should we be focused on finding out how to improve education with different types of students in specific places at particular times of their development? These questions reflect an on-going struggle between those who view this field as a science and those who regard it as a craft. The aforementioned Panel on Educational Technology of the President's Committee of Advisors on Science and Technology (1997) listed as one of its six major strategic recommendations that the government "initiate a major program of experimental research....to ensure both the efficacy and cost effectiveness of technology use within our nation's schools" (p. 5). Reeves (1999) contends that a wiser course would be to support more development research (aimed at making interactive learning work better) using a wider range of quantitative, qualitative, critical, and mixed methods and less empirical research (aimed at determining how interactive learning works) using experimental designs.

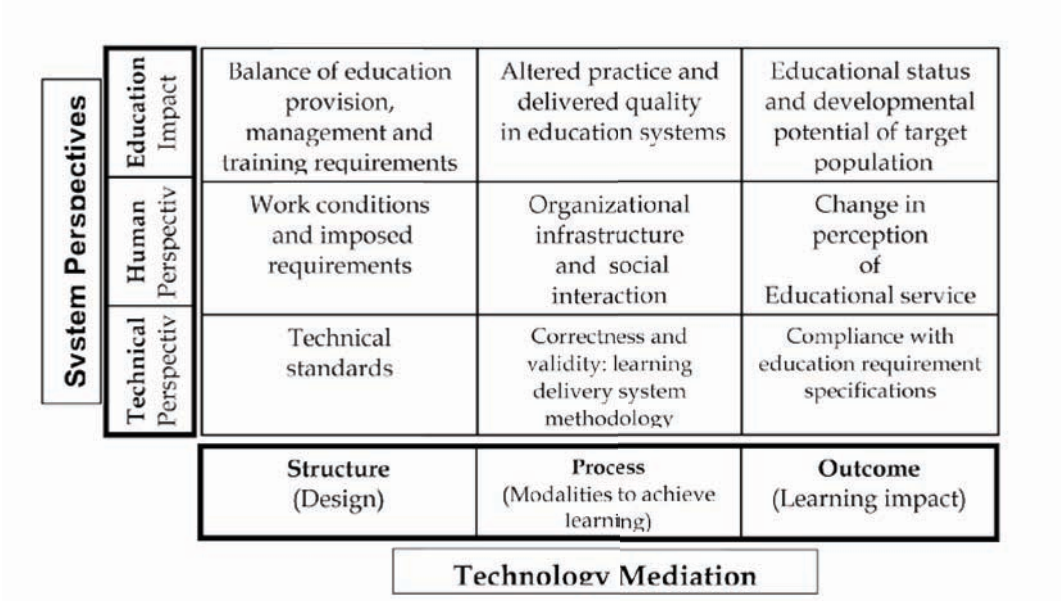
From a baseline of practice of attempting to evaluate many e-learning programmes, one of the biggest problems has proved to be handling the number of variables which potentially impact on the effectiveness of the programme and deciding what constitutes dependent, independent and irrelevant variables in a given situation.

According to (Hughes 2003), over several e-learning evaluation projects, five major clusters of variables have emerged; individual learner variables, environmental variables, technology variables, contextual variables and pedagogic variables.

Individual learner variables include physical characteristics, learning history, learner attitude, learner motivation, and familiarity with the technology. On the other hand learning environment variables include the immediate (physical) learning environment, the organisational or institutional environment and the subject environment. Contextual variables include socio-economic factors, the political context and cultural background. Technology variables have been identified as the kind of hardware and software used and the connectivity, media and mode of delivery employed. Pedagogic variables touch on issues concerned with the level and nature of learner support systems, accessibility issues, methodologies employed, flexibility, learner autonomy, selection and recruitment criteria, assessment and examination, and accreditation and certification issues.

The issues discussed above are incorporated and aggregated into a two-dimensional structured framework that considers pedagogic and contextual issues see Omwenga & Rodrigues (2006). We note that the two issues cut across human and technical provisions of an educational system. Attitude, and socio-political issues that define the context and environmental issues that define cultural values, all result in human perspective A matrix-like framework is shown below in figure 1

Figure 1: Technology Mediated Learning: Evaluation



The framework in figure 1 indicates that any technology mediation for educational purposes has a structure, a process and a learning outcome (SPO) which can be viewed from three main systems perspectives: that of the technical system functioning, human perspectives of those involved, and the overall impact on the education system. A matrix diagonal resulting from the overlay are technical standards; organisational infrastructure and social interaction; and educational status and developmental potential of target population. By extending this framework more fully into a matrix we may explore and study broader and more comprehensive relationships. For example, when the structure (or design) of a system is considered in terms of its overall impact on the educational system, one can appreciate the opportunities and challenges that arise in implementing interactive learning across emerging economies. Let us consider each of these matrix items (levels and components) in turn.

Level 1. The system's functioning (Technical).

This may be referred to as the raw efficiency of the system itself. Broken down into the evaluation procedure, the following aspects may be considered for an e-learning system:

- **Structure:** What are the hardware requirements and is the software structure understandable? Does the full set of system components work together in a technical sense? Has the system implemented pedagogic requirements that mimic classical principles of instruction?
- **Process:** Is the method by which an instructional system brings about learner behavioural change from a state of "not knowing" to a state of "knowing" given specific learner entry behaviour.
- **Outcome:** Here we want to know if the results are relevant, applicable and reliable. We ask: do they meet the requirement specifications?

Level 2. Human perspectives.

This includes the acceptability of the system by the various stakeholders, and considers how the system's functions affect them. Foster & Conford's (1992) technology evaluation in health services recognises at least three roles under human perspectives. In other situations, there may be more roles worthy of consideration, in which case, the number of stakeholders has to be increased. Assessing human perspectives of information systems is not easy and these aspects are not easily measurable. Researchers must allow themselves both the freedom to identify sufficient stakeholders and the freedom of using qualitative judgments in their analyses when quantitative measures cannot be obtained. We identify three stakeholders in this framework:

The user (Instructor). This is the primary agent in the system implementation, who is indispensable for its proper functioning. Within the SPO dimension, this poses questions such as:

- **Structure:** What are the changes to working conditions, in terms of the physical environment, skill requirements etc.?
- **Process:** How is the user's mode of operation changed? Are these changes seen as desirable to the user as an individual, and to the user's organisational role?
- **Outcome:** Is the overall effectiveness of the user within the education system enhanced?

The Learner. This is the person who the system is expected to benefit, and who is often directly or indirectly affected by its implementation.

- **Structure:** Are learners required to modify their behaviour in any way?
- **Process:** How is the learner's experience altered at the point of contact with the system?
- **Outcome:** Does the use of the system result in changes in the quality of service and better education for the recipient?

The Administrator. This is the person responsible for the general management of the e-learning unit. Note that, since this is under the human perspective heading, assessment at this level is focused on the person responsible for the management of the individual e-learning unit rather than the whole education system. Thus, it is limited to the administrator's immediate concerns.

- **Structure:** Is the system a reasonable, cost-effective and efficient alternative to existing structures?
- **Process:** Does the system imply change in the delivery of activities for which the administrator is responsible'? Does it change the character of the administrator's job?
- **Outcome:** Does the system improve specific education provision on a reasonable metric?

Level 3. Education system.

This involves a consideration of the impact of a system's use on the education system as a whole, and on e-learning itself. It concerns the national developmental level in its widest possible sense.

- **Structure:** Does it change the balance between the functions of the different education providers?
- **Process:** Does it affect practice and delivered quality of education provision?
- **Outcome:** Does it improve the education status and development potential of the population it serves?

Conclusion

The justification for the approach described here can be made from two main angles. Firstly, it provides a procedure that assesses the technical and the social aspects of a system, as well as the long-term impact on improving education provision. Secondly, it is a standardized procedure for reporting evaluation results that may be widely applied. In summary, the evaluation framework proposed here permits a structured view of interactive learning projects which recognises both the need to link the technology grounded applications with ones that involves a more fundamental understanding of the broader concept of the value of education.

Applying the evaluation model to two e-learning case studies, see Omwenga & Rodrigues (2006), has shown that, to a large extent, e-learning is a sustainable and viable mode of instructional process. The structure, process, and outcome in each of the three levels of evaluation has not brought out any fundamental flaws that threaten the mode of delivery. It is however, not a matter of conjecture that higher learning institutions need to make initial substantial resource input in implementing e-learning in their institutions.

These observations are of interest to policy makers. Quite clearly, technology has a positive impact in supporting education provision. E-learning is a viable mode of delivery. However, recommendations on the form, structure and process of introducing universal e-learning in institutions of higher learning needs a much broader evaluation in terms of capacity to sustain the technology and enabling access to it that will not exclude others. Moreover, although many researchers have claimed that technology-mediated education can play a foundation role of all attempts at poverty alleviation, better health, and larger access to education,

governments, and technocrats may argue that provision of basic teaching materials such as books, laboratory equipment, and even qualified human resource manpower are more urgent issues to deal with in poor economies.

References

Clark, R. E. (1994). Media will never influence learning. Educational Technology Research and Development, 42(2), 21-29. [Clark 1994]

Clark, R. E. (1983). Reconsidering research on learning with media. Review of Educational Research, 53(4), 445-459. [Clark 1983]

Clark, R. E., & Estes, F. (1998). Technology or craft: What are we doing? Educational Technology, 38(5), 5-11. [Clark & Estes 1998]

Farley, F. H. (1982). The future of educational research. Educational Researcher, 11(8), 11-19. [Farley 1982]

T Foster, D., & Conford T. (1992). Evaluation of Health Information systems: Issues, Models and Case studies, Bhatnagar, S.C. & Odedra M. (eds.), Social Implications of computers in Developing Countries. McGraw-Hill. 304-309. [Foster, D., & Conford 1992]

Hughes, J., & Attwell, G. (2003). A framework for the Evaluation of E-Learning, 2003, European Seminar Series on Exploring Models and Partnerships for eLearning in SMEs, Brussels.

Jonassen, D. H., & Reeves, T. C. (1996). Learning with technology: Using computers as cognitive tools. In D. H. Jonassen (Ed.), Handbook of research for educational communications and technology (pp. 693-719). New York: Macmillan. [Jonassen & Reeves 1996]

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. Educational Technology Research and Development, 42(2), 7-19. [Kozma 1994]

Omwenga E.I., & Rodrigues A.J. (2006). Towards an Education Evaluation Framework: Synchronous and Asynchronous E-learning Cases. Journal of the Research Centre for Educational Technology Spring 2006. http://www.rcet.org [Omwenga & Rodrigues 2006]

Perelman, L. J. (1992). School's out: Hyperlearning, the new technology, and the end of education. New York: William Morrow. [Perelman 1992]

Postman, N. (1995). The end of education: Redefining the value of school. New York: Alfred A. Knopf. [Postman 1995]

President's Committee of Advisors on Science and Technology. (1997, March). Report on the use of technology to strengthen K-12 education in the United States http://www.whitehouse.gov/WH/EOP/OSTP/NSTC/PCAST/k-12ed.html. Washington, DC: The White House. [President's Committee of Advisors on Science and Technology 1997] [Reeves 1999] Key Note Address ED-MEDIA '99

Salomon, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. Educational Researcher, 20(6), 10-18. [Salomon 1991]

Tanner, D. (1998). The social consequences of bad research. Phi Delta Kappan, 79(5), 345-349. [Tanner 1998]

23
M-Learning: The Educational Use of Mobile Communication Devices

Paul Birevu Muyinda , Ezra Mugisa , Kathy Lynch

This is a position paper which explores the use of mobile communication devices in teaching and learning. We especially undertake a crosswalk around the fast evolving field of mobile learning with a view of positioning it as new paradigm for learning. We deduce that successful development and implementation of any mobile learning solution requires a deep understanding of the learning environment which is greatly influenced by the learners learning styles, theories behind m-learning, the technology at play and the institutional/organisational culture. Interplay of all these factors enables the contextualization of learners and provides a learner centered learning environment. We eulogize the need to blend mobile with fixed communication devices in order to bridge the digital divide in electronic learning and finally conclude that mobile technological innovations per se are not enough to propel mobile learning but a match in growth need to be realised in the mobile learning pedagogies, theories, philosophies and organisational attitudes towards its use in education.

Introduction and Background

The use of mobile communication devices in education has led to the evolution of a new paradigm in electronic learning (e-learning) called mobile learning (m-learning). M-learning, is a form of e-learning that specifically employs wireless portable communications devices to deliver content and learning support (Brown, 2005) [4]. Advances in mobile computing and handheld devices (ipod, cell phones, smart phones, PDA, notebooks, etc), intelligent user interfaces, context modelling, wireless communications and networking technologies (WI-FI, Blue Tooth, GPS, GSM, GPRS, 4G) have precipitated mobile learning (Sharples, 2000 [30]; Knowledge Anywhere, 2002) [15].

M-learning is gaining prominence because of the increasing desire for lifelong learning which is usually undertaken by learners with other life obligations related to work, family and society. Such learners are constantly on the move and require devices that facilitate learning on the go. Typical e-learning systems have failed to provide on-the-go learning because they are usually situated in fixed environments (Woukeu et al., 2005) [36]. They are tailored toward PC based web access and are not customized for mobile devices (Goh and Kinshuk, 2006) [11]. Hence m-learning comes in handy to support the on-the-move learner.

M-learning just like its parent field, e-learning, has not fully matured. Consequently, m-learning is attracting considerable research (Woukeu et al.,

2005) [36]. Unfortunately, research in young fields has been criticised for lacking appropriate theories and clear epistemological stand view points (Mitchell, 2000) [23]. As an evolving research area, many fundamental issues in M-learning are yet to be exhaustively covered (Goh and Kinshuk, 2006) [11]. Such fundamental issues have been identified by Conole (2004) [6]. According to him, new learning technologies are best appreciated if one can understand the technology at play, learning styles of the technology users, the pedagogical aspects of using the technology for teaching and learning and organizational or institutional attitude towards the technology. Goh and Kinshuk (2006) [11] are in consonance with Conole (2004) [6] as they re-emphasise the need for further research in pedagogical practices generated from simple wireless and mobile technologies. It is therefore evident that the design and implementation of m-learning applications has not stabilised since its theoretical and philosophical underpinnings are just in the making. We use Conole's (2004) [6] prerequisites for introducing new learning technologies to explore the possibilities and challenges of integrating mobile communications technologies into the teaching and learning process. Our position is based on works around mobile communications devices for learning, m-learning and learning styles, m-learning theories, mobile technological innovations and learning, organizational culture and m-learning, challenges to m-learning and future projections for m-learning.

Possibilities of mobile communications devices for learning

We use the phrase mobile communication devices to refer to those wireless communication devices that can be used while on-the-move. These devices provide "wearable computing environments" (Sung et al., 2005, pg 2) [32]. They come in a multitude of models, sizes, capabilities and purposes (Attewell, 2005) [2]. Mobile phones, smart phones, PDA, iPods, notebooks and zunes are examples of such devices. Their permeability varies from country to county and from one economic class to another. However, the ownership of the mobile phone has been democratised as a wide spectrum of the populace, irrespective of race, economic status and country, have embraced its use (Prensky, 2004) [29]. The mobile phone is a necessary device of life (Muyinda, 2007) [26].

Apart from contributing to their orthodox purpose of communication, these devices are widely being used in commerce and entertainment (Keegan, 2005) [13]. The iPod, for instance, is a great source of entertainment for both the digital natives and migrants. High end mobile phones are now being used to access Internet and a number of organisations have developed mobile phones applications (Sung et al. 2005 [32]; MobiLearn, 2005 [25]; Goh and Kinshuk, 2006 [11]; Thornton & Houser, 2005) [33]. This has been induced by the capabilities of mobile phones and their widespread acceptability and permeability. By 2004, there were 1.5 billion mobile phones in the world, a number which was three times the number of PCs (Prensky, 2004) [29]. The growth in number has equally been matched with growth in processing capacity. The processing capacity of high end phones

is comparable to that of mid 1990 PCs, a capacity which was required to land the spaceship on the moon in 1969 (Archibald, 2007) [1]. This processing power has provided an impetus for use of mobile phones in education. Therefore the bulky of this paper is concentrated on the mobile phone as a tool for m-learning.

Several large scale initiatives for example MobiLearn (MobiLearn, 2005) [25], MLearning (MLearning, 2005) [24] and From e-Learning to M-learning (Ericsson, 2002) [8] have been investigating the potential benefits of this new pervasive approach to learning. An m-learning survey in UK's schools and higher education has suggested that young adults (16-24) are switched onto learning by mobile phones and PDAs (LSDA, 2003) [18]. Goh and Kinshuk (2006) [11] have cited several m-learning initiatives including games-oriented implementation for m-portal (Mitchell, 2003) [22]; class room of the future (Dawabi et al., 2003) [7]; hands-on scientific experimentation and learning (Milrad et al., 2004) [21]; mobile learning system for bird watching (Chen et al., 2003) [5] and context-aware language learning support system (Ogata and Yano, 2004) [28]. At Kinjo Gakuin University in Japan, mobile phone have been used in the teaching and learning of English language (Thornton & Houser, 2005) [33] while at the University of Pretoria, m-learning has been used for extending administrative support to distance learner (Brown, 2005) [4]. Reminders for critical dates and events are sent to distance learners as SMS messages, snippets of audio messages are recorded on telecommunications companies' servers for students to call in and listen, textual study materials are augmented with short objective type questions that students are required to answer with real-time feedback being provided. Attempts are being made to have content delivered on mobile communications devices but with little pedagogic practices (Sung et al., 2005) [32].

Goh and Kinshuk (2006) [11] have identified several other applications of m-learning. For instance, through the use of interactive games and contests installed on mobile devices, learners can construct their own knowledge and share among themselves. In the classroom, m-learning integrates with online learning management systems to provide tools for brainstorming, quizzing, and voting. In the laboratory, m-learning supports individual learning as well as collaborative learning. Mobile devices can be of benefit to laboratory environments for data gathering and control. In field trips, mobile devices support learning by collecting pictorial and textual data. Their mobility enables learning to take place in the field. In distance learning mobile devices support the delivery of synchronous and asynchronous learning while in informal settings the devices support incidental and accidental learning. M-learning supplements formal learning and teaching.

As already alluded to mobile communications devices come in variety of sizes, types, designs and models. The varied designs are meant to cater for varied customer tastes. Attewell (2005, pg 2) [2] confirms thus:

The modern mobile phone market caters for a wide variety of customer tastes and lifestyles. Some phones are tiny and discrete, some are chosen for their appearance (like a fashion accessory, with alternative covers that allow that appearance to be

changed to match the owner's outfit), some just offer basic functionality while some others provide a wide range of business and leisure services to their users. Manufacturers are marketing diverse product ranges, including devices that specialize in providing particular services or are aimed at particular users. Instead of describing a product as a mobile phone, manufacturers often use descriptions like 'game deck', 'communicator' or 'mobile multimedia machine'.

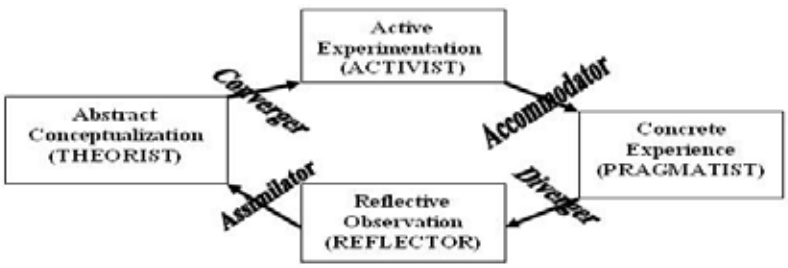
This implies that applications designed for use on mobile phones must take cognisance of user preferences. In teaching and learning, the application should conceptualise the learner (Conole, 2004) [6] and take care of the different learning styles and learning preferences.

M-learning and learning styles

Learning is the process of retaining/remembering and understanding the material in order to implement it (Muyinda, 2007) [26]. The philosophy underpinning the design and implementation of an online course recognizes the need for: a conducive learning environment, dedicated blocks of time on the part of the learner for e-learning, course developers focusing on the course's main objective, targeting users' profiles, defining the course level, repetitions to allow concepts to sink and be learned, having an engaging presentation, hands-on participation/ practice, immediate feedback and evaluation (Mescan, 2006) [20]. M-learning can particularly enable repetition of facts, immediate feedback and evaluation, and can actively engage learners as it emphasizes learner centeredness.

Further, the hands on practice enable experiential learning as is espoused in Kolb's (1984) [16] experiential learning theory. According to this theory, ideas are not fixed, but are formed and modified through the experiences we have and by our past experience, hence making learning process cyclic in nature. M-learning can abet Kolb's (1984) [16] cyclic learning process in which the learner working from his/her experience can come up with several reflections that can be used in formulating new theories which he/she can experiment with in order to get new experiences (Figure 1). Whenever, we travel from one location to another, we gain different experiences which we can record on our mobile devices and reflect on when conceptualizing new theories that we later experiment with to gain further experience.

Fig 1: Learner contextualization (Honey and Mumford, 1982 [12]; Kolb, 1984 [16]; Wolf and Kolb, 1984 [35]; McKimm, 2002 [19])



McKimm (2002) [19] strengthened Kolb's learning cycle by identifying the learning abilities needed at each point of the cycle. He (McKimm, 2002 [19]) contends that the learning cycles require four kinds of abilities, namely:- i) concrete experience - where learners are enabled and encouraged to become involved in new experiences; ii) reflective observations - where time must be availed to learners to be able to reflect on their experiences from different perspectives; iii), abstract conceptualization where learners must be able to form and process ideas and integrate them into logical theories; and iv) active experimentation where learners need to be able to use theories to solve problems. A community of practice sending educative SMS messages to each other via a mobile phone is likely to benefit from the aforementioned abilities (Figure 1).

Learning abilities have been strongly associated with Wolf and Kolb's (1984) [35] learning styles namely: the accommodator, diverger, assimilator and converger. These have been juxtaposed in the learning cycle in Figure 1. The accommodator carries out plans and tasks that involve them in new experiences, the diverger has good imaginations and generates new ideas, the assimilator creates theoretical models and makes sense of disparate observations while the converger applies ideas in a practical way. These styles march Honey and Mumford's (1982) [12] categorization of learners as either activists or reflectors or theorists or pragmatists. The activist responds most positively to learning situations that offer challenge and which include experiences and problems, the reflector responds most positively to structured learning activities in which time is provided to think, the theorist responds most positively to logical, rational structure and clear aims, and pragmatist responds most positively to practically based immediate relevant learning activities which allow them to practice and use the theory. These are also mapped onto the learning cycle in Figure 1. Understanding the different learning styles and abilities enables the formulation of an all inclusive m-learning theory.

M-Learning Theories

Literature indicates lack of m-learning theories, but grounds are being prepared for their development. Sharples et al., (2005) [31] have suggested four pre-requisites for the formulation of an m-learning theory. The first pre-requisite requires one to distinguish between what is special about m-learning vis-à-vis other types of learning. The second pre-requisite is to determine the amount of learning that occurs outside the classroom with a view of m-learning embracing it. Vavoula (2005) [34] in a study of everyday adult learning discovered that the majority (51 percent) of learning episodes took place at home or at place of work - learners usual environment, 21 percent - outside the office, 5 percent - outdoors, 2 percent - in friend's home, 6 percent - place of leisure, 14 percent - places of worship, the doctor's room, cafes, hobby stores and cars. Only 1 percent occurred on transport. The learning occurring on transport suggests that m-learning is not necessarily associated with physical movements. The third pre-requisite is the need to bear in mind the contemporary accounts of practice such as learner centeredness,

knowledge centeredness, assessment centeredness and community centeredness. The fourth pre-requisite is that an m-learning theory must take account of the ubiquitous use of personal and shared technology. This negates Keough's (2005) [14] pessimism about the working of m-learning arising from its association with the ever changing mobile technology.

Sharples et al., (2005) [31] enlist questions to provide a criterion against which an m-learning theory could be tested. Is it significantly different from current theories of classroom, workplace or lifelong learning? Does it account for mobility of learners? Does it theorize learning as a constructive and social process? Does it analyse learning as a personal and situated activity mediated by technology? A clear distinction between classroom and mobile learning ought to be drawn. Vavoula (2005) [34] reports that the MobiLearn European project while reflecting on formulating a theory for m-learning identified that: it is the learner who is mobile rather than the technology, learning is interwoven with other activities as part of everyday life, learning can generate as well as satisfy goals, the control and management of learning can be distributed, context is constructed by learners through interaction, m-learning can both complement and conflict with formal education, and that m-learning raises deep ethical issues of privacy and ownership.

In absence of concrete theoretical underpinnings for m-learning, existing theories can be harnessed to provide a rich learning experience in M-learning. Naismith et al. (2004) [27], have, during their review of M-learning literature, proposed to solve the dearth in m-learning theories by considering new practices against existing learning theories - behaviourist, constructivist, situated, collaborative, informal and lifelong learning theories.

The behaviourist learning theory emphasizes activities that promote learning as a change in learner's observable actions. The learning should invoke a stimulus and a response. In the case of m-learning, an SMS message, for example, invokes a stimulus which may lead to an action as a response.

The constructivist learning theory emphasizes activities in which learners actively construct new ideas or concepts based on both their previous and current knowledge. With a mobile phone learners can construct their own knowledge and share it freely with peers at anytime in any place. This in m-learning is referred to as 'participatory simulations' (Naismith et al., 2004) [27].

The situated learning theory emphasizes activities that promote learning within an authentic context and culture. Mobile devices are especially well suited to context-aware applications simply because they are available in different contexts, and so can draw on those contexts to enhance the learning activity.

The collaborative learning theory emphasizes activities that promote learning through social interaction. Through conversations on mobile phones collaboration can be enhanced.

The informal and lifelong learning theory promotes activities that support learning outside a dedicated learning environment and formal curriculum. Mobile

technologies can support informal learning which may be intentional or accidental (Sharples, 2000) [30]. Intentional learning may be acquired through, for example, intensive, significant and deliberate learning efforts, while accidental learning may be acquired through conversations, TV and newspapers, observing the world or even experiencing an accident or embarrassing situation. As was found by Vavoula (2005) [34], the majority of learning episodes in adults is informal. Continuous innovations in mobile communications technology are precipitating informal learning.

Mobile Communications Technology Innovations and Learning

The hype usually attached to new learning technologies often shoots the technology in its own foot as critical issues related to its "usability, flexibility and extensibility are often over shadowed by the need to quickly demonstrate the new features of the technology" (Sung et al., 2005, pg 1) [32]. Hence new technologies are embraced on the surface with no deep understanding of their fullest potentials (Graham, 2004) [10]. M-learning being a young field, its impact and capabilities have not been fully explored. It is well known that most computer users exploit only a small proportion of the technology available to them, and that immensely powerful machines are often used as little more than hi-tech typewriters and calculators. Keegan (2005) [13] has observed that the mobile phone has been around for a couple of years with little regard to its potential for learning.

For an innovation which necessitates technological change and social re-organization, Graham (2004) [10] proposes a framework to answer questions such as: i) What the anticipated benefit of the innovation will be and whether there will be genuine additional benefits; ii) whether the chance of its being implemented successfully is much higher than the chance of its failure; iii) what the cost of its introduction would be in terms of disruption to existing systems that are known, tried and reliable; iv) how stable the circumstance in which the proposed innovation is to be made; and v) whether there are recurrent patterns of behaviour that would give some pointers to its likely reception?

Relatedly, Conole (2004) [6] while considering underpinning technology of e-learning also asks questions such as: i) what are the new and emerging technologies and how can they be used to support learning and teaching? ii) What learning platforms are being used and how do they compare? iii) What are the emerging new software and hardware systems? iii) How can we explore mobile and smart technologies? and iv) What ways are in-built tracking mechanisms within m-learning systems giving rise to surveillance issues?

These questions bear directly on the lives of people for whom the innovation is intended. These questions ought to be answered before undertaking any new technological innovations. IT projects have been undertaken whose results have not benefited the intended users. Care ought to be taken because; in the name of technological improvement, a huge cost in terms of personnel as well as money can be incurred by an organisation quite pointlessly when all conditions in the organisation are not favourable for m-learning.

Organizational culture and M-learning

Lee (2003) [17] observes that an information system is not the information technology alone, but the system that emerges from the mutually transformational interactions between the information technology and the organization. This implies that innovations in mobile technology per se can not propel m-learning. There must be a positive organisational attitude towards the technology and the people in the organisation must be enabled to use the technology. Hence, to effectively use m-learning in an organization, m-learning system designers need to understand: i) how the different stakeholders (academics, support staff, administrators, senior managers and students) currently work; ii) the mechanism and procedures for developing shared knowledge banks of expertise and information; iii) the need to outline roles and responsibilities for m-learning activities - management, technical, research, dissemination, evaluation and training; iv) the different views to m-learning and its role - academics vs. support staff; v) how the institution divides roles and the responsibilities for m-learning; and vi) how much training and support the staff are to get (Conole, 2004) [6].

Challenges to M-learning

Keough (2005) [14] is pessimistic about the functioning of m-learning. Keough (2005) [14] has advanced seven reasons as to why m-learning will not work. According to him m-learning as a concept alone is doomed to failure because as a learning model it appears:

- To be technology driven: M-learning alone is a technology driven concept
- Not cogniscent of market usage: We know too little about what mobile devices are used for
- Yet to adopt discoveries in Cyber psychology: We know too little about flow and learning relationships/Networks or the Transactional Analysis of Mobile Relationships
- Not to change entrenched institutionalised education Models: Cultures of education and communications reflect government control measures
- To rely on nascent consumer technology: Mobile devices are inherently dissatisfying by never quite meeting every promised need for the consumer.
- To be short on standards to overcome cultural differences: while standards are slow to emerge governments are rapidly regulating and limiting the use of mobile communications technology
- To lack a mobigogy: teaching and learning models are needed

--- Keough (2005, pg 1) [14]

Others such as Boone (2007) [3] cite examination malpractices as one of the challenges that will make the wearable technologies not desirable for learning especially among administrators of educational institutions.

> *...schools [in the US] started banning cell phones, realizing students could text message the answers to each other. Now, schools across the country [US] are targeting digital media players as a potential cheating device. ...Devices including iPods and Zunes can be hidden under clothing, with just an ear bud and a wire snaking behind an ear and into a shirt collar... Some students use iPod-compatible voice recorders to record test answers in advance and play them back (Boone, 2007, pg 1) [3]*

With such developments in the minds of school/college/university administrators, it will take considerable effort to lobby for the acceptance of the mobile phone or any other wearable communication gadget as a tool for learning. Where they have to be accepted, considerable restrictions will have to be imposed, sometimes to the detriment of learning.

Another challenge to m-learning stems from the habit of showcasing new communication technology as a learning tool. In many instances, efforts are mainly concentrated on delivering content to these mobile devices with little consideration to the rich potential for more interactive learning paradigms (Sung et al., 2005) [32]. These showcases usually over shadow practical issues related to usability, flexibility, and extensibility in favor of quickly demonstrating the new features of the technology. This, in most cases deals a killer blow to m-learning adoption.

The pessimism resulting from the challenges identified here and elsewhere is uncalled for. The World is dynamic. Technology, cultures, teaching and learning models, methods, just to mention but a few, are not static. M-learning just like any immature field requires time to grow and to be understood. Besides, advocates for m-learning recognize the fact that it can not be used alone in its entirety (Brown, 2005) [4]. It has to be blended with other methods of delivery including face to face, print and online learning if the digital divide is to be bridged. With ubiquitous computing in sight, we can not delineate our selves from m-learning (Muyinda, 2007) [26]. It will only require time to have solutions devised for its productive use.

However, the truly big challenge for the educators and technology developers of m-learning is to find ways to ensure that this new learning technology is highly situated, personal, collaborative and long term; in other words, truly learner-centred learning.

Future Projection of M-learning

The future of m-learning is forecasted to be bright. The capabilities of mobile phones, PDAs and smart phones are always on the move to higher ends. Research endeavours in this field are magnanimous (Woukeu et al., 2005 [36]; Goh and

Kinshuk, 2006 [11]). Integrated context-aware capabilities will transform everyday activities by providing the ability to capture details about the time, location, people around you and even the weather (Naismith et al., 2004) [27]. The entire internet will become both personal and portable. Such technologies will have a great impact on learning. Learning will move more and more outside of the classroom and into the learner's environments, both real and virtual and the m-learning is well positioned to champion these innovations.

As such research into m-learning is leaning towards m-learning in games and competitive learning, classroom learning, laboratories learning, field trip learning, distance learning, informal learning, m-learning pedagogy and theories, learning and teaching support, m-learning architecture and m-learning evaluation, requirements, and human interface (Goh and Kinshuk, 2006) [11].

Conclusion

The need to utilise previously unproductive time in our daily lives and the need for lifelong learning can not be overemphasised (Geddes, 2004) [9]. Advances in mobile communications technology have provided necessary conditions for m-learning. Mobile communication devices are capable of abetting independent learning for learners with different learning styles as it provides access to learning during previously unproductive times, it allows for more flexible and immediate collaborative options, it allows for controlled learning in contextual situations, and provides greater options for teachers to observe and assist in independent learning. Successful development and implementation of any mobile learning solution however, requires a deep understanding of the learning environment.

References

Archibald, J. (2007). "Cell phones as an educational tool". Available at http://www.tectonic.co.za/view.php?id=1396. (Accessed 20th March 2007).

Attewell, J. (2005), "Mobile technologies and learning: A technology update and m-learning project summary", London: Learning and Skills Development Agency. Available at www.LSDA.org.uk. (Accessed 30th May 2006).

Boone, R. (2007). "Schools banning iPods to beat cheaters". Available at http://news.yahoo.com/s/ap/20070427/ap_on_hi_te/ipod_cheating. (Accessed April 27, 2007).

Brown, H. T. (2005). Towards a Model for M-Learning. *International Journal on E-Learning, 4 (3): 299-315.*

Chen, Y. S., Kao, T. C., and Sheu, J. P. (2003). A mobile learning system for scaffolding bird watching learning. *Journal of Computer Assisted Learning*, 19:347-359.

Conole, C. (2004). E-Learning: The Hype and the Reality. *Journal of Interactive Media in Education,* 2004 (12). Available at http://www-jime.open.ac.uk/2004/12/conole-2004-12.pdf. (Accessed 30th May 2006).

Dawabi, P., Wessner, M., and Neuhold, E. (2003). "Using mobile devices for the classroom of the future". In Attewell, J. and Savill-Smith, C. (Eds.), *Learning with mobile devices. Research and development.* London: Learning and Skills Development Agency.

Ericsson (2002), "From e-Learning to M-learning EU project". Available at http://learning.ericsson.net/mlearning2/project_one/index.html. (Accessed 31st May 2006)

Geddes, S.J. (2004). Mobile learning in the 21st century: benefit for learners. *Knowledge Tree e-journal*, 30 (3): 214 - 228

Graham, G. (2004). E-learning: a philosophical enquiry. *Education + Training,* 46 (6/7):308-314

Goh, T. and Kinshuk (2006). Getting Ready for Mobile Learning — Adaptation Perspective. *Journal of Educational Multimedia and Hypermedia,* 15 (2):175-198.

Honey, P. and Mumford, A. (1982). *The manual of learning styles*, Peter Honey, Maidenhead

Keegan, D. (2005). "The incorporation of mobile learning into mainstream education and training". *In Proceedings of the 4th World Conference on MLearning* (M-Learning: 2005), SA, 25-28 October.

Keough, M. (2005). *"7 reasons why MLearning doesn't work"*. Available at http://www.mlearn.org.za/CD/papers/McMillan-Keough.pdf. (Accessed 1st June 2006).

Knowledge Anywhere (2002). *"Flexible Learning: Mobile Learning Objects"*. A White Paper. Available at http://www.ottersurf.com/MLO-WP.pdf. (Accessed on 31st May 2006).

Kolb, D.A. (1984). *Experiential Learning*, Prentice-Hall, Eaglewood Cliffs, New Jersey

Lee, A. S. (2003). "Re-introducing the Systems Approach to Information Systems". Keynote Address at *ISOneWorld* Las Vegas, NV.

LSDA (2003). "Mobile phones switch young people on to learning", *Mlearning project report*, Learning and Skills Development Agency. Available at http://www.lsda.org.uk/files/pdf/press/7feb2003.doc. (Accessed 31st May 2006).

McKimm, J. (2002). "Developing yourself as a leader, learning cycles and learning styles". In Judy McKimm (ed). *Curriculum development strategies and models: learning theories, in curriculum design and development module. London Deanery Clinical Teaching.* Available at http://www.le.ac.uk/sm/le/projects/fdtl/Resources/module3/Learning (Accessed 31st May 2006).

Mescan, S. (2006). *The Philosophy of E-Learning, Progressive Information Technologies,* York County Industrial Park, Emigsville, PA. Available at http://www.pitmagnus.com/pitmagnus/news/wp_elearn_phil.pdf. (Accessed 27th March 2006).

Milrad, M., Hoppe, U., Gottdenker, J. and Jansen, M. (2004). "Exploring the use of mobile devices to facilitate educational interoperability around digitally enhanced experiments". In Roschelle, J., Chan, Kinshuk, and. Yang, S. J. H (Eds.), *Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2004), Mobile Support for Learning Communities,* Taoyuan, Taiwan, March.

Mitchell, A. (2003). "Exploring the potential of a games-oriented implementation for m-portal". In Attewell, J. and Savill-Smith, C. (Eds.), *Learning with mobile devices. Research and development.* London: Learning and Skills Development Agency.

Mitchell, P.D. (2000). "The impact of educational technology: a radical reappraisal of research methods". In Squires, D., Conole, G. and Jacobs, G. (Eds), *The changing face of learning technology*, Cardiff: University Wales Press, 51-58.

MLearning (2005). "MLearning: learning in the palm of your hand". Available at http://www.m-learning.org/. (Accessed 31st May 2006).

MobiLearn (2005). "MobiLearn Project". Available at http://www.mobilearn.org/. (Accessed 31st May 2006).

Muyinda, B. P. (2007). Mlearning: pedagogical, technical and organisational hypes and realities. *Campus-Wide Information System*, 24 (2): 97-104

Naismith, L., Lonsdale, P., Vavoula, G., Sharples, M. (2004). "Literature Review in Mobile Technologies and Learning". *Report 11. A Report for NESTA Futurelab.* Available at http://elearning.typepad.com/thelearnedman/mobile_learning/reports/futurelab_review_11.pdf. (Accessed 30th May 2006).

Ogata, H., and Yano, Y. (2004). "Context-aware support for computer-supported ubiquitous learning". In Roschelle, J., Chan, T-W., Kinshuk, and Yang, S. J. H. (Eds.), *Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2004), Mobile Support for Learning Communities,* Taoyuan, Taiwan

Prensky, M. (2004). "What Can You Learn from a Cell Phone? – Almost Anything". Available at http://www.marcprensky.com/writing/Prensky-What_Can_You_Learn_From_a_Cell_Phone-FINAL.pdf. (Accessed 29th May 2006).

Sharples, M. (2000). "The Design of Personal Mobile Technologies for Lifelong Learning". *Computers and Education*, 34:177-193.

Sharples, M., Taylor, J., Vavoula, G. (2005). "Towards a Theory of Mobile Learning". *In Proceedings of the 4th World Conference on Mlearning (M-Learning 2005)*, SA, 25th – 28th October.

Sung, M., Gips, J., Eagle, N., Madan, A., Caneel, R., DeVaul, R., Bonsen, J. and Pentland, A. (2005). Mobile-IT Education (MIT. EDU): m-learning applications for classroom settings. *Journal of Computer Assisted Learning*, 21:229–237.

Thornton, P. & Houser, C. (2005). Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning*, 21: 217–228

Vavoula, G.N. (2005). "D4.4: A Study of Mobile Learning Practices", *Report of MobiLearn Project.* Available at http://www.mobilearn.org/download/results/public_deliverables/MOBIlearn_D4.4_Final.pdf. (Accessed 30th May 2006).

Wolf, D.M. and Kolb, D.A. (1984). "Career development, personal growth and experiential learning". *In organizational psychology: Readings on human behavior*, 4th edition, eds, D. Kolb, I. Rubin and J. Maclntyre, Prentice-Hall, Eaglewood Cliffs, New Jersey.

Woukeu, A., Millard, E.D., Tao, F., Davis, C. H. (2005). "Challenges for Semantic Grid based Mobile Learning", *IEEE SITIS* 2005. Available at http://www.u-bourgogne.fr/SITIS/05/download/Proceedings/Files/f135.pdf. (Accessed 1st June 2006).

24
Implementation of E-learnin in Higher Education Institutions in Low Bandwidth Environment: A Blended Learning Approach

Nazir Ahmad Suhail, Ezra K Mugisa

Higher Education around the world is becoming networked and fundamental changes are taking place in Higher Education Institutions. There is no geographical isolation at the university or college level. When Higher Education Institutions are in the process of implementation of e-learning, a number of factors come into play. Some factors are about the technology, others about the prospective users, still others about the local context of use and the associated costs. On the other hand there are many aspects of the socio-economic and technological environment taken for granted in developed countries that need to be explicitly addressed during technological transformation in developing countries. These include; among other things, connectivity (low bandwidth) and accessibility, inadequate telecommunications infrastructure and lack of reliable power supply. This paper reviews various factors and processes with an emphasis on university settings and after analyzing, synthesizing and making a comparative study of the frameworks and models, the paper proposes a gradual transition model for implementation of e-learning in Higher Educations Institutions in Least developed countries followed by a comprehensive framework adaptable in low bandwidth environment using a blended learning approach. Implementation process of the framework is also explained.

Introduction

The growth in Internet has brought changes in all walks of life including the education sector through e-learning. The globalization of Higher Education is increasing rapidly; students attend courses of study from all over the world, employees work and study globally. Seufert (Seufert 2000) explains, "Due to the inter-activity and ubiquity of the Internet, learning is possible without space and time barriers. The long-term implications are a worldwide network and a real market place for university and college level education. This will expand naturally into vocational and adult training as well and Education might become a major export factor between countries".

When Higher Education Institutions start the process of implementation of e-learning, a number of factors come into play; "Some factors are about the technology, others about the prospective users, still others about the local context of use and the associated costs"(Wilson et al 2002). Alexander (Alexander 2001) views that successful e-learning takes place within a complex system composed of many inter-related factors. On the other hand (uys et al 2004) pointed out that during technological transformation in the developing countries; there are many aspects of the socio-economic and technological environment taken for granted in developed countries that need to be explicitly addressed. These include among other things; connectivity (low bandwidth) and accessibility, inadequate telecommunications infrastructure, and lack of reliable power supply.

This paper reviews various factors and processes with an emphasis on university settings and after analyzing, synthesizing and making a comparative study of the frameworks and models, the paper proposes a gradual transition model for implementation of e-learning in Higher Education Institutions in Least Developed Countries, followed by a comprehensive framework adaptable in low bandwidth environment, using blended learning approach, which addresses the issues associated with developing countries.

Global Trend

According to the report (Norman et al 2003), the global market for e-learning in various parts of the world which include; USA, Europe, Asia, and Africa is significant and increasing. It is reported that e-learning is one of the fastest growing sectors in the U.S and Europe education and training market with the total dollar value of all e-learning products and services projected to reach dollars 40.2 billion and 6 billion respectively in 2005. By giving the details of e-learning developments which have taken place in Asia and Africa, the report concludes that more than 120 universities in Japan have installed a communications satellite system for organizing lectures, seminars, and meetings, while developing countries are also making extensive use of distance learning.

Statement of the Problem

Least Developed Countries (LDCs) fall under the category of low bandwidth environment, where the average bandwidth available to a user is much lower than that in the developed world (aidworld 2006) and average university pays 50 times more for their bandwidth than their counterparts in other parts of the world (Steiner et al 2005). Due to inadequate infrastructure and scarce resources, the Higher Education Institutions(HEIs) in these countries do not have the capacity to meet the growing demand of higher education, which is expanding exponentially (Goddard1998) throughout the world. Volery et al (Volery et al 2000) hold that capacity constraints and resource limitations can be overcome through the implementation of e-learning and creating a new opportunity to satisfy this growing demand in the mature student market.

Claudia (Claudia 2002) argues that all these obstacles together with high priced services, are obstacles which can be called a 'vicious circle' to Internet penetration in the country, and this vicious circle cannot be broken without decisive intervention of one or more of the above mentioned constraints. Claudia views International IP connectivity as a critical barrier to Internet, which is possibility for a user on an electronic network to communicate with other networks, and it precedes access to and use of Internet. He further adds that the width of this digital route is "bandwidth", i.e. the maximum amount of information (bit/sec) that can be transmitted along a channel (data transmission rate).

According to (infobrief 2003), by recognizing that bandwidth is a valuable institutional asset that needs to be managed, and conserved, this approach puts emphasis on how to explore ways to control and manage the many bandwidth hungry Internet applications, uses, and practices. However, as implementation of e-learning in HE institutions leads to fundamental shift in learning styles (Singh et al 2001), in that regards, (Bates 2000) argues that for the universities to initiate a change in their methods of teaching, an over reaching framework is needed to ensure that the use of technology needs are embedded within a wider strategy for teaching and learning. The existing frameworks do not address the issue of Internet infrastructure, among other limitations. Therefore for successful transformation, (Uys et al 2004) suggest that e-Learning needs to be implemented within a strategically developed framework based on a clear and unified vision and a central educational rationale, hence a need for a comprehensive and strategic framework with particular emphasis on bandwidth management by the HE institutions in low bandwidth environment which can facilitate the fundamental shift from once for life learning model to life long learning style.

Existing E-learning Frameworks

A number of Frameworks and models exist for implementation of e-learning in higher education. But they are not static rather they are dynamic and have evolved from classroom based teaching towards models that incorporate technology and pedagogical issues. Elmarie (Elmarie 2003) noted that, "While the first e-learning models emphasised the role of the technology in providing content, delivery and electronic services, more recent models focus on pedagogical issues". Some of the existing E-Learning frameworks and models are listed below:

Framework for Rutgers University (USA)

According to authors Triveni et al (Triveni et al 2003) 'The Learning Framework Study Group' recommended the framework proposed by Khan (Khan 2001) of The George Washington University for Rutgers University Libraries. Khan's framework as shown in Figure1 has 8 dimensions: Institutional, Pedagogical, Technological, Interface design, Evaluation, Management, Resource support, and Ethical considerations. Table1 briefly explains the dimensions of the framework. Khan's frame is located in Khan's publications, namely, B.H.Khan (Ed.), Web-based training (pp.355-364; Englewood Clffs,NJ: Educational technology Publications, which is translated

into many languages of the world and it is very popular especially in developed countries. Barry (Barry 2002) noted that, "Various issues within the eight dimensions of the framework were found to be useful in several studies that were conducted to review e-learning programs resources and tools".

Fig 1: An e-learning framework: Adapted from (Khan 2001)

Table1: An e-learning framework: Adapted from (khan 2001)

No	Dimension	Explanation
1	Institutional	Institutional readiness, Institutional matters, Collaboration, Administrative matters, Organisational, Academic, Infrastructure availability, and Planning
2	Technological	Availability of Technology Infrastructure
3	Pedagogical	Teaching/Learning requirements, Content Management Systems
4	Resource Support	Online, Offline technical support
5	Evaluation	Assessment of learners, Instructions and programs
6	Interface Design	Overall look and feel of E-learning programs
7	Management	Maintenance of learning environment, Distribution of information
8	Ethical considerations	Social and Cultural diversity, Copyright and so on

Khan argues that this framework can be applied to e-learning of any scope. Each dimension has further sub dimensions and each of these are inter related, e.g., after handling all matters concerning staff, students, and planning in Institutional dimension, next step is to put in place the necessary technology to support the e-learning programs, followed by e-learning teaching requirements etc. Author believes that a meaningful e-learning environment can be created for a particular group, by putting each stake holder group (such as learner, instructor, support staff etc.) at the center of the framework and raising issues along the eight dimensions of the e- learning environment as shown in Figure1.

Framework for Adaptation of Online Learning

The Framework for Adaptation of Online Learning was developed by Faridha (Faridha 2005). It is a modification of Bates (Bates 1997) ACTIONS model which has elements: Access (A), Cost(C), Technologies (T), Interactivity (I), Organization (O), Novelty (N), and Speed (S). Faridha grouped online learning issues into three categories; Educational, Managerial, and Technological.

Educational: This factor addresses the issues concerning; curriculum development, instructional design, and delivery.

Managerial: All organizational matters and constraints for implementation of online learning are looked at in this factor.

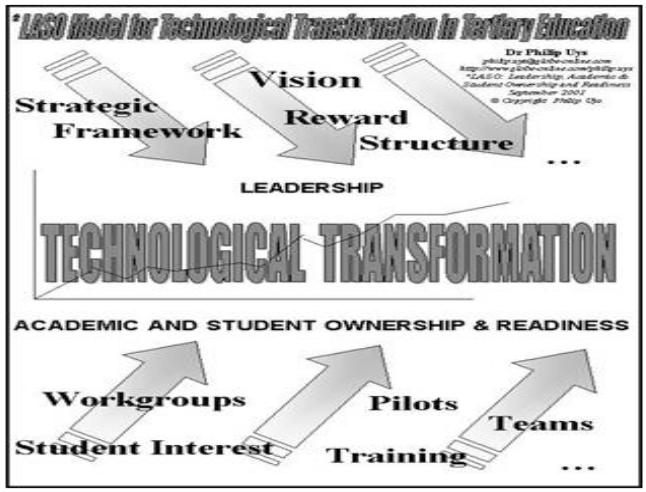
Technological: This factor handles issues like: access, integration, usability, and flexibility.

The LASO (Leadership, Academic and Student Ownership and Readiness) Model for Technological Transformation in Tertiary Education

The LASO Model for Technological Transformation in Tertiary Education is based upon major findings of the author Uys (Uys 2004) at his Doctoral research in New Zealand. The Model has the elements: Leadership, Academic, Student Ownership and Readiness, as shown in Figure2. According to this model technological transformation occurs when leadership is integrated with academic and student ownership and readiness, whereas Leadership is achieved through mechanisms which define a clear vision for the transformation, providing incentives for the staff engaging in the change process and the creation of a strategic framework to guide the transformation. The author argues that the strategies such as; pilot projects ,extensive training, establishing workgroups in every faculty/school, teams for courseware development should be used to achieve the Ownership and Readiness for change by both students and academic staff. A ragged line shown in the figure "signifies the complexities and dilemmas with which technological transformation is often associated".

The LASO model is proposed for developed and developing environment. MacNaught et al (in uys et al 2004) state, "The LASO model for technological transformation is one where management provides for the requisite vision, direction, organization, focus and control over the resources needed and thereby empowers the staff for action and ownership of the transformation". The model also includes an inside-out dimension as it attempts to address the affective domain such as motivation of staff and students.

Fig 2: A Framework for developing and developed environment: Adapted from (uys 2001)



A Framework for Success

The Framework for Success was proposed by Jennifer (Jennifer 2005) and has five elements: Technology, Content, Administration and support, Communication, and Financial analysis.

Technology: There are two types of technologies; synchronous and asynchronous. Synchronous technologies involve real-time interaction between an instructor and learner and they are like a broadcast with a time and a "channel" (web URL) for tuning in, and include webcasts, webinars, and chats, which can be recorded and replayed, and the recordings would be considered as asynchronous. The author argues that it is necessary to make IT department a partner in the technology decision making process.

Content: Content can be developed internally or can be bought from vendors. Therefore organizations should decide as what content to buy vs. build internally.

Administration and support: Further to content development what follows is administration and student support. According to author, it is necessary for someone to be there full time for student support to receive queries, issue of identity cards, and to facilitate registration process etc.

Communication: Two factors are to be considered when communicating e-learning strategy to learners; change management and marketing communications.

Financial analysis: Much emphasis should be put on financial analysis, as this is the factor which determines sustainability of the e-learning program, and financial analysis should include all related costs including; cost of technology, authoring tools, course development, support, and administration.

The demand-driven learning model

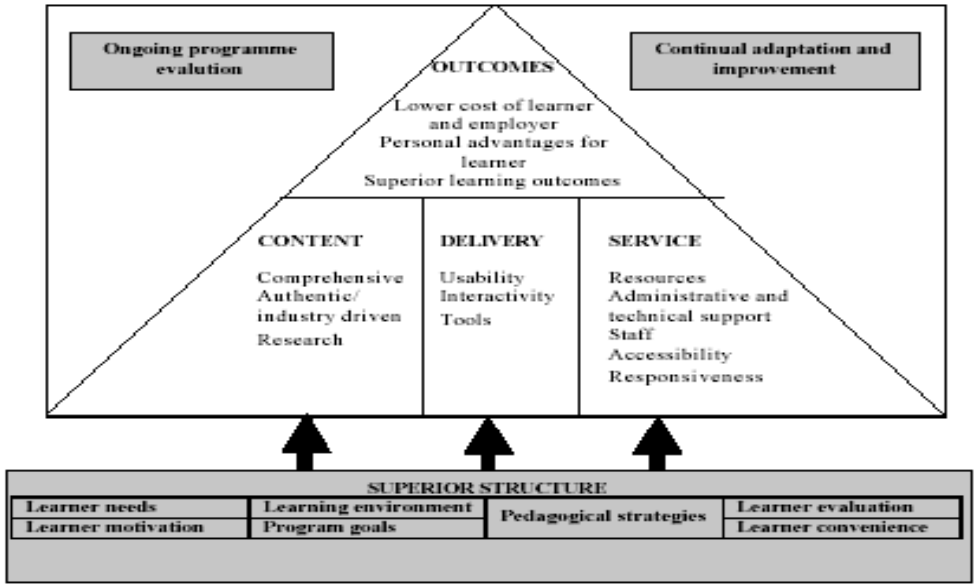
The demand-driven learning model was developed in Canada by MacDonald et al (MacDonald et al 2001) as a collaborative effort between academics and experts from private and public industries. It emphasizes the three consumer (learner) demands: High quality content, Delivery, and Service. As shown in Figure 3, the superior structure has its components; (learners needs, learner motivation), (learning environment, program goals), pedagogical strategies, and (learner evaluation, learner convenience). Other components of superior structure are: (content, delivery, and service), which result into superior learning outcomes (lower cost of learner and employer, personal advantage of learner), with emphasis on ongoing program evaluation and continued adaptation and improvement.

Content: Content should have qualities like; authenticity, comprehensiveness, and should be research based.

Delivery: A web-based delivery is recommended which should have user-friendly interface with communication tools to support interactivity.

Service: Service should include; the provision of resources for e-learning, administrative and technical support.

Fig 3: The demand-driven learning model: Adapted from (MacDonald et al 2001)



Critique to Literature Review

In terms of key factors important commonalities are identified between existing frameworks and models listed above; they refer to, and suggest, similar factors. Some frameworks such as (Macdonald 2001) and (Faridha 2005) combine several

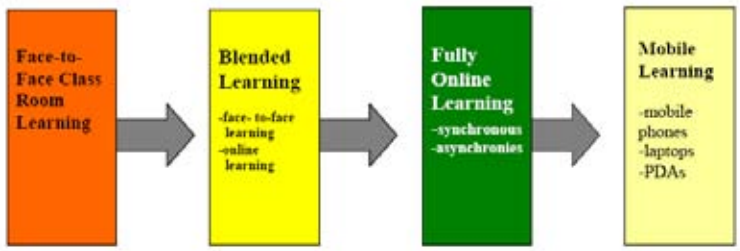
elements under one factor while others like (Jennifer 2005) and (Khan 2001)) refer to each of these areas as a key factor. In addition to that, it is noted is that some frameworks have a starting point whereas some other frameworks do not specifically point to a starting point when looking at e-learning implementation.

Finally, and more importantly, the reviewed literature reveals that the core factors; accessibility and connectivity (low bandwidth), irregular or non existent power supply, inadequate telecommunication infrastructure, high cost of hardware, software, economic conditions, and cultural issues, associated with implementation of e–learning in least developed countries, which fall under the category of low bandwidth environment and which need solution are not considered. Therefore, there is a need for a comprehensive framework which facilitates the implementation of e-learning systems within higher education in low bandwidth environment, addresses the above issues, and works as a road map for transformation of once for life learning style to life long learning model.

Proposed Gradual Transition Model for Implementation of e-learning in HE Institutions in LDCs

Implementation of e-learning technology in higher education provides a wide range of new opportunities for development by increasing flexibility in time and location of study (Ravenscroft 2001). Although e-learning has the potential to contribute in the educational advancement of developing countries, but the strategies and techniques of introducing it differ significantly than those used in developed countries (Ahmad 2004), due to different cultural and economic conditions. Therefore, to overcome such socio-economic and infrastructural constraints associated with these countries, a gradual phased transformation (Naidu et al 1996) from conventional face-to-face learning to e-learning is required in the context of university settings. In order to facilitate this gradual transition without compromising the quality of education provided by close classroom interaction, we propose a gradual transition model as shown in Figure 4. The model represents a continuum of educational technology integration into the various kinds of learning styles in higher education system. It originates from conventional face- to- face learning mode to the supplemental use of technology in the classroom, through blended or hybrid learning, to fully online distance learning environment, followed by Mobile learning (M-Learning) (Susan 2003).

Fig 4: Proposed Gradual Transition Model for Implementation of e-learning in HE institutions in LDCs



In order to enter into the arena of e-learning, according to our proposed Transition Model, the first phase after Traditional face-to-face learning mode is Blended learning. Harvey (in Balarabe 2006) defines Blended learning as, "a mix of different types of training; Synchronous and Asynchronous components, Instructor-facilitated and Self-paced components, and e-learning and Traditional face to face learning".

The next phase after blended learning is online learning or e-learning. There is no widely agreed upon definition of e-learning. However, (CTAL 2001) in a comprehensive sense, defines e-learning as, " Instruction and learning experiences that are delivered via electronic technology such as the Internet, audio- and video-tape, satellite broadcast, interactive TV, and CD-ROM. Web-based learning, computer-based learning, and virtual classrooms are some of the processes and applications used to distribute e-learning". E-learning can be Synchronous learning, a real-time, instructor-led online learning in which all participants are logged on at the same time and communicate directly with each other or Asynchronous learning, self-paced, in which instructor and learner interact with a time delay(Webb et al 2004).

The last phase of the Model is Mobile learning or M-Learning. M-learning can be defined as "learning that is mediated by mobile devices such as mobile phones, Personal Data Assistants (PDAs), handhelds, wearable devices and laptops"(Doherty 2003). M-learning is useful for administration and organization in higher education (Wood 2003) and can be used to complement other teaching and learning methods or to replace them. Many people across the globe have access to theses devices. On the other hand, currently, most developing countries do not have necessary infrastructure to support M-learning and it cannot be implemented as yet. Although, (Masters 2004) proposes that institutions within these developing countries should establish and commence mobile learning efforts as soon as possible.

Proposed Comprehensive Blended learning Framework

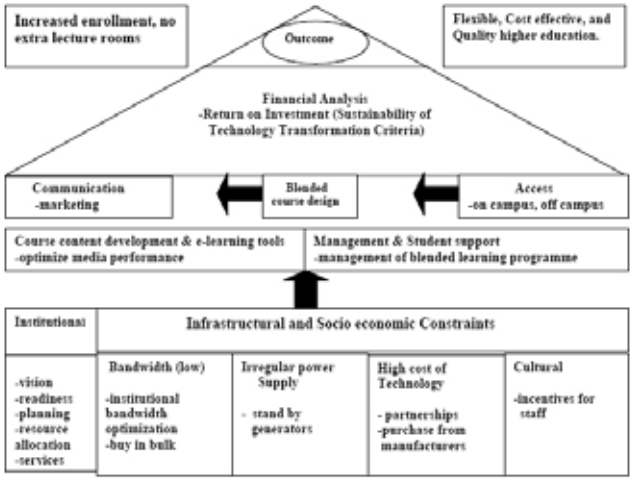
Our Proposed Framework is based on Khan's (Khan 2001) Blended e-learning Framework having dimensions; Institutional, Pedagogical, Technological, Interface design, Resource support, Evaluation, Management and Ethical. Harvey (Harvey2003)

argues, "Organizations exploring strategies for effective learning and performance have to consider a variety of issues to ensure effective delivery of learning and thus a high return on investment, while Khan's framework has capacity to serves as a guide to plan, develop, deliver, manage, and evaluate blended learning programs". On the other hand, in this frame work, many aspects of the socio-economic and technological environment such as connectivity (low bandwidth) and accessibility, inadequate telecommunications infrastructure, and lack of reliable power supply are taken for granted that need to be addressed during technological transformation in the context developing countries (uys et al 2004).

In terms of factors, our proposed "Comprehensive Blended Learning Framework adaptable by the Higher Education Institutions in Low Bandwidth environment" as shown in Figure5 is a modified form of Khan's Blended learning framework. The proposed framework considers those issues and shows that constraints like; scarcity of resources and inadequate infrastructure, including insufficient bandwidth cannot be the barriers towards the process of e-learning implementation (IICCIT 2006).

The proposed framework has dimensions; Institutional, Infrastructure, Bandwidth, Cultural, Content Development and e- learning Tools, Management and Student Support, Communication, Access and Financial Analysis. All the dimensions are inter-related and are in sequence of logical order. According to the framework, first step is to handle Institutional matters and putting in place the necessary Infrastructure, addressing low bandwidth and cultural issues and overcoming other constraints., which is followed by Content Development & e-learning Tools and Management of blended learning programme & Student support dimensions. Next, the developed courses are put online and are communicated, which are accessed by on campus and off campus students. The last dimension of the framework is Financial Analysis- the sustainability criteria of the technology transformation. The outcome of the implementation of the framework is increased enrolment with no extra lecture rooms for HE institutions and Flexible, Quality Higher Education at affordable cost for students.

Fig 5. Proposed Comprehensive Blended learning Framework adaptable in Low Bandwidth environment



Institutional: The Institutional dimension addresses issues concerning; organizational readiness, a vision for e-learning at the institution & development of technology development plan, formation of steering committee, human and financial resource allocation, staff and student affairs (khan 2001). Recruitment, library services, collaboration with other institutions, maintenance of infrastructure, and general administration, student ownership, copy right and learners needs, offering each trainee the learning delivery mode independently as well as in a blended program, are also part of the Institutional dimension.

Infrastructure: After the institutional vision for implementation of e-learning in their program offerings, readiness, and resource allocation, and handling other administrative matters, the next step is to put in place the necessary Technology Infrastructure. The term "infrastructure" according to (Blinco et al 2004), is highly contextual in its meaning and in e-learning contexts, "e-learning infrastructure", "technical infrastructure", and "ICT infrastructure" all convey a range of meanings. The basic requirement for implementation of e-learning is the availability of regular power supply, computers, telecommunication infrastructure, reliable Internet connection, and bandwidth. The necessary Technology Infrastructure also includes; high-speed access to the university network and the Web, including access from off-campus, provision of appropriate classroom technologies, and student computing abilities. Bates (Bates 1997) argues that, "While technology infrastructure strategy is absolutely essential, it is often the first and sometimes the only strategy adopted by universities build it and they will come".

Bandwidth: Bandwidth is part of necessary infrastructure for implementation of e-learning. But in LDCs the insufficient bandwidth that supports the educational needs of students and university, adversely affect delivery and teaching using e-Learning technologies that rely entirely on a high-speed campus backbone (Claudia 2002). In the implementation of e-learning process bandwidth is required by the institution for the development of e-learning course materials (content development), and by the learners who access those materials. Institutional bandwidth can be conserved through Bandwidth optimization. From the perspective of bandwidth, all media are not created equal. Asynchronous e-learning uses web based learning modules but does not support real time interaction between the instructor and the students. Synchronous e-learning consists of on-line real-time lectures, which typically have to be joined by students at the time of their delivery. Additional asynchronous functions typically support the learning environment. Most demanding in terms of bandwidth are forms of collaborative e-learning in which students have to interact continuously to solve problems or engage in other learning activities. Table2 below illustrates that only certain forms of e-learning require broadband support.

Table 2: Broad band requirement for e-learning: Adapted from (Bauer et al 2002)

	Application	Network demand	Complementary Functions and Tools
Asynchronous	Computer Based Training, Multimedia Database Support System	POTS ISDN	E-mail Automatic upload of Educational materials
Synchronous	Remote Lecture Room, Interactive Home Learning	Up to 6 ISDN channels, ATM, Internet protocol stack	Bulletin board, videoconference systems, e-mail, chat room, file exchange tool
Collaborative	Remote Interactive Seminars	Up to 6 ISDN channels, ATM, Internet protocol stack	Bulletin board, videoconference systems, e-mail, chat room, file exchange tool

Text and simple graphics can be downloaded quickly even in low bandwidth environment, whereas complex media require more bandwidth (Bruce 2001), which can be acquired through bandwidth optimization; exploring and controlling bandwidth hungry applications, filtering undesirable traffic from reaching backbone. Other possible solutions according to (ATICS 2004) are; formation of bandwidth consortium-which could cost half the cost of bandwidth, management of centralized network and technical capacity, improved regulatory policies regarding educational bandwidth. Although, bringing Internet access into the remote rural village in a least developed country is still a challenge different than optimizing it at a university, in its library, labs, and offices, or on the desktops of government or business officials in a capital city (David 2004).

Cultural: Implementation of e-learning changes the perception of teaching and learning, by providing entirely new educational culture (Karen 2006). It reconstitutes the roles for faculty members such that ; faculty members become e-Learning content developers, instructors, content experts, instructional designers, graphic artists, media producers, and programmers. Some incentives should be put in place to reward them.

Cultural change is a complex and one of the biggest challenging subjects in any medium and most particularly in the context of LCDs. People fear from the technology. Cultural awareness extends to appropriate design that takes into account the different learning styles. For example design that presents characters, thoughts, and speech in both audio and text format can address: accessibility to technology, different learning styles, and consideration of language needs for non native speakers of the language being used, and for native speakers with unfamiliar accent.

Another important factor to be considered in any training product design is learner motivation.

Content Development and e-learning Tools: Once the technology infrastructure is in place, the faculties interested in offering their programs in blended learning, are to develop and design the courses according to learners needs, offering each trainee the learning delivery mode independently as well as in a blended program (Harvey 2003).

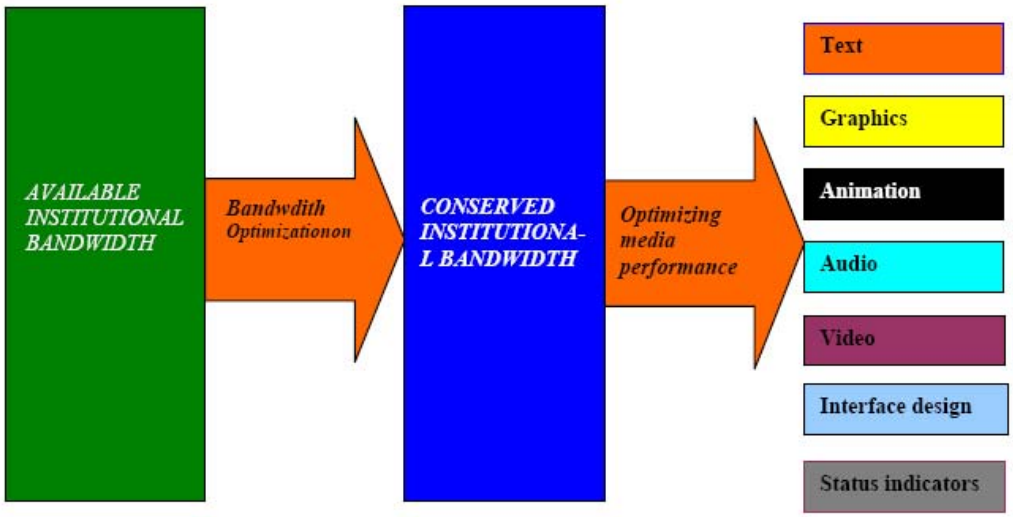
E-learning content development is a team work (John 2004). It includes: Instructional designers, Subject matter experts, Software developers, Graphic designers, Project managers, Database specialists, and Translators. Multimedia specialists, Distance learning specialists, professors, and instructors are also part of the content development team. Other team members are: Information security and privacy experts and Legal advisers.

Developed courses are to be put on line which can be accessed both by on campus and off campus students. Other requirements, according to (John 2004) include: Virtual Learning Environment (VLE), use of research facilities and resources, video and audio streaming, web conferencing, Computer Assisted Assessment (CAA). Learning Management System (LMS) and Learning Content Management System (LCMS) are also part of the e-learning tools. LMS is a program that manages the administration of training, typically includes functionality for course catalogues, launching courses, registering students, tracking student progress and assessments, and (LCMS) is a web-based administration program that facilitates the creation, storage and delivery of unique learning objects, as well the management of students, rosters, and assessment.

Proposed Two level Bandwidth Optimization Model

Figure 6 below shows Two level Bandwidth Optimization Model; at first level optimization of available institutional bandwidth by controlling bandwidth hungry applications and uses and at second level by optimizing media performance.

Fig 6. Proposed Two Level Bandwidth Optimization Model



Optimizing Media performance

There are two main ways to improve the course's performance: media optimization and streaming. The content that is presented in a continuous stream as the file downloads is referred as Streaming media. The streamed file starts playing before it has entirely downloaded. It is an effective way to deliver bandwidth-intensive content without making the user waiting. The streaming technologies can be used to reduce the bandwidth, but the rule of authoring is to make the courses small, which is called optimization. To optimize various media types effectively, techniques used, according to (Bruce 2001) are the following:

Text: Text files are small and perform well at low bandwidth, users can search for specific words, and content can be updated easily. Using anti-aliased text avoids having to create display text as a graphics file, which can make the course size much larger.

Graphics: Graphics are optimized by modifying file attributes, such as decreasing the resolution, size, and number of colours. Web graphics should have a bitmap resolution of 72 pixels per inch. Using graphics saved at a higher resolution will make the file unnecessarily large. The size of imported graphics should not be changed directly in an authoring tool and large graphics can be resized in an image-editing application.

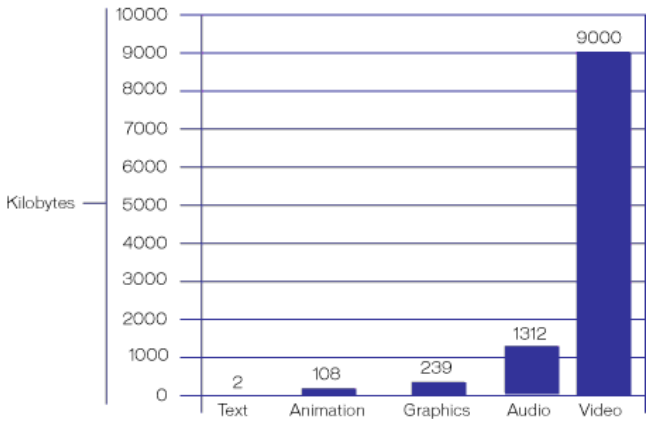
Animation: The animation file size is dependent on the size and file type of the graphics being animated. Techniques for optimizing animations are similar to those for optimizing graphics.

Audio: Large audio files can be optimized for efficient playback. Audio can be optimized by balancing sound quality and file size while musical audios by use of a short file that loops rather than one long audio track. Several files can be looped to play throughout your piece. Because Mono audio files are significantly smaller than stereo audio files, therefore files should be saved as mono unless it is necessary to use stereo audio.

Video: High bandwidth is required to download a video. Three standard digital video formats are: QuickTime, Video for Windows, and MPEG. Streaming video format, such as Real Video, requires a special server. Video files tend to be large and are not appropriate for delivery on modem connections- low bandwidth. Users can turn on bandwidth detection to automatically receive video at the highest quality their bandwidth supports. Video is captured, edited, and optimized in video editors such as Adobe Premiere. If video is too bandwidth intensive, it can be substituted with still graphics and audio, which will considerably decrease the size of your course. As Flash files are considerably smaller, substituting animation can also make downloading more efficient.

A careful decision should be made by the users when viewing various types of media, as text and simple graphics can be downloaded even at any available bandwidth. The chart below illustrates the file size required to present approximately one minute of media.

Fig 7: File size requirement for presenting one minute media: Adapted from (Bruce 2001)



Authoring guidelines

Following are some of the techniques to deliver low-bandwidth courses.

Interface design: It should be noted that a clean and simple interface design can make courses more compact. By using the authoring program's drawing tools, create large blocks of colour rather than importing bitmaps. The graphics

without gradients compress better than heavily shaded graphics. Moreover, only those graphics should be used that are necessary for learner comprehension. For self paced training, CD-ROM and DVD are usually created on 800 x 600 pixel screen resolution but online courses require a smaller screen resolution that will download quickly.

Status indicators: Status indicators should be added to inform users when they are waiting for files to download. Progress bars will show how quickly a file is downloading. Loader movies presented in a very small window, present a short introduction or entertaining animation that can hold the user's attention while files are downloading. The Loader movie can be a lightweight main menu that loads other portions of the course.

Management and student support: The Management dimension deals with issues related to the management of a blended learning program, such as infrastructure and logistics to manage multiple delivery types. Harvey (Harvey2003) argues that delivering a blended learning program is more work than delivering the entire course in one delivery type. The management dimension also addresses issues like registration and notification, and scheduling of the different elements of the blend.

Communication: After putting the courses online. There are some factors to consider in communicating e-learning strategy to your learners which include: Real time communication, change management and marketing communications (Jennifer 2005). Due to reduced bandwidth, which is the information carrying capacity of a communication channel, real-time communication is a bit problematic issue for LDCs. However, universities can collaborate, form partnership among them and buy high bandwidth in order to achieve this benefit which is very important. Universities should use both print and electronic media to publicize their on line courses. Some resistance in the beginning is expected.

Access: In order to implement e-learning in the universities, students are required to have access to computers and Internet, whereas access to computer technology is a major issue, particularly in developing countries. In these countries many students cannot afford to purchase computers and network access. Those who have computers have machines that are not suitable for multimedia or Internet access. Several strategies can be used to provide support for student access to computers, like providing computer labs on campus for students. Although, "It is a useful start-up strategy, but it becomes unsustainable in the long run as the primary source of student support but relying on computer labs for access has some drawbacks. For example, as the need to use computers for learning increases, either capital investment costs get out of control, or students' lining up for access reaches unacceptable levels. Secondly, with the technological change, computers in labs can get outdated (Bates 1997).Other strategies to increase the accessibility of computers and networks for learners are the development of government-funded

educational networks, equipped with advanced technologies including generators to help learners who stay in remote areas with non existent or irregular power supply.

More important, the main advantage of e-learning technology is flexibility- a learner can choose as where, when and how to learn but students to access learning from a specific place, often at a specific time, if they have to book, thus removing one of the main advantages of using technology- its flexibility. However, in the long run the most flexible and most cost-effective approach is to encourage students to provide their own computers and Internet access. Governments should provide loans to the students.

Financial Analysis: The return-on-investment is the most important factor, which determines whether e-learning program receives the investment it needs to succeed and grow. The financial analysis should include: Costs for technology, authoring tools, course development, support, and administration (Jennifer 2005).

Implementation of the Proposed Framework

The implementation process of the proposed framework include: Strategic target, Need analysis, Plan and Design, and Implement and Improve (Ingrid 2003).

Strategic target: Collis (Collis2002) analyzed the most frequent objectives of ICT policies in higher education institutions, among which he found objectives regarding pedagogical (enhancing quality of learning), economical (e.g. enhancing cost-effectiveness, generating institutional income), business (e.g. enhancing competitiveness, enhancing status and reputation of the institution) and organizational (e.g. enhancing flexibility) aspects.

Need analysis: In order to realize the strategic approach, it is important to know the specific change needs of the organization. Needs assessment program should also be concentrated on identification of possible courses to be delivered through e-learning framework based on market demand.

Planning and Designing: In addition to planning the technical, financial and organizational infrastructure, the human factor is a critical success factor. An innovation will only be adopted, if the key stakeholders are motivated and competent to manage the change. Motivation and ability have to be fostered on the individual, department and the board level (Ford 1996). Following are some of the planning considerations to be taken in this phase: learning and teaching (identify e-Learning scenarios for university teaching culture, establish a stakeholder-management), technology (building technical architecture e.g. selection of central LMS, support for standard authoring tools, provide networked workstations for staff and students).

Implement and Improve: The program based on the results of the needs assessments and satisfying the local constraints should be designed. The course design included the following components: On line components would be applied in limited extent

due to inadequate accessibility, course material would be presented on CD and used for self-study, e-mail would be used for disseminating information about the course, assignment, course upgrades. It would be mandatory for all participants to have an e-mail address, all materials and information would be available online as well as on CD, Discussion boards would be available for student communication, Facilities for communication with the instructors would be available and especial attention would be provided to ensure interaction and communication. Important aspect in the implementation phase is the identification and handling of resistance (Brake 2000). This research is still going on.

Conclusion

Many factors point the way to successful implementation of e-learning in the universities in low bandwidth environment. Learning those factors—both external and internal and how those factors work in the context of inadequate infrastructure will be essential to a sound strategy. This paper has provided a comprehensive review of e-learning strategy implementation literature and, based on this, proposed a blended learning comprehensive framework, adaptable in low bandwidth environment. In order to facilitate a gradual transition from conventional classroom-based learning to e-learning without compromising the quality of education provided by close classroom interaction, gradual transition model is also proposed. The Implementation process of the framework is also highlighted.

This framework is of high practical importance and usefulness. Its implementation is a means of reducing the cost of higher education, enhancing the quality of teaching, and widening access to flexible, and quality higher educational opportunity to millions of working and non working adults in least developed countries. It will constitute an important and novel source of new knowledge and provide a better understanding of the area to both researchers and managers.

References

Ahmad,El-Sobky (2004).“Emerging e-learning in Developing Countries: Challenges and Opportunities.” Information for Development (i4d), 2004

Alexander, S. (2001).Education + Training. In emerald Volume 43 Number 4/5 2001 pp. 240-24

American Heritage Dictionary of the English Language, Fourth Edition. Houghton Mifflin Company, 2004 aidworld (2006). Low Bandwidth Guidelines: Why Should Websites is optimized for Low Bandwidth? URL:http://www.aptivate.org/index.html (Retrieved on 12May 2007)

Balarabe, Y. (2006).The Effects of Blended Learning on Mathematics and Computer Attitudes in Pre- Calculus Algebra. In TMME, vol3, no.2, p.176 .The Montana Mathematics Enthusiast, ISSN 1551-3440, Vol. 3, no.2, pp. 17 6-183 2006

Barry, B. (2002).ISD and the e-learning framework. URL:http://www.wit.ie/library/webct/isd.html(Retrieved on January 24, 2003)

Bauer,J.M. P., Gai, J. Kim, T. A., Muth, S. S., Wildman (2002).Whiter Broad band Policy . Department of Telecommunication, Michigan State University East Lansing, Michigan 48824, USA.URL:http://tprc.org/papers/2002/72/Broadband_v1.pdf (Retrieved on 22 july,2006

Bates, T. (1997).The Carnegie Foundation for the Advancement of Teaching-What Kind Of University? 1997 London, Englan: Restructuring the University for Technology

Bates, A.W. T. (2000). Managing Technological Change: Strategies for College and University Leaders, San Franciso: Jossey-Bass. ISBN 0-7879-4681-8 CTAL (2001).Commission on Technology and Adult Learning (2001): A vision of e-learning for America's work force, Alexanria, American Society for training and development,Washington DC National Governors' Association (ED 455 432).URL:http://www.nga.org/cda/files/elearning report.pdf (Retrieved on 19 February 2006)

Blinco,K., Jon Mason,J., Neil McLean,N., Scott, W. (2004).Trends and Issues in E-learning Infrastructure Development: A White Paper for alt-i-lab 2004 Prepared on behalf of DEST (Australia) and JISC-CETIS (UK)

Bruce, B.(2001). Developing media for Low Bandwidth. In ASTD's Source of E-learning; Learning Circuits.

Brake, C. (2000). Politikfeld Multimedia. Münster, Waxmann.

Claudia Sarrocco (2002).Improving the IP connectivity in the least developed countries: breaking the Vicious circle. Claudia Sarrocco Journal: info ISSN: 1463-6697

Collis, B., van der Wende,Marijk (2002). Models of Technology and Change In Higher Education, Center for Higher Education, Policy studies CHEP,Twente. Url: http://www.utwente.nl/cheps/documenten/ictrapport.pdf)

Doherty, C. (2003) "IT goes top of the class," Birbeck University of London Press Releases October 8, 2003. URL: http://www.bbk.ac.uk/news/prarchive/infotech.html (accessed October 26, 2005).

Elmarie, E.(2003). A look at e-learning models: investigating their value for developing an e-learning strategy Bureau for Learning Development, Unisa Progressio 2003 25(2):38-47

Faridha, M. (2005). Towards Enhancing Learning with Information and Communication Technology in Universities; A Framework for Adaptation of Online Learning. A Research Dissertation Submitted to the Graduate School in Partial Fulfillment of the Requirements for the award of the Degree of Master of Science in Computer Science of Makerere University, Kampala-Uganda

Ford, P. (1996). Managing change in higher education learning environment architecture. Buckingham, Society for Research into Higher Education and Open University Press

Goddard, A. (1998). Facing up to market forces; Times Higher Education Supplement

Harvey, S. (2003). Building Effective Blended Learning Programs. Educational Technology, 43(6), 51-54

Ingrid, S. (2003). Sustainable implementation of E-Learning as a change process at universities paper presented at the Online Education 2003.

John, D. (2004). An E –learning strategy for the University of Warwick Version 2.1 [20.5.02] Information Technology Policy Committee

Jennifer, De Vries (2005). E-learning strategy: An e-learning Framework for Success. Blue Streak learning, quality learning solutions bottom line results.elearningcentre.typepad. com/whatsnew/2005/08/strat.html

IICCIT(2006).The proceedings of the Ist International conference on CIT for development, education, and training, UNCC, Addis Ababa, Ethopia , May 24-26, 2006

infobrief (2003). INASP (2003), "Optimizing internet bandwidth in developing countries higher education", INASP infobrief, 1 July 2003. available at: www. emeraldinsight.com/.../viewContentItem.do.

Khan,B.H and Ealy,D (2001). A framework for Web-based authoring systems. In B.H.Khan (Ed.), Web-based training(pp.355-364).Englewood Clffs,NJ: Educational technology Publications

Karen S.(2006).Challenges and opportunities for practicing E-learning Globally;Buckingamshire Christian University College, Buckinghamshire.

McNaught, C., Kennedy, P. (2000). "Learning technology mentors: bottom-up action through top-down investment", The Technology Source, November/ December, Url: http://ts.mivu.org/default.id=820

MacDonald et al 2001). A demand-driven learning model. In a look at e-learning models: investigating their value for developing an e-learning strategy. Elmarie Engelbrecht Bureau for Learning Development, Unisa Progression 2003 25(2):38-47

Masters, K.(2004). Low-key m-learning: a realistic introduction of m-learning to developing countries. URL: http://www.fil.hu/mobil/2005/Masters_final.pdf (Retrieved on 3 October 2005)

Naido S. and Oliver, M.(1996). Computer Oriented Collaborative Problem Based Learning: An Instruction -al Design Architecture for Virtual Learning in Nursing Education, Journal of Distance Education, Vol. XI, No.2, 1996

Norman, L. Latham, M. (2003). The Promise of E-Learning in Africa: the Potential for Public-Private Partnerships. IBM Endowment for the Business of Government.

OECD (2005). E-learning in tertiary education, The OECD Policy Briefs prepared by the Public Affairs Division, Public Affairs and Communications Directorate. United States OECD Washington Center Washington DC. 20036-4922

Ravenscroft, A. (2001). "Designing E-learning Interactions in the 21st Century: revisiting and rethinking the role of theory", European Journal of Education , 36 (2), 133-156.

Seufert, S. (2000). Trends and future developments: Cultural Perspectives of Online Education. In: Adelsberger, H; Collis, B., & J. Pawlowski (Eds.) International Handbook on Information Technologies for Education & Training, Springer, Berlin et.al., Germany

Singh, G. and Priola, V. (2001). Long distance learning and social networks: An investigation into the social learning environment on online students. Proceedings of the Sixth Annual ELSIN Conference. 158-164.

Susan E. M. (2003).The Ohio State university E-Learning Implementation Strategy and plan. Url:http://telr.osu.edu/resources/ITelearning.pdf.(Retrievedon 12 May 2007)

Steiner, R., Tirivayi, N., Jensen, M. Gakio, K. (2005).African tertiary institution connectivity survey. Nairobi: African Virtual University.UrL:http://www.avu. org/documents/Partneship9(Retrieved 24 Feb 2006)

Thriven K., Rebecca G., Roberta. (2003). Frameworks for Rutgers University Libraries.URL: www.rci.rutgers.edu/ ~kuchi/files/Recommendations%20of% 20the%20Learning%20framework%20Study%20Grou p.pdf dUys, P.M. (2001). Quality in the Management of Technological Transformation: A Framework for Developed and Developing Environments. Proceedings of the 2003 EDEN (European Distance Education Network) Annual Conference, 15-18 June 2003, Rhodes Island, Greece

Uys, P. M., Nleya, P., Molelu, G. B (2004). Technological Innovation and Management Strategies for Higher Education in Africa: Harmonizing Reality and Idealism. In Educational Media International Volume 41, Number 1 / 2004 Pp: 67 – 80

Uys, P., Cheddi, K. Mothibi, J. (2004). Implementing the LASO model: development of a pilot online course at The Faculty of Engineering and Technology, University of Botswana. In: Campus-Wide Information Systems,volume 21 number 3 2004 pp125-131

Volery and Lords(2000). Critical success factors in Online Education. The International journal of Education Management.

Webb, E., Jones, A., Barker, P. & Schaik, P. (2004). Using e-learning dialogues in higher education. Innovations in Education and Teaching International, 41(1), 93-103.

Wilson, B., Sherry, L., Dobrovolny, J., Batty, M., & Ryder, M. (2002). Adoption of learning technologies in schools and universities. In H. H. Adelsberger, B. Collis, & J. M. Pawlowski (Eds.), Handbook on information technologies for education & training. New York: Springer- verlag

Wood, K. (2004). "Technolgy for E-Learning: Introduction to Mobile Learning (M-Learning)," Fe first. URL: http://ferl.becta.org.uk/display.cfm?page=65&catid=192&resid=5194& printable=1(Retrieved on 12 May 2007).

25
Towards a website Evaluation Framework for Universities: Case Study Makerere University

Michael Niyitegeka

To what extent do organisational websites contribute to the business value? The development of an organisational website should be guided by the objectives/core functions of the organisation. Websites are tools that are expected to add value to the processes in the organisation. They present an interface between the wider community and the organisation. The assumption that is held when an organisation develops a website is that it has information that it wants the public to access without a hustle. This access should in one way or the other translate into business benefits[1] for the organisation. Academic institutions are in the business of creating and sharing knowledge and one tool that is available for usage is the website. Fortunately in Makerere University like many other Universities in Uganda and the world at large this tool is being deployed. All faculties in Makerere University have a website that is hosted and can be accessed. The question however is; are the websites of any relevance to the growth of the Faculty and University in terms of creating opportunities and disseminating knowledge? Presently there is no evaluation tool that can be used by faculty and university management to evaluate the websites. This paper attempts to propose an evaluation framework that could be used by management at the faculties and the University.

Introduction

Websites have become common phenomena across the world. They are used by different organisations to fulfil different objectives or purposes. Common to all websites is the purpose to provide information to the wider community that would have been otherwise difficult to access due to physical location of the organisation.

A website is a collection of web pages, typically common to a particular domain name or sub domain on the World Wide Web on the Internet (Wiki; 2006). A website is resident in cyberspace implying that it is not restricted by geographical location. "A website enables broad dissemination of information and services" (Ivory Y.M, & Megraw R.; 2005). It is this ability to reside in cyberspace that makes the website a critical tool for information dissemination.

[1] Business Benefits in this context means benefits that accrue out of more people being able to access information about the organisation and then use that information to make an informed decision to the benefit of the organisation.

An organisation's website will essentially provide basic information about the organisation. This information could include the history, mission and philosophy, products and services, contact, location, current events among others. "However, in many cases, it is difficult to see how a company's website aligns with its general business strategy and in extreme cases, a website's only purpose is to provide the company with a presence on the internet" (Human Level Communications; 2005). The implication is that there is limited value derived from such investment because it does not contribute to the wellbeing of the organisation.

Important to note is the fact that all websites have an intended purpose that they intend to achieve. In most cases the purpose is to disseminate information to the different strategic partners. You develop a website because you have information that you think adds value to the existence of the organisation. Websites enable in bridging the communication gap between the organisation and the clients/ customers. Websites therefore create and improve closer customer relationships, and thereafter generate excitement about the organisation (Kotler; 2000). To what extent is this excitement being created?

Organisations are investing in; developing and deploying websites because of the immense potential websites have in contributing to the growth of the organisation. In the era of the information age, people always want to make informed decisions about the available choices and thus the kind of information that is available on the WWW has significant implications on the value contribution of the website. According to University metrics-Global University rankings one of the premises held is that "an organisation becomes its website" This implies the totality of the web content of a university is more than a reflection of the total embedded knowledge of the university. (www.universitymetrics.com/g-factor).

Cognisant with the fact that globalisation is a reality now than ever before through the Internet, the ability to be accessed is a requirement if an organisation is to participate or be part of the global world. "No matter what the size or business niche of enterprise, the creation of the company's presence on the web is a must do action" (Tsygankov; 2004).

Makerere University is regarded as an International University and thus there are expectations from the world. Having a website is more than necessity, but the kind of information that you have on the website is more important than just having a website. The fact that the websites present an interface between the wider public and the organisation, the kind of information that you have on the website should enable decision-making in your favour. This implies that the kind/quality of information that is available on the website should be the concern of everybody in the organisation but most specifically management.

The core functions of a University are to teach and research and thus by default a university website should enhance these core functions. The credibility of a university website would therefore emanate from its ability to provide information about the academic programmes and its research competence or credentials. It is incumbent on management both at the faculty level and university to ensure that the websites enable them to promote their core function.

In the world of competition, the organisation that provides sufficient information to it's would be clientele is most likely to become the market leader. Whereas universities are essentially not business entities, being able to attract funders, students and parents entirely depends on what information they can access. Websites therefore present an opportunity for Universities and particularly Makerere University to avail as much information as possible to the wider community to enable the decision making process.

Makerere university has no specific guidelines for website development and thus individual faculties or units have the discretion to decide what their website looks like and the kind of information that is posted on the website.

It is therefore important that an evaluation framework be developed so that it is possible to evaluate the credibility of University Websites. This would enable management to constantly check content on the respective websites.

Rationale for the Paper

This paper is intended to propose an evaluation framework for university websites. The paper draws its strength from the fact that majority of the members of management both at faculty and university levels are not technically grounded to use the scientific models. Thus there is need to develop a framework that can be used at management level for quick decision making. The framework that is proposed therefore reflects the issues that are likely to concern management.

Methodology

A total of 17 websites were evaluated out of the 24 links on the University intranet as of November 2006. The evaluation focused its efforts on websites that originate from academic units. These websites were accessed through the University intranet i.e. http://webmail1.mak.ac.ug/cgi-bin/sqwebmail?index=1 . The parameters that were used in the evaluated originated from literature that was reviewed.

Web Credibility

The success of most websites hinges on credibility. Those who create web pages hope people will adopt specific behaviours such as the following; register personal information, purchase things online, fill out surveys, click on ads, contribute content to a community, download software, bookmark the site return often (Fogg, B.J.; *et al., 2002*).

Web credibility is relative and entirely depends on the person who is accessing the website. The quality of information on the website is fundamental to the credibility of the website. Quality is one of the universal goals of product and information designers as well as developers (Mandel; 2002). It is important to note that in most cases designing a website has been more art than science, leaving many web designers to rely on intuition alone (Fogg B.J.; *et al.,* 2001). Over emphasising the art rather than content quality renders the website irrelevant to the core functions of the organisation.

In a study of 2,500 person's, conducted to ascertain how users evaluate the credibility of websites; "design look" ranked highest followed by information design/structure and information focus (Fogg B.J.; *et al.*, 2003). From this study one is able to appreciate what the website visitors regard as important. The design look is important because it creates the first impression and makes one to continue searching for more information. How you present your information is equally critical and thus information design/structure requires attention because a number of websites have a lot or limited information that sometimes it becomes extremely difficult to navigate.

Inherent with web credibility is the issue of quality. "Quality is an intrinsic and multifaceted characteristic of a product. The relevance of each facet can vary with the context and time because people can change their points of view and update their references related to an object or a subject" (Albuquerque & Bechior; 2002). It is however important to note that benchmarks for quality could be established especially with regard web site quality.

University Websites-World Rankings

Webometric is an organisation that ranks university websites based on the publications posted on the web. The primary objective of webometrics rankings is to promote publications on the web by universities and other research institutions (Webometrics; 2006). Webometrics collects data from search engines such as google, yahoo, search.live, teoma. The methodology used by webometrics can be accessed from http://www.webometrics.info/methodology.html.

Results from webometrics as of July 2006 indicate that African universities are not doing very well in terms on web publications and hence web presence. The University of Cape Town in South Africa is only university website in the top 500 world websites. Whereas Makerere University is at number 17 in Africa it is at number 3,577, with Uganda Christian University following at number 2 in Uganda, 88 in Africa and 7,449 in the world (Webometrics; 2006)

Since webometrics concentrates on measuring rankings using publications, which is one of the core functions of a university it means that African universities have not, used the web as a tool to disseminate their research outputs.

Proposed Evaluation Framework Using Makerere University Websites

There have been attempts to come up with criteria for evaluating webpages;

http://www.library.cornell.edu/olinuris/ref/research/webcrit.html

http://credibility.stanford.edu/guidelines/index.html

These attempts have focused on the business/commercial websites. Webometrics specifically looks at publications available on the website and does not consider other areas of concern like appearance, content among others.

This paper intends to propose an evaluation framework that could be used by Organisational managers. This framework will enable persons in positions of authority to evaluate their websites and there after make informed decisions.

The framework used the following criteria to evaluate the unit websites at Makerere University. These attributes are not scientific for a reason that majority of the website evaluation frameworks are technical and do not enable management to take decisions independently.

1. **Privacy Policy**- Each website ought to have a policy that protects its information. Lack of a policy implies that your information could be used for or against you.
2. **Last Update** – This demonstrates how regularly the website is improved or even the validity and accuracy of the information on the website.
3. **Content** – The kind and quality of content as well as accuracy is examined.
4. **Authority** – The author of the content is examined. Looking at the domain name whether it is a valid name and that it can be trusted as the source of the information does prove enable proving the authority of the content.
5. **Appearance** – first impressions matter, does the homepage have any relevance with the activities of the unit? The way the homepage is designed will influence the desire to keep on searching for more information.
6. **Publications** – is there any link that has a list of publications of the unit? Internet publication is becoming more popular as compared to the hard copy approach and thus it is to the advantage of the academic units to make public their research reports.
7. **Contacts** – can the unit be contacted from the website? Since the website is an interface between the organisation and the public visitors to your website should be able to access persons in authority. It is important that options are availed such that people can easily contact specific officers.
8. **Any other Comments**- this is intended to provide other comments that may not have been captured under the seven dimensions.

Duration of the study

The websites evaluated in this study were accessed between December 11th 2006 and 3rd January 2007. If there are any changes after 3rd January 2007 they have not been captured by the evaluation.

Scope of the Study

The study was limited to units/faculties that have websites at Makerere University and can be accessed through the University intranet. Emphasis was placed on academic units within the University.

Approach to the Evaluation

Each website was visited in order to gather information about the specific website. The information gathered was tabulated in table 1. The proposed framework was used guide data collection. The author was the principal investigator in this study.

Presentation of Findings

Makerere University has a website as well as an intranet, http://www.mak.ac.ug and http://webmail1.mak.ac.ug/cgi-bin/sqwebmail?index=1 respectively. The university homepage has basic information about the university. This link will take the visitor to the different programmes under the academic units. There are 20 academic units of these 8 units have active links for undergraduate programmes and 12 for postgraduate programmes. The implication is that for the undergraduate programmes one can only access further information from 8 units and the rest are not accessible and 12 for postgraduate programmes. Makerere University homepage is the gateway to the University on the web; if there is not sufficient information then most likely further search would be inhibited.

In regard to research at Makerere University there is limited information that one would enable one to evaluate the University research competence. A lot of information on these links was posted when the websites were being designed and thus may not be a true reflection of what is happening at the university.

The findings indicate that majority of the faculties or units pay limited attention to what is on the website. Whereas a number of websites show 2006 as year of last update, in most cases one or two postings may have been done, and no attempt is done to review the existing documents. In almost all Faculties the course codes and related content that are posted on the respective websites are wrong.

Majority of the websites were designed by other persons who are not resident in the faculties and this could explain why majority of the websites are not up to date. This casts a question; who should be in-charge of the website? In majority of these units there is no specific person who is charged with maintaining and updating the websites on a regular basis. It is only the Faculty of Computing and IT, Technology and Forestry and Nature Conservation that have contacts for the web master. The Faculty of Veterinary Medicine Prof. E. Katuguka is still reflected as the Dean, almost two years since he moved to Graduate school (http://vetmed.mak.ac.ug). It is surprising that majority of the units do not have an active link of research or publication.

It is therefore evident from this simple analysis that majority of the units have not realised the potential of the websites and thus consider it as a one time activity. The way websites are designed is critical and a lot of attention ought to put into coming up with a design that will catch the eye and create a lasting impression.

Majority of the websites are static and lack dynamism of a modern website. The only website that exhibits some dynamic tendencies is the Faculty of Computing & IT website where the student's link is enabled with database applications. Students are able to view results, submit complaints, evaluate course lecturers and select course electives among others.

In order to create a web presence that is credible, it is incumbent upon management that sufficient information that is accurate is posted on the website. It is desirable that at the time of designing the purpose of developing a website is known. A website should be designed for users and thus keen interest ought to be placed in getting the right information to the intended recipients.

In conclusion, it is imperative that guide ought to be developed which can be used by the different units when developing their websites. The current situation or approach could actually be expensive since these websites occupy substantial space that could be for something else needless to add the website hosting costs.

This proposed evaluation framework, could be used by management and the Directorate of Information and Communication Technology (DICTS) to ascertain whether their websites are achieving the intended purpose.

Research Recommendations

1. Carry out a credibility study of the University websites in the region.
2. Validate/Improve on this evaluation framework.
3. Evaluate the cost of having a website that is not credible
4. Develop broad criteria for ranking academic websites.

Table 1: Table showing Faculty Website Evaluations

Unit Name/URL	Privacy Policy	Last Updated	Content	Authority	Appearance	**Publications**	**Contacts**	**Other Comments**
Faculty of Arts http://arts.mak.ac.ug	Not Available	2006	Course Codes are not updated. No course curriculum. Limited content available on the website.	Available	Relevant One can easily relate it to the Faculty.	Not Available	Available for Dean and Departmental Heads	No student link. The content is not recent as the website especially the course codes. There has been a redesign of the homepage without changing content.
Faculty of Social Sciences http://ss.mak.ac.ug	Has a disclaimer link although there is not content.	2006	Content is relatively well presented. Course codes are not updated. A number of links could not be accessed.	Available	Relevant The available information is easy to navigate through.	Not Available although a link is available on research.	Not available although there is a link on "contact us".	The Website is elaborate if all the links were active.
Faculty of Agriculture http://agric.mak.ac.ug	Available	17th /12/04	Course codes are not up to date. No course curriculum Faculty prospectus link has no content.	Available	Relevant, it demonstrate the activities undertaken at the Faculty.	Available under staff profiles last updated in 2004.	Dean's available	There are links that have no content. Home page still has information of 2005 conference advert.
Faculty of Forestry and Nature Conservation http://forestry.mak.ac.ug	Not available	4/05/06	Courses do not have adequate information. Prospectus is of 2003 and has only course outlines.	Available	Basic appearance, the banner is not complete. The design lacks a professional touch.	Available for only one person 1998-2002.	Dean's and administrator's available	Banner is not properly seen.

Faculty of Science http://sci.mak.ac.ug	Not Available	23/06/04	Academic programmes have no description. There is no link for students. The available content is basic about the different departments and what they do.	Available	Basic Appearance	Not Available	Dean Heads of department	Does not provide adequate information like subject combinations. The homepage still carries conference ad of 2003.
Faculty of Technology http://tech.mak.ac.ug	Not Available	Aug 2006	Updated course codes. An outline of courses available. Have adequate links with content.	Available	Relevant although the homepage is crowded with a lot of information.	Available for two persons	Available	Still has Prof. Ssebwufu as VC under the Gatsby project.
Faculty of Computing & IT http://www.cit.ac.ug	Not Available	**23/11/06**	Detailed curriculum available. Has an active News Link. Has an elaborate students links. Relatively easy to navigate through the information. Detailed calendar for the faculty programmes.	Available	Relevant, information structure and position is ideal, dynamic in nature.	Available as well as technical reports.	Available	The About CIT link should be moved upwards.

School of Education http://ww2.mak.ac.ug/educ	**No information available: "This domain is parked, pending renewal, or has expired. Please contact the domain provider with questions.**							
Faculty of Law http://makerere.ac.ug/law	Available	2006	Detailed course description available. Basic links are available. Staff profiles are available.	Available	Relevant although basic in design.	Available under staff profiles.	Available	Some of the links have information that requires updating e.g. the e-mail address for the Academic Registrar. acad-muk@infocom.co.ug
Institute of Public Health http://www.iph.ac.ug	Not Available	2006	Basic information available about the programmes. Elaborate information on the activities in the Institute is available.	Available	Good, there is a design flaw on the homepage it has not followed the conventional design of a homepage. The picture gallery on the homepage could be moved upwards.	Available under staff profiles	Available	Has links to other websites. Has links that have no content.

School of Graduate Studies http://gradauateschool.mak.ac.ug	Not Available	July 2005	Sufficient information about graduate programmes. Some units that were formerly institutes and now faculties are still reflected as institutes.	Available	Relevant with the relevant links.	Page still under construction	Available	Faculty links are not updated especially with new programmes. Conflict in the names of the school, Graduate School or School of Graduate Studies both are on the homepage. Some Faculties have new programmes that not reflective e.g. Arts & Computing & IT.

References

Albuquerque A.B. & Belchior A.D (2002); E-Commerce Website Quality Evaluation; Proceedings of the 28th Euromicro Conference (EUROMICRO'02) IEEE

Fogg B.J., Marshall J., Laraki O., Osipovich A., Varma C., Fang N., Paul J., Rangnekar A., Shon J., Swani P., Treinen M., (2001); What makes websites Credible? A Report on a Large Quantitative Study; Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems. New York; ACM Press. Seattle, WA (USA) 31 March – 5 April, 2001: 61-68 ACM Press

Fogg B.J., Soohoo C., Danielson R. D., Marable L., Stanford J., Tauber R. E., (2003); How do users Evaluate the Credibility of Websites? A study with over 2,500 participants; ACM

Fogg B.J., Kameda T., Marshall J., Sethi, R., Sockol, M., & Towbridge, T., (2002); "Stanford-Makovsky web Credibility Study 2002: Investigating what makes Web sites credible today." A Research Report by the Stanford Persuasive Technology Lab & Makovsky & Company. Stanford University. Available at www.webcredibility.org 16/12/06

http://arts.mak.ac.ug accessed on 16/12/06

http://ss.mak.ac.ug accessed on 18/12/06

http://agric.mak.ac.ug accessed on 18/12/06

http://forestry.mak.ac.ug accessed on 18/12/06

http://cit.ac.ug accessed on 20/12/06

http://sci.mak.ac.ug accessed on 20/12/06

http://tech.mak.ac.ug accessed on 20/12/06

http://ww2.mak.ac.ug/educ accessed on 27/12/06

http://makerere.ac.ug/law accessed on 3/01/07

http://www.iph.ac.ug accessed on 02/01/07

http://graduateschool.mak.ac.ug accessed on 02/01/07

Human Level Communications (2005); Measuring and Improving the Performance of a Website; Accessed on http://www.humanlevel.com/resources.asp?IdNoticia=18 on 17/05/07

Ivory Y.M. & Megraw R. (2005); Evaluation of website Design Patterns; ACM Transactions on Information Systems; (23), (4)

Kotler Philip (2000); *Marketing Management*; Prentice Hall New Delhi; Millennium Edition

Mandel Theo (2002); Quality Technical Information: Paving the way for usable Print and Wed Interface Design; *ACM Journal of Computer Documentation* **(26)**

Tsygankov A. Victor (2004); Evaluation of Website Trustworthiness from Customer Perspective, A Framework; *ICEC'04, Sixth International Conference on Electronic Commerce;* ACM

Webometrics (2006); World *Universities' Ranking on the Web*. Available at www.webometrics.info 03/01/07

Wikipedia (2006); Website; Accessed on www.en.wikipedia.org/wiki/website 15/12/06

www.makerere.ac.ug accessed on 03/01/07

26
Standards-based B2B e-Commerce Adoption

Moses Niwe,

This study presents the preliminary findings from an explorative study of Industry Data Exchange Association (IDEA) concerning challenges, benefits and problems in adoption of standard based business-to-business e-commerce. IDEA is a standards organization facilitating business to business e-commerce in the electrical Industry.

1. Introduction

Information and knowledge have become key strategic resources, upon which organizations across all industries make their decisions. Trends that have made information systems of strategic importance include globalization and competitive pressures for increased quality with lower costs (Chen, 2002; Clarke, 2001; Laudon and Laudon, 2006).

In the global business environment, businesses should see the enhanced role of electronic business as particularly increasing the importance of information systems. The Internet with its open environment, and other networks have made it possible for the organization to access and exchange enormous amounts of electronic information both inside in the organization and around the world with minimal time resulting in lower communication and coordination costs.

With the arrival of the Internet, words synonymous with "e", standing for electronic are the buzzword for state of the art products and services today, from e-traveling to e-banking (Alter 2002). The first idea of "e" goes back to the 1960s in electronic data processing. Later this concept developed as electronic mail and then electronic data interchange (EDI), which later transformed into electronic commerce (e-commerce) applications that dealt with the electronic transfer of funds in the early 1970's. However, the applications were limited to large corporations like the financial institutions that could afford the big expenses. These inter-organizational systems (IOS) expanded from financial transactions to other kinds of transaction processing and extended the types of participating companies to manufacturers, retailers, services, and other forms of business. As the World Wide Web arrived in the 1990s, it grew explosively with new forms of e-commerce. As a result both business to consumer (B2C) and business to business (B2B) examples of electronic business models have emerged.

Originally e-commerce was related with B2C, which dealt with basic forms of online purchase transactions. However B2B is bringing in more revenues. Exponential figures from Forrester, Gartner, IDC, Jupiter and other technology

related research groups predict that worldwide B2B e-commerce revenue alone will exceed $15 trillion by the year 2010 (Turban, E. et al 2006). Interest has broadened to B2B transactions that are governed by inter-organizational linkages, and high value. For this reason, B2B e-commerce technologies have a key strategic role in organizations across all industries in the global Internet based economy (Lawer et al., 2004; Porter, 2001; Thatcher and Foster, 2002).

There are many possible activities for e-commerce, such as: production, distribution, marketing, sale, etc. In fact, when discussing e-commerce, most people refer to either activities between an organization and its customers, B2C, or activities between two or more organizations, B2B. This thesis focuses on B2B. Business-to-business e-commerce, also known as electronic B2B or just B2B, refers to transactions between businesses conducted electronically over the Internet, extranets, intranets, or private networks. The working definition of B2B to be used in this thesis is electronic versions of documents in a standardized format such as X12 or the flat file equivalent being moved from computer to computer (Wise and Morrison 2000; Clayton and Waldron 2003). Hence, the term EDI is expanded to encompass various applications of B2B e-commerce. In this case an extensible markup language (XML) document; a meta language written in standard generalized markup language that allows one to design a markup language, used to allow for the easy interchange of documents on the World Wide Web is really EDI (Varon, 2002). Flat file or web form, Excel spreadsheets electronically are forms of EDI (Domaracki, 2001). This is an important issue to understand because most companies, especially small and medium sized, do not have the infrastructure to tag documents in the classic EDI sense. They do not want to spend the money on it because it is too expensive, difficult, and complex (Subramaniam, and Shaw 2002). Small companies want other ways to work around the costly issues but still get the benefits of doing B2B e-commerce. However, it is still believed that the general adoption of B2B is heavily influenced by small and medium sized enterprises (SMEs) (Wagner, 2003). The efficiencies and benefits are still in involving the SMEs because everybody gains that way. What we have seen in the last 5 to 6 years is a progression of EDI being redefined unlike the old fashioned way.

With this background, the principal objective of this paper is to examine the issues and challenges faced by firms in adopting B2B e-commerce. To understand the solutions qualitatively, we examine a standards organization in the electrical industry. The data was collected through electronic mail correspondences, interviews and company documentation. The case selected was largely due to their willingness to participate in this study. From previous work, (Niwe, 2007b), the organization adoption of X12 standard as the widely accepted U.S. B2B e-commerce standard in the last 20 years was used as the unit of analysis.

1.1 Electronic Data Interchange

Electronic Data Interchange as a type of IOS is a foundational block for understanding B2B e-commerce. In EDI, organizations use proprietary value

added network (VAN) infrastructure to share business document forms like invoice, and shipping schedule, between a sender and receiver computer, for business use (Riggins, and Mukhopadhyay, 1999). The condition is that both trading and business partners have to meet all the necessary basic requirements for communication. For this to happen companies involved have to make step by step refinement, to their business processes and systems. According to (Chan and Swatman, 2000; 2004) this is considered as the first step in the e-commerce implementation process.

1.2 B2b E-commerce

The value of B2B e-commerce technology as a solution to cutting costs and maximizing profits is appreciated by most firms in the digital economy. Traditional supply chains such as EDI with their inefficiencies have been responsible for the need for firms to find better options and hence the keen interest in the B2B e-commerce model that works at addressing these problems (Niwe, 2007a). There are documented benefits for B2B e-commerce adoption across all yielding industry sectors (Archer, and Yuan, 2000). As (Kehal and Singh 2005; Berthon et al, 2003) point out B2B e-commerce systems have resulted in lower transactions costs. (Lucking-Reiley and Spulber, 2001) mention productivity and efficiency gains. B2B e-commerce also has facilitated entry into new markets plus extension of existing markets. In addition, use of electronic representations of business transaction documents can reduce processing and handling, thereby reducing processing costs, data entry errors and cycle times as electronic commerce is based on background, system to system communication and document processing (Ratnasingam, 2002; Amit, and Zott, 2001). Therefore there is no doubt that this technology is beneficial. However, issues still remain for organizations, e.g. why we are not seeing the full benefits after all the hype about the tangible benefits this technology provides.

Assessing B2B involvement by industry reveals different patterns of growth by sector. Starting with the financial sector as the earliest adopters; for about three decades big banks have provided corporate clients with electronic banking services over private networks using B2B. However, this was a limited service to only a few collaborations due to the high cost involved (Yan and Paradi, 1999; Tassabehji, 2003). This changed over time with the advent of the Internet and rise in computer technology, because the traditional e-commerce moved to the web with all its advantages. Through the Internet, new products and services are reached and delivered with B2B e-commerce technologies, addressing the concerns of costs, and creating many more business opportunities. B2B performance for other sectors is better defined by region and private versus public platforms. For example, in the United States (U.S.) the Health sector, is probably one of the few sectors that has attracted government participation more than any other. Speculations could be based on the reasoning that the other sectors are more driven by the private rather than the public systems. Most of the sectors are driven by their functional needs. For example, the high need of connecting business partners and their goods and

services, internationally has caused the transportation sector to develop very fast. This is all aimed at making trade faster and cost effective. Other sectors driven by their business function include manufacturing, and apparel. The manufacturing sector is reported to be leaders in B2B e-commerce adoption (IIE Solutions, 2001)

There are literature sources on inter-organizational systems in the form of B2B e-commerce. However few studies have concentrated on the successful adopters of B2B e-commerce by industry and the companies' experience. Despite the predictions and promises of B2B e-commerce, it is still at the beginning of the adoption process (Gurunlian and Zhongzhou, 2001). Furthermore, regarding perspective, the continuous improvement of hardware and software changes the focus of organisational performance from technological to strategic issues. For many organizations in their quest to adopt B2B e-commerce technologies, emphasis has been placed on operational and implementation issues and ignoring the strategic aspects such as the industry pressure (McEwan, 2001; Chan and Swatman 2004; Gattiker, et al 2000). Hence understanding strategic issues in B2B e-commerce adoption for stakeholders has become important.

1.3 B2b Technology And Standards

B2B technologies used in web-based IOS standards includes system interoperability technologies such as applicability statement 2 (AS2), a file format specification about how to transport data securely and reliably over the Internet. File transfer protocol (FTP) is a protocol used to transfer files over a transmission control protocol/Internet protocol (TCP/IP) network, e.g. after developing the hypertext markup language (HTML) pages for a web site on a local machine, they are typically uploaded to the web server using FTP. However, B2B data usually is EDI messages though it may be of any other message type. AS2 specifies how to connect, deliver, validate and acknowledge data. It creates an envelope for a message which is then sent securely over the Internet. Security is achieved by using digital certificates and encryption. It has provided many benefits including removal of value-added network (VAN) costs

Standards are among the key technological factors for successful e-commerce transactions for different trading partners (Reimers, 2001). Because standards contribute to improving business processes, reducing purchase and inventory costs, increasing productivity and market efficiency, and taking advantage of new business opportunities with market intelligence techniques (Choudhury, 1997, Nelson and Shaw 2005, Medjahed et al, 2003). Using propriety standards business documents such as purchase order, invoice, shipping schedule, and claim submission are being exchanged via networks between the business partners. There is general agreement that adopting electronic communications based on standards is a goal worth attaining. However, when day-to-day business operations are effected, there are a variety of factors which force organizations to adopt other options. Governments and other standard-setting bodies, i.e standard developing organization (SDOs) have made significant accomplishments in developing standards to support B2B.

They play a coordinating role in developing the infrastructure necessary to support standard-based communications. SDOs are those organizations accredited and who operate under the procedural jurisdiction of ANSI in the U.S. and ISO internationally and who produce standards that are recognized as national body standards or ISO standards. For example the Institute of electrical and electronics engineers (IEEE) is an ANSI Accredited Organization, and they receive limited immunity from anti-trust in the US because their procedures are audited by ANSI on a regular basis to make sure they embrace the vision that ANSI endorses. There are a lot of organizations, including not just IEEE and the Internet engineering task force (IETF) but also organization for the advancement of structured information standards (OASIS) and the world wide web consortium (W3C), that are very well run, produce extremely valuable results, and are probably as open as (and sometimes more so than various accredited or recognized organizations. The IETF, W3C, OASIS and all other organizations are standards setting organizations, a super-set of organizations from SDOs, consortia, commercial joint ventures, alliances who create specification in collaboration with other organizations and entities other than themselves (Söderström, 2002).

Internationally electronic data interchange for administration commerce and transport an ISO standard for EDI was proposed as the worldwide standard. Another standard XML, supports B2B transactions and has become the format for EDI and Web services. ANSI X12 and OASIS are two of the most publicized cross-industry SDOs. OASIS is a non-profit, international consortium that drives the development, convergence, and adoption of e-business standards. Members themselves set the OASIS technical agenda, using a lightweight, open process expressly designed to promote industry consensus and unite disparate efforts. The consortium produces more web service standards than any other organization along with standards for security, e-business, and standardization efforts in the public sector and for application-specific markets (OASIS, 2007). OASIS is developing electronic business using eXtensible Markup Language (ebXML) for the formatting of XML-based business messages. Electronic business XML is a modular suite of specifications that enables enterprises of any size and in any geographical location to conduct business over the Internet. Using ebXML, companies now have a standard method to exchange business messages, conduct trading relationships, communicate data in common terms and define and register business processes (ebXML, 2007). RosettaNet is an example of an industry focused SDO organizations that aims at creating open e-business processes. RosettaNet is a non-profit consortium of cross-industry companies working to create, implement and promote open e-business process standards (RosettaNet, 2007).

2. Methodology

The study uses an integrated approach of qualitative techniques comprising in depth interviews and analysis of expert opinions. Qualitative research approaches are suitable for answering research questions of exploratory nature (Myers,

2005; Trochim, 2005; Miles and Huberman, 1984). We chose this approach, because the main body of empirical knowledge that is relevant to our research objectives is tacit – it lies in people's heads, experiences and work practices, most of which is not documented (Niwe, 2006a). Hence, in this study we interview experienced experts in the industry. We use interviews, e-mail correspondences and documentation reviews for data collection. Expert opinions are used, to have the sector's perspective central to the research process, thus providing means of validation of the research results (Niwe, 2006b).

To address the research objective data collection used a three-phased approach. The data was collected by a fieldwork study in the U.S for two months period during June and July 2006. The U.S is the region in the world responsible for the highest volumes of B2B transactions (67%), and highest volumes of B2B revenues which continue to drive the global adoption rate (McGann, et al 2005). The research process begins with developing the research questions that are used in data collection. Each interview was approximately forty five minutes. First they were asked about scope and e-commerce applications that they have implemented in their respective roles and future plans. Secondly, we asked them to recall general problems that they encountered from the adoption process, and then we specifically discussed the strategic issues. The interviews were organised in a pattern as to look out for similarities for the analysis stage. The interviewees are top managers such as, president or heads of e-commerce division. The interviewees were asked about their firms' e-commerce operation. The next step involved selecting the experts to be used and the appropriate data gathering techniques. This was accomplished with the help of an Interview guide, using structured and unstructured interviews with experts that were or are involved in the e-commerce implementation of IDEA. Documentation sources included the organization's documentation both past including archival records and present documentation. The next step involved evaluating and analyzing the data with the help of identifying keywords. These activities were iterative and to some extent simultaneous.

The case selected was largely due to its willingness to participate in this study. The adoption of X12 in the last 20 years was used as the unit of analysis. Also the entire supply chain was considered which includes manufacturer, distributor and retailer. The case examined aimed at building a richer understanding of a single supply chain. As not many research perspectives exist on the adoption of B2B e-commerce over the entire supply chains, single company cases are an appropriate research approach (Yin, 1989).

3. Idea And B2b Practices

Industry Data Exchange Association (IDEA) is the e-business standards organization professionally focused on the electrical parts industry consortium in the U.S. In 1998 Industry Data Exchange Association was created by National Electrical Manufacturers Association (NEMA) and the National Association of Electrical Distributors (NAED) to manage the development of network systems,

which foster e-commerce for their customers and members in the U.S. Though it is expanding to other industries on the retail side, the principal focus in the wholesale is electrical. Industry Data Exchange Association's principal B2B e-commerce network is Industry Data Exchange (IDX2), an Internet business communication service (extranet) that enables trading partners to exchange business documents such as purchase orders, advance ship notices and claim submission, securely very cost effectively. The IDX2 also enables the delivery and access to IDEA's industry data warehouse (IDW2). The IDX2 provides all of the traditional electronic data interchange services. IDX2 provides interconnections to the traditional VANs allowing customers to trade with partners that are not part of the IDX2 community as well as direct connect to other exchanges. We interviewed IDEA president, Mike Rioux and IDX2 Manager, Tom Guzik.

IDEA emphasizes providing e-commerce based solutions in standards, services and training. IDEA accomplishes this with the Industry Data Warehouse (IDW2), the Industry Data Exchange (IDX2) and e-Business data and transaction Standards developed by IDEA. Mike Rioux, IDEA President "If somebody calls up and wants a solution we give it to them if they want standards we give to them. If they want just information or training we give to them and our objective is make money like everybody else in the capitalist world so we are trying to sell them services. If we sell service then our owners reap the benefits and our owners have members who are distributors and suppliers. And they end up with a lower supply chain cost. Our model is a little different we are not a privately owned company we do not have stocks and dividends but we have rates and charges to the customer that belong to the organizations."

The IDEA service suite maximizes supply chain efficiencies for companies, allowing them to conduct business electronically with 100 percent of their suppliers, customers and strategic partners. Mike Rioux, IDEA President says "Our products are services intended to provide proven business data and business-to-business (B2B) solutions that drive down supply chain costs, slash cycle times and enhance customer satisfaction by cutting across the supply chain. We also offer them an opportunity to drive up their sales because that what it is really about, to sell more products."

IDEA started with opening up their networks in April 2001 with 50 customers doing over a million kilo-characters worth of B2B. Today they have over 270 customers and doing six million kilo-characters that they process across their B2B IDX2 network. This is done in various formats (versions of B2B). A web form on a web page allows a user to enter data that is, typically, sent to a server for processing and to mimic the usage of paper forms. Forms can be used to submit data to save on a server (e.g., ordering a product) or can be used to retrieve data (e.g., searching on a search engine).

In terms of the network, IDEA started with a frame relay network. Despite the positive aspect of security, the network provided, they had to deal with the disadvantage of hardware and software cost being too high, hence this turned out

to be a bad experiment. The results of this did not go well partly because the target group was limited to the very large purchase expense in large companies. In 2000, IDEA got an Internet based Applicability Statement 2 (AS2) communications network being among the very first adopters, when they launched in April 2001. When AS2 EDI over the Internet (EDI-INT) was just starting out it gave them a leading edge and since then IDEA has been looking at enhancements on this Internet based AS2 communication network and consequently watching a steady growth with B2B EDI solutions adopters.

In distribution, there is a lot more they can get out of e-commerce because distribution did not adapt the Internet based AS2 communication network. Retailers such as Home depot, and Sears all went on 100 percent with the AS2 communications network because of the volume of transaction that they could handle. Hence they are already fully mature in the adoption process doing peer to peer AS2 connections, XML and all their variations. Most of electrical part manufacturers are convinced of the benefit of EDI, flat file or B2B, thus growth is evident for IDEA. The lead adopters come in and as they become more technically secure in the way of doing B2B e-commerce transactions, there is room to move on to other avenues such as B2B integration.

IDEA's growth rate without sales and marketing is between 12 to 15 percent annually. IDEA has very few non-electrical companies using their IDX2 network for example grocery and pharmaceutical but they do not go out and sell the services to them. Such instances arise when one of their known customers compels their clients, that they could be selling electrical products to, for example, a warehouse, because it is good and saves them money, hence, as (Hart, and Saunders 1997) propose, issues of power and trust arise in the B2B EDI technology adoption. Tom Guzik, says "We just landed an account with a corrugated cardboard box company called Crew Wall, they are not in the electrical industry but still there going to benefit from the network." However IDEA does not focus on that market because they are owned by the electrical industry. It is noted that they do not have all the electrical manufacturers and distributors on their network hence their focus is still to maximize the electrical industry. Mike Rioux agrees that "the principal reason for the 12 to 15 percent annualized growth is the Internet." Value Added Networks (VANs) are expensive and require leased lines and if the Internet is used as the transport means to send an EDI document it takes some costs out of the equation. This is what the IDEA network does with the help of the Internet.

Issues include the telecommunication infrastructure, and challenges with the B2B tools being complicated. Furthermore IDEA, concern also includes the companies assuming that it is normal course of doing business without collecting matrix on error rate, e.g. orders or invoices. The business relationships between many companies have been unique to their business processes, hence their concerns over the new ideologies that the X12-XML would address all the bottlenecks in electronic business. In addition, concerns with data security and the reliability of the standard still arise.

4. Concluding Remarks And Implications For The Future

B2B is a very complex heavy set of message standards, with so many variations that they differ from network to network, from Industry to Industry, and are quite expensive to interface with all its variations. They have generally been designed for batch transmission and processing. Just adding this to the Internet does not really do much to solve the problem, or add any real benefit. It is would be like saddling the Internet with a paper-based system - e.g. sending faxes to people rather than emails. IDEA companies need to learn how to extend the core applications such as customer relationship management with their business partners.

In the management of the supply chain the key resource is the information format. For most companies the approach to adoption is being done the other way round. Like many other companies IDEA members are being pressured into adopting by the industry rather than seeing the benefits. Hence, failure rates attributed to failure to change the internal procedures. The B2B systems are not integrated with the internal systems. These adopters need to map strategic business process reengineering plans for their B2B technology to be improved.

From the study many further interesting issues arise, such as whether companies buy in to standard based B2B e-commerce as a method of adoption. Places shown where they do include hardware or software interfaces. Also, do vendors use standards if they can do something better on their own and sell it? Are purchase decisions impacted by the presence of standards? If the answer is yes, then we can proceed to the next question. Do firms that embrace standards in their products do better financially than firms that reject standards? There is no proof (that the billions that various industries pour into standards really produce some significant payback) that standardization rewards it proponents in the IT industry.

Furthermore, to advance the research of B2B e-commerce adoption beyond the U.S., we propose to look at a comparative study of government case studies of U.S., versus European Union and its member states. Eliminating paper based business transactions with its expenses has been the main motivation behind B2B adoption. As we have seen, the more capable larger organizations have seen tremendous growth in doing their B2B electronic transactions over the widely accepted standards of the US (ANSI X12) for the U.S. organizations. For European Union, a case study of Sweden and, UN EDIFACT, could be examined to compare the different firms, and present a synopsis of the adoption for the two predominant standards.

References

Alter, S. (2002) Information systems: foundation of e-Business, 4th ed., Upper Saddle River, NJ: Prentice-Hall

Amit, R. and Zott, C. (2001). Value creation in e-business. Strategic Management Journal, 22.

Archer, N. and Yuan, Y., (2000). Managing business-to-business relationships throughout the e-Commerce procurement life cycle. Internet Research, 10(5).

Ariba, Inc. (2001) "Ariba Selects Syncra for Supply Chain Collaboration to Broaden Commerce Platform and Deliver Improved Value Chain Efficiencies. Press release, Ariba, Inc., Sunnyvale, CA, February 28,.

Berthon, P. and Ewing, M., Pitt and Leyland and Naude, P. (2003). Understand B2B and the web: the acceleration of coordination and motivation, Journal of Industrial Marketing Management, 32.

Chan, C. and Swatman, P.M.C. (2004), "B2B E-Commerce Stages of Growth: The Strategic Imperatives," hicss, vol.08 (8), p. 80230a, Proceedings

Chan, C. and Swatman, P.M.C. (2000), "From EDI to Internet commerce: the BHP steel experience," Internet Research, 10 (1)

Chen, T. (2002). 'Globalization of E-Commerce: Environment and Policy of Taiwan'. Centre for Research on Information Technology and Organizations, University of California, Irvine, CA.

Choudhury, V. (1997). Strategic Choices in the Development of Interorganizational Information Systems, Information Systems Research, 8(1).

Clarke, S. (2001). Information Systems Strategic Management: An Integrated Approach. Routledge Information Systems Textbooks. Routledge, Francis and Taulor Group, London and New York.

Clayton T and Waldron K (2003) e-Commerce adoption and Business Impact, a Progress Report. Economic Trends ONS

Domaracki, G.S. (2001) the Dynamics of B2B e-Commerce, AFP Exchange, 21(4).

Gattiker, U.E., Perlusz, S. and Bohmann, K. (2000) Using the Internet for B2B Activities: A Review and Future Directions for Research, Internet Research, 10(2).

ebXML, http://ebxml.org,

Gurunlian, J., and Zhongzhou, L., (2001). E-commerce and development report. United Nations, New York and Geneva.

Henriksen, H. Z., Andersen, K.V., and Pedersen, T., (2002), IS innovation: Adoption of B2B e-commerce. in Towards the Knowledge Society: eCommerce, eBusiness and eGovernment. Proceedings of the second IFIP conference on eCommerce, eBusiness, eGovernment (I3E 2002) (pp. 569-81). October 7-9. Lisbon, Portugal.

IIE Solutions; (2001), Manufacturers lead B2B e-commerce adoption. 33 (7).

Kehal, H. and Singh, V. (2005), Digital Economy: Impacts, Influences and Challenges. Hershey, PA: Idea Group Publishing.

Laudon, K. and Laudon, J. (2006), Management Information Systems; managing the digital firm, 9th Ed, Pearson Prentice Hall.

Lawer et al (2004). A study of web services strategy in the financial services industry, EDSIG

Lucking-Reiley D. and Spulber D. F, (2001). "Business-to Business Electronic Commerce," Journal of Economic Perspectives, 15(1)

McEwan, L. (2001), "Lessons Learned the Hard Way: The Development of Web-based Business Networks," BIG Conference, online at: http://www.mori.com/pubinfo/pdf/lee5.pdf,

Medjahed, B. B., Bouguettaya B., Ngu A., and Elmagarmid. A. K. (2003).Business-to-business interactions: issues and enabling technologies. The VLDB Journal The International Journal on Very Large Data Bases, 12(1)

Miles, M. B. and Huberman, A.M. (1984). Qualitative Data Analysis: A Sourcebook of New Methods. Beverly Hills, CA: Sage.

Myers, M. D. (2005) "Qualitative Research in Information Systems," MIS Quarterly (21:2), pp. 241-242. MISQ Discovery, updated version, last modified: July 26 2005

Nelson, M., and Shaw, M. (2005) "Interorganizational system standards diffusion: The role of industry-based standards development organizations," http://www.business.uiuc.edu/Working_Papers/papers/05-0126.pdf

Niwe M., (2006a) "B2B E-Commerce Adoption: in the Financial Services Sector,"in Proceedings of the International Resource Management Associaion International Conference Medhi Khosrow-Pour(ed.), Washington D.C, U.S.A May 21-24, pp, 1075-1077.

Niwe M., (2006b) "Business-to-business electronic-commerce adoption: theory building," in Proceedings of the 5th International Conference on Perspectives in Business Informatics Research Lina. Nemuraite, Benkt Wangler, Rita Butkiene(eds.), Kaunas, Lithuania October 6-7, (2006), pp, 55-59.

Niwe M., (2007a). "Business to Business e-commerce adoption: a case study approach" in Proceedings of the International Conference on Web Information Systems and Technologies poster, Joaquim Filipe, Jose Cordeiro, Bruno Encarnacao and Vitor Pedrosa(eds.), Barcelona, Spain March 3-6, pp, 196-199.

Niwe M., (2007b) "Diffusion of the Business to Business transaction accredited standards committee X12 standards." To appear in proceedings of International Resource Management Association-upcoming, Vancouver, Canada May 18-23, (2007),

OASIS (2007), http://www.oasis.com

Porter, M. E. (2001), 'Strategy and the Internet', Harvard Business Review

Ratnasingam, P. (2002), "Perceived versus Realized Benefits in E-Commerce Adoption," Malaysian Journal of Library and Information Science, 7(2).

Reimers, K. (2001). Standardizing the new e-business platform: Learning from the EDI experience. Routledge, 11(4).

Riggins, F. J., and Mukhopadhyay, T. (1999), Overcoming Adoption and Implementation Risks of EDI, International Journal of Electronic Commerce, 4(1).

RosettaNet (2007) http://www.rosettanet.org

Söderström, E. (2002), Standardising the Business Vocabulary of Standards, In The ACM Symposium on Applied Computing, Madrid, Spain,

Subramaniam, C. and M. J. Shaw, (2002) "A Study of the Value and Impact of B2B e-Commerce: The Case of Web-Based Procurement," International Journal of Electronic Commerce, 6(4).

Tassabehji, R. (2003) Applying E-Commerce in Business, London, GBR: Sage Publications, Incorporated

Thatcher, S. and Foster, W. (2002). B2B e-commerce adoption decisions in Taiwan: The interaction of organizational, industrial, governmental and cultural factors. 38th HICSS.

Trochim, William M., "The Research Methods Knowledge Base," 2nd Edition. Internet WWW page at URL http://www.socialresearchmethods.net/kb. (version current as of January 16th 2005)

Turban E., King D., Lee J. K., Viehland D., (2006) Electronic Commerce: A Managerial Perspective, Prentice Hall

Varon, E., (2002) The ABCs of B2B, CIO Magazine, Available online at http://www.cio.com/research/ec/edit/b2babc.html.

Wagner, B. A., Fillis, I. and Johansson, U. (2003) E-Business and E-Supply Strategy in Small and Medium Sized Businesses (SMEs), Supply Chain Management: An International Journal, 8(4).

Warkentin, M (Editor) (2001). Business to Business Electronic Commerce: Challenges and Solutions. Hershey, PA, USA: Idea Group Publishing.

Wise, R. and Morrison, D. (2000) 'Beyond the exchange: the future of B2B', Harvard Business Review, 78(6).

Yan, G. and Paradi, J. (1999) Success Criteria for Financial Institutions in Electronic Commerce, Proceedings of the 32nd Hawaii International Conference on System Sciences,.5,

Yin, R. K. (1989), Case Study Research: Applied Social Research Method Series, 5. California: SAGE Publications, Inc.

PART 2

Information Systems

11
An Ontological Approach to Domain Modeling for MDA-oriented Processes

Dilip Patel, Michael Kasovski, Shushma Patel

There exists a gap between objects which exist in the real world and the elements which represent them in a software system. This paper addresses this gap with a domain modeling process able to construct, manage, and negotiate the appropriate domain concepts within a domain model. This process is based on an ontological approach and its constituting components. It was developed as a The Model Driven Architecture (MDA) provides a framework of models which can store the various aspects contained within a software system. Within the context of the MDA, the domain modeling process and models will be situated within the Computational Independent Model. This approach is illustrated by its application to a simple situational case study.

1. Introduction

Many organisations employ the use of software as a means to support their continued existence (Patel, 2002). Software can assist in accomplishing their day to day business activities. It can promote the communication and organisation of its employees. Customers can also benefit from the services provided by the system. Hence, the software system must be able to incorporate such issues in its development to provide a well suited system for the organisation.

There exists a gap between the existing business concepts in the real world and the ones which exist within the software (Daniels, 2002). Business elements require an appropriate representation in the software system which will determine how concepts within the system should exist and describe the relationships between one another. Domain modeling is a method which is concerned with the capture and representation of business concepts in a software system (Agrawal, Karsai, and Ledeczi, 2003). It provides an articulation of the meaning behind concepts which are implemented in the system. The resulting domain model acts as a formalised context behind the nature of the elements in the developing system.

The Model Driven Architecture focuses on the use of models as the central artifact in the development of software systems (Miller and Mukerji, 2003). It provides a hierarchy of models with the purpose of separating the aspects of the developing system into each model. The system progresses in the development lifecycle as it undergoes a series of transformations between models.

An ontological approach within the context of the MDA can provide the necessary depth in developing a domain model. The domain model should not be hindered by computational dependencies in order to exhibit the meaning behind related concepts in the system. Hence, the ontological approach will contribute

to the development of a domain modeling situated as part of the Computational Independent Model (CIM) of the MDA.

This paper will provide a domain modeling method which is able to be incorporated into a MDA oriented process. We first discuss in section 2 the underlying concepts which support this research. Namely, there will be a discussion on the notion of ontology, the model driven framework, and the relationship between them. We exhibit previous works which have influenced the development of our process in section 3. In section 4, the domain modeling process is described. We apply the process to a case study in section 5 to express its practical implications. Section 6 examines research currently undertaken in this research area. We discuss future research opportunities which would stem from this work in section 7. Finally, section 8 outlines the conclusions of the paper.

2. Underlying Concepts

This section discusses the theoretical underpinning of the concepts behind the development of the domain modeling process. The topics discussed will be the notion of ontologies, the MDA, and the relationships between them.

2.1 Ontologies

The study of ontologies has undergone much research over the past few decades (Guarino and Welty, 2002). It has contributed to the advancement of many fields within computer science, including AI, knowledge management and software engineering. There are many recent developments which address the various activities within a software development process (Benjamin, Patki and Mayer, 2006; Kaiya and Saeki, 2005; Hoss and Carver, 2006). An ontology provides many of the qualities which are essential for developing a valuable domain model.

Gruber can be seen as providing much of the foundations concerning modern ontology development. According to Gruber (1993), "an ontology is an explicit specification of a conceptualization". He believed that ontologies articulated a formalised account of what is said to exist within a given area of existence. This area of existence stems from its context, or purpose which dictates the representations constituting a simplified description of the real world. Relationships within an ontology form a bond between concepts which could then be interpreted as knowledge (Gruber, 1993). Contraints on ontological elements would ensure the semantic integrity of the model.

Guarino's (1998) seminal paper builds upon the foundations of ontological development, debating some of Greuber's views on the definition of an ontology. From his perspective, Guarino argues that it is the semantic meanings, or axioms which define the ontology, rather than its vocabulary. Since the axioms determine the validation of ontological elements, one must adhere to the axioms in order to be considered part of the ontology. According to Guarino, a vocabulary developed by concepts and relationships may envelop models and implementations which are outside the context of the ontology. The axioms provide a filter from which only the appropriate knowledge may be extracted.

Kaiya and Saeki (2005) apply Guarno's ontological theory in an attempt to analyse and evaluate requirements. Their ontology was created from the semantics and relationships existing within a given set of requirements. Their research does support the relationship between requirement and domain concepts. However, their process generates the ontology form the requirement, whereas we argue that the ontology should serve as the basis for defining a requirements model.

Nguyen and Corbett (2004) approach the notion of ontologies from a mathematical perspective. They formally define the notion of concepts, relations, and properties in order to mathematically analyse system knowledge. However, their formalisms add robustness to the model at the cost of model comprehension. Our process will place a stronger emphasis on model comprehension, as a reference to the meaning behind elements in a system model.

An ontological-based approach can provide the necessary tools and information to develop a simple, yet meaningful domain model. A separation of concerns must exist between the domain e and the system elements which will apply them. This alleviates complexity when understanding the meaning encapsulated within the domain concepts.

2.2 The Model Driven Architecture (MDA)

The MDA is a modelling framework developed by the OMG which address the development of software systems (Miller and Mukerji, 2003). Models in the MDA represent information about the system at varying levels of abstraction. Development processes which incorporate the MDA develop systems by the development and transformation of the system models until a complete software product is generated. The MDA was designed to encourage the reuse and interoperability of system models and their information to develop quality systems.

Each model within the MDA articulates the developing system according to a particular view (Miller and Mukerji, 2003). Each view consists of a level of abstraction, and a set of system concepts which are relevant to the model's purpose at that stage. This information is articulated as a metamodel, whose purpose is to formally define the platform upon which the model rests upon. Elements within the system models are defined by the semantics and formalisms associated by their corresponding concepts defined in the metamodel.

The models within the MDA are the Computational Independent Model (CIM), the Platform Independent Model (PIM), and the Platform Specific Model (PSM). Each model is dependent on its relative abstract model, yet is independent if its implementing, or more specific model. The CIM contains a model which is independent of any design, or computational implementations. It contains elements which articulate the system's requirements and environment. The PIM expresses the system independent of any implementing technology. The PSM expresses the designed system according to a specific implementing technology. As a system model transforms into a more specific mode, so too must the metamodel to accommodate the shift in level of abstraction. Eventually, the final

transformation generally transforms the PSM into a technology-specific format executable by the computer system.

This process develops a domain model which exists from a computational independent viewpoint. It serves as a reference of meaning behind the concepts existing in a CIM. This acknowledges the domain model's role as metamodel for the CIM within a MDA oriented process. As the CIM translates into a PIM, so too must the domain model transform to maintain the relationship between element and concept throughout the MDA framework.

3. Themes and Rationale

This section presents previous works which have influenced the domain modeling process. The underlying themes and rationale behind the development of the method will be discussed.

The framework is designed to address the notion of representing the domain of a software-based system. A domain in this paper is a representation of the elements which exist within the organisation and environment (Neusibeh & Easterbrook, 2000). The domain model may be subject to reuse across multiple projects. Also, the organisation may evolve over time to adapt to changing needs. Therefore, a domain model must contain mechanisms which will allow it to adapt alongside the changing business environment. The domain is an important aspect to communicating the nature of the organisation in order to create the appropriate system to address it.

A domain model would encapsulate the concepts which exist within the real world and their meaning, which would be incorporated within the system models (Daniels, 2002). The domain model does not contain the information concerning how and where the domain will be used by the system. Within the context of the MDA, a domain model would then be represented as a metamodel due to its use in encapsulating the meaning of system concepts.

This separation creates a distinction between the nature of the organisation and the system which will address the organisation. This separation allows for easier management of the domain independent of the system which will incorporate it. However, since the system addresses the needs of the organisation, its elements are dependent on the representations of the real world elements within the domain model.

Reuse has been shown to be a factor of great importance to the success of developing a domain model (Agrawal, Karsai and Ledeczi, 2003). The systems which are developed for an organisation should incorporate similar concepts which represent it in the domain. The reuse of a domain model would allow for quicker software development by the immediate deployment of a previously made domain model. Also, reuse of a domain model could provide easier communications across different systems due to the standardised understanding of the business concepts generated from sharing a common model (Miller and Mukerji, 2003).

An organisation may change over time to adapt to its changing needs (Patel,

2002). Since a domain model represents elements from the organisation, it must also be able to adapt in order to be consistent with the organisational changes. Therefore, the domain model should contain mechanisms to handle the evolution of models over time. Elements which were previously included in the domain model may require modification to incorporate subtle, yet previously unrecognisable properties. The model may also undergo an extention to encompass concepts whose importance was previously unknown to the organisation (Gause, 2005).

Domain models generally appear in the early stages of a development process, when there are multiple perspectives of the system and its environment (Lamsweerde, 2000). This would entail an elicitation process which would extract the necessary information from the real world into the model (Neusibeh and Easterbrook, 2000). Furthermore, in order to consolidate the different perspectives, the domain model must be able to be readily communicated and managed between project members. To ensure communication, a balance must be made between the simplicity of comprehending the model with the robustness of the information which it contains.

A domain model is most appropriate within the CIM of an MDA oriented process. The CIM was designed to hold the elements which express the requirements and environment of the system (Miller and Mukerji, 2003). Elements within a CIM are defined in terms of its real world representations within the domain model. Therefore, the domain model will act as the metamodel to the CIM. As the CIM transforms into a PIM, the domain model must also transform into a PIM-compliant metamodel. To accommodate this, elements within the model must adhere to MOF conventions and be UML-compliant.

4. Development method

This domain model will be situated as a metamodel within the CIM of an MDA-oriented process. It will contain the ontological aspects necessary to support the elements of a requirements CIM. The domain model in this work will serve as the context to and environment of the requirements as a representation of the real world organisation. It will be independent of its application as a PIM, hence it will not contain elements of data structures or operational behaviour. This process contains cyclical elements which pose to reexamine and improve the quality of the developing domain model. The domain model will be UML-compatible so that it can easily be transformable into a PIM metamodel for further development.

4.1 Elicitation and Documentation Phase

The purpose of this phase is to capture as much information from the organisation as possible for definition within the domain model. This phase is influenced by the requirements engineering activities in (Neusibeh & Easterbrook, 2000). In order to depict an accurate representation of the organization, the domain concepts should be elicited and modeled. This resembles Silva's (2000) preliminary phase in its purpose to gather information about the situation. The domain concepts

should correspond to an element existing in the real world. However, the domain concepts should not contain any attributes or properties. At this stage, importance is stressed on modeling as many concepts as possible to allow for thorough examination later. Once the concepts are defined, relationships between concepts will then be identified. Each relationship is an identified binary link between two concepts. The relationship will also include a description of the relationship in order to better comprehend the connection between concepts. Since the domain model is independent of the system design, cardinality will not be defined on any of the relationships. Once all of the possible relationships are identified, the constraints will then be incorporated into the model. Constraints will determine any properties of the domain which represent similar limitations or rules which exist in the organisation. Constraints in this model will contain a description of the constraint and the domain elements to which it belongs. A constraint may be attributed to a single domain element, to multiple elements at once, or as a global constraint applicable across the domain.

4.2 Domain Audit Phase

This phase is concerned with the examination and improvement of the domain model. Within the first phase, emphasis was placed on eliciting as much information about the domain as can be found. In this phase however, the domain modeler who had elicited the elements will now examine them to determine their relevance to the domain and to remove apparent inconsistencies. The model has a greater chance of undergoing a more rigorous assessment due to the modeler's previous experience with the domain. Elements should be assessed in the same order to which they were elicited; domain concepts, then their relationships, and finally the domain constraints. This ensures a uniform strategy towards the domain assessment. When an element in the domain model is deemed to be unnecessary to the domain model, any directly associated elements must be reviewed to determine whether it still applies to the model. For instance, a constraint related to a removed concept should be examined to determine whether it still applies to other elements within the model or be removed along with the concept. In the case which a concept is found to not be relevant within a domain model, it should be removed from the model along with any associated relationships. This removes any unrelated elements from the domain model. Any constraints related to the removed concept should be examined to determine whether it still applies to another element within the model or be removed along with the concept.

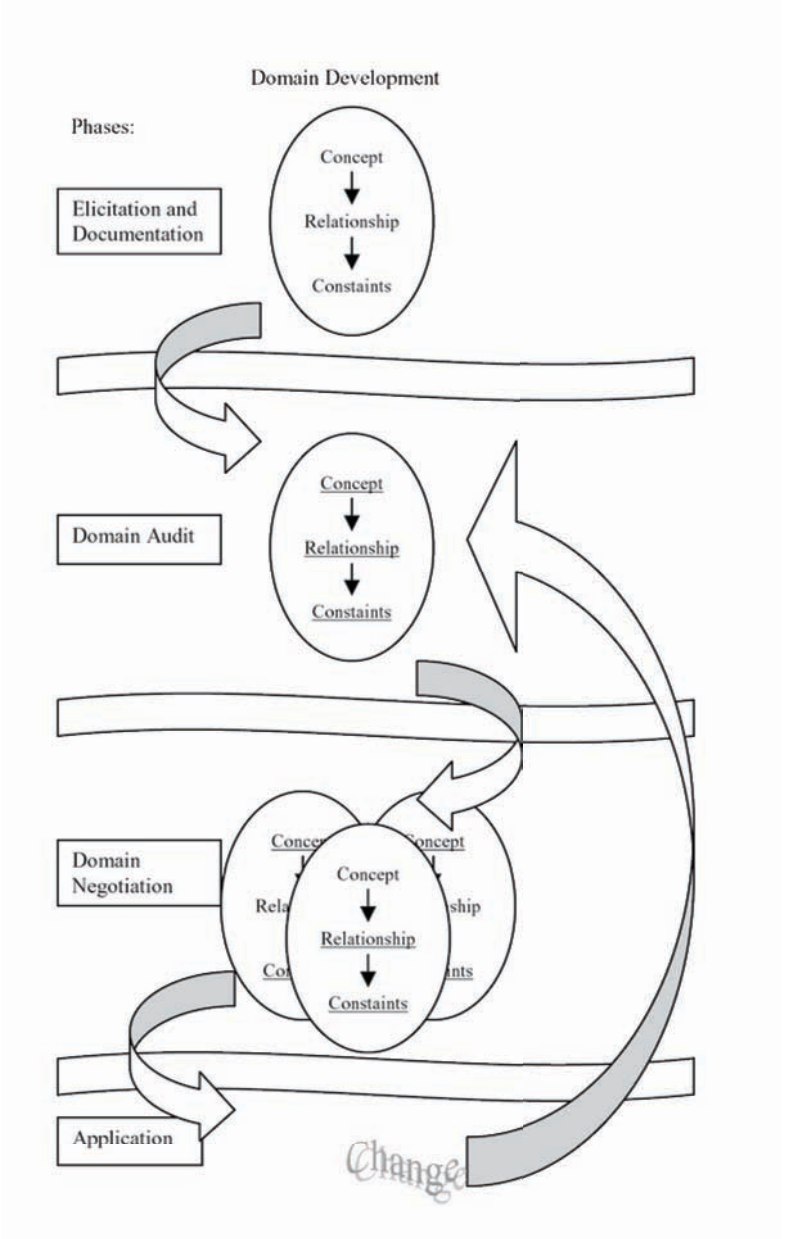
However, in some occurrences removed elements may turn out to exhibit its relevance under a future examination. To minimize the effects of such errors, a mechanism will exist which will contain the removed errors. Each removed element would still possess its information as well as the related elements where the element was placed before the removal. If the situation arises where a removed element is found to have belonged to its original place, the replacement of the element back into the model would be easier to manage.

4.3 Domain Negotiation Phase

Different participants in this development process observe the domain through their own perspectives (Silva, 2002). This may lead to various interpretations of the same domain. A consensus must be made on the nature of the domain model among the participants to ensure a consistent understanding of the environment and its meaning. Once a domain model has undergone examination by the modeler, it must then undergo a collaborated inquiry among all of the participants. Any domain models which other participants have developed will aid in determining points of comparison and difference in interpretations. When negotiating the different interpretations of the domain, any removed elements should be placed separately with a description over its previous placement, similar to those removed in Phase 2. Once the domain is agreed upon, it is ready for application and referrence by elements within the CIM.

As time passes, the real world organisation may undergo changes to adapt to the evolving environment. This may prompt a modification of the domain model to reflect this change. The model is sent back to a domain expert to determine the most appropriate alteration and its effect on the domain. Once the modification is made, it undergoes another mediation among the participants to consolidate the new interpretation of the domain for further use.

Figure 1. The Domain Development Process

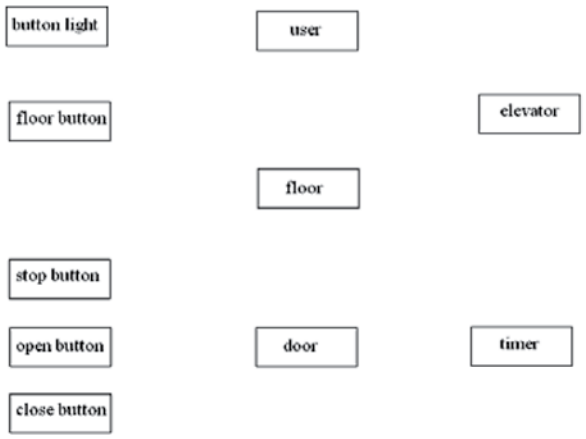


5. Case study

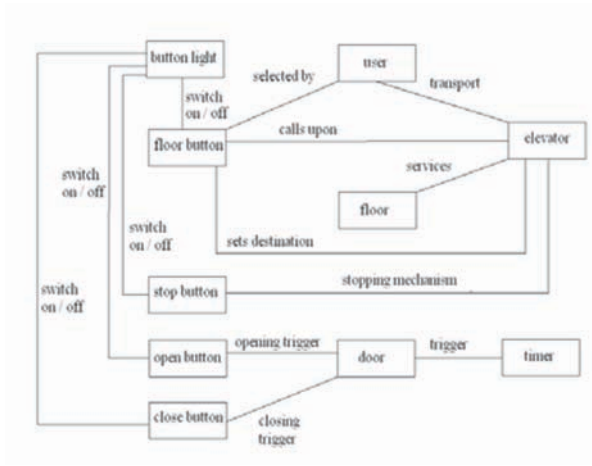
This work will present its approach on a case study examined in (Hoss and Carver, 2005). It provides a simple scenario which can clearly exhibit the domain modeling process and visualised representations of the developing domain model. The following figure will show the progress in the initial phase of the process.

Figure 2(a, b, c). Domain model at various stages during the Elicitation and Documentation Phase

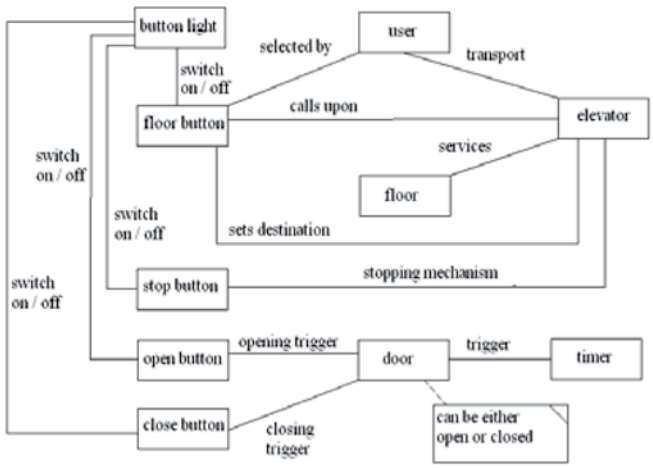
(a)



(b)

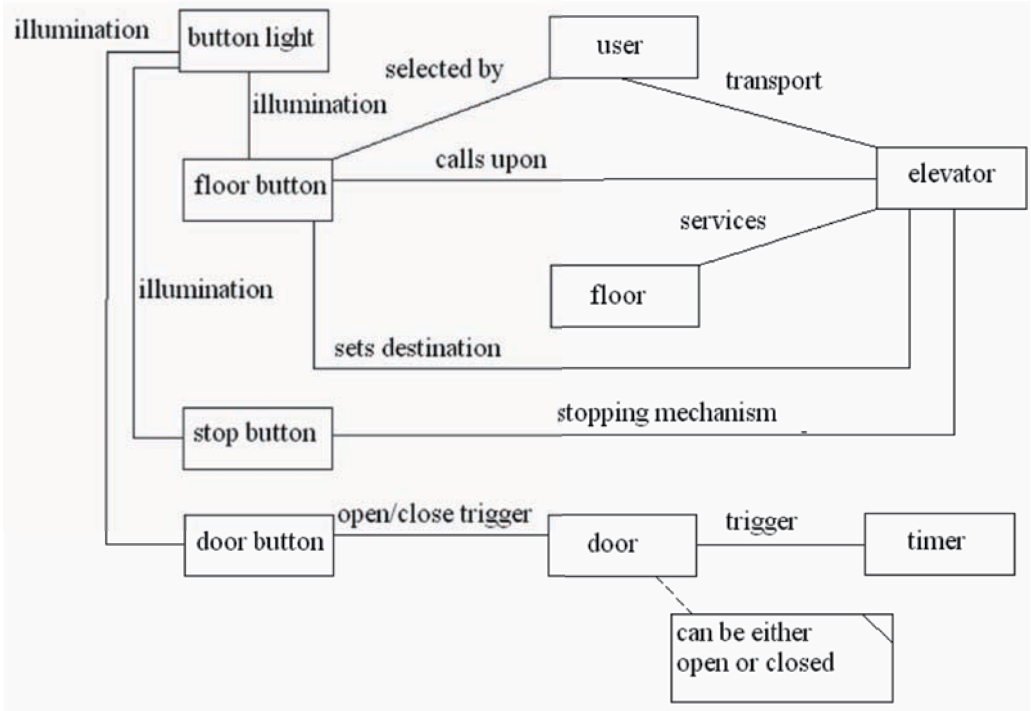


(c)



In figure 2a, the domain expert begins by identifying the many concepts on the model. Any concept which may be uniquely described, such as the open and closed elevator buttons are expressed separately on the domain model. Figure 2b exhibits the domain model after the initial attempt at relationship modeling. The relationships specified in this model are articulated in terms of phrases which express modeler's understanding of the domain expressed in the case study. Finally, this phase concludes with the addition of constraints seen in figure 2c. A constraint was added on the door concept which limits the possible states of the door to being open or shut.

Figure 3. Domain model after Domain Audit Phase



Once the modeler becomes familiar with the domain, he or she may provide insight onto improvement of the initial domain model. One major change consists of the consolidation of the open and shut buttons into one concept encompassing the trigger(s) for elevator door movement. The open and shut buttons are therefore removed from the model, however are still stored for future reference. Also, the relationships between the buttons and their lights may have been better articulated as one illuminating the other rather than a simple switch. Once the modeler is satisfied with their domain model, it is brought up for discussion among the other participants. There, it undergoes the scrutiny of other people as well as other domain models which may have been made by others. Once completed, it is ready for application and reference within the CIM.

6. Related works

Model Integrated Computing (MIC) is an approach which employs domain concepts in a framework of models in order to develop software (Szipanovits and Karsai, 1997). From their perspective, domain models in MIC articulate a system's design. They are based on the formalisms and language constraints expressed in the metamodel. The intricacy behind the language formalisms in the metamodel increases the amount of information contained in the metamodel at the cost of complexity.

Moon, Yeom, and Chae (2005) emphasize the importance commonality and variability within domain requirements. Their research presents a domain as the set of similar constructs among a group of products and the variants which may exist from product to product. The concepts within the domain do not accommodate for constraints, but rather are articulated a part of the domain requirements. This weakens the separation of concerns as the domain requirements become burdened with the additional complexity entailing the inclusion of constraints.

Garrido et al. (2007) approach domain ontologies and developing computation independent models from a groupware perspective. Their research stresses on the collaboration and communication aspects of knowledge sharing and requirements specification. However, they do not emphasize on the issues of ontology complexity and its impact on the development process.

7. Implications and Future work:

This paper introduced a process which addresses the notion of domain modeling for incorporation within an MDA-oriented process. It delves into the articulation of real world concepts at the early stages of development, where the system is independent of computational constraints. At this level of abstraction, due to the lack of computational rigour, a domain model may be evaluated according to the degree to which it can communicate its knowledge to other people. What also should be considered is the extent to which another person is able to comprehend about the domain and about the system from the domain model.

The process contains elements which promote the idea of collaboration with other participants and iterative domain development. A study could be constructed to determine the capabilities of such a process as a learning tool towards understanding the organisation via the domain model. This may lead to future applications of learning-based methods into MDA-oriented processes. Not only would the learning curve to MDA processes decrease, but it could aid in bringing different participants such as stakeholders to directly be able to contribute to the system development process.

The domain modeling process proposed in this paper contained a mechanism which would store previously removed domain elements for future reference. The paper outlined the necessary information to be stored, yet does not delve into the effectiveness of incorporating such a mechanism. A question which may be

raised is in regards to the amount of value to which such a mechanism may bring to the development process, and where best to deploy it.

This approach discusses the metamodeling aspects of developing a CIM appropriate for an MDA oriented process. However, this paper does not explain in detail the nature of the model which will employ the formalisms developed in this process. Future work may include developing a requirements model which would be able to capture the needs of the system, while articulating them in terms of relevant domain terminology. Such a framework would then be applied to MDA-oriented processes to examine the extent to which the models can integrate and the value it adds to model driven development.

8. Conclusions

This paper focused on developing a domain modeling method suitable for an MDA-oriented process. We identified a link between ontology development and model driven development using the common notion of domain modeling. The process promoted iterative and collaborative development of models representing the elements present in the real world. It promotes the idea of incorporating domain development into a CIM as a component of an MDA-oriented process. We also exhibited the process by applying it to a basic case study and observing the domain's development. This paper supports the notion that ontology-based domain development can become a valuable asset in capturing and comprehending the meaning behind model driven systems development.

References

Agrawal, A., Karsai, G., and Ledeczi, A. (2003). An end-to-end domain-driven software development framework. Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications. Anaheim, CA, USA, ACM Press.

Nuseibeh, N. and Easterbrook, S. (2000). Requirements engineering: a roadmap. Proceedings of the Conference on The Future of Software Engineering. Limerick, Ireland, ACM Press.

Daniels, J. (2002). «Modeling with a sense of purpose.» Software, IEEE 19(1): 8-10.

Gause, D. C. (2005). «Why context matters - and what can we do about it?» Software, IEEE 22(5): 13-15.

Gruber, T. (1993). "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, vol. 5, no. 2, pp. 199-220.

Guarino, N. (1998). Formal ontology and information systems. In Proceedings Formal Ontology and Information System.

Hoss, A. M. and Carver, D. L. (2006). Ontological approach to improving design quality. Aerospace Conference, 2006 IEEE.

Kaiya, H. and Saeki, M. (2005). Ontology based requirements analysis: lightweight semantic processing approach. IEEE Fifth International Conference on Quality Software, 2005.

Karsai, G., M. Maroti, et al. (2004). «Composition and cloning in modeling and meta-modeling.» Control Systems Technology, IEEE Transactions on 12(2): 263-278.

Ledeczi, A., et al. (2001). «Composing domain-specific design environments.» Computer 34(11): 44-51.

Miller, J. and Mukerji, J. (eds.) (2003). "MDA Guide Version 1.0.1" OMG. http://www.omg.org/docs/omg/03-06-01.pdf.

Moon, M., Yeom, K. and Chae, H. S. (2005). «An Approach to Developing Domain Requirements as a Core Asset Based on Commonality and Variability Analysis in a Product Line.» IEEE Trans. Softw. Eng. 31(7): 551-569.

Nicola, G. and Christopher W. (2002). «Evaluating ontological decisions with OntoClean.» Commun. ACM 45(2): 61-65.

Patel, N. V. (2002). Global ebusiness IT governance: radical re-directions. Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002.

Perakath, B., Patki, M., and Mayer, R. (2006). Using ontologies for simulation modeling. Proceedings of the 37th conference on Winter simulation. Monterey, California, Winter Simulation Conference.

Sztipanovits, J. and Karsai, G. (1997). «Model-integrated computing.» Computer 30(4): 110-111.

van Lamsweerde, A. (2000). Requirements engineering in the year 00: a research perspective. Proceedings of the 22nd international conference on Software engineering. Limerick, Ireland, ACM Press.

12
Enhancing Immunization Healthcare Delivery through the Use of Information Communication Technologies

Agnes Semwanga Rwashana and Ddembe Willeese Williams

The role that Information Communication Technologies plays in improving the efficiencies and effectiveness of healthcare delivery and particularly immunization coverage through the use of information and communications technologies has been well established. The paper examines the effectiveness of current immunization systems and challenges, then goes on to examine broader views regarding the interplay of political, social, economic and technology forces that influence the level of immunization coverage. It is this inter-play of forces that makes the problem of immunization coverage complex and also affected by time delays to deliver some of the functions. The paper presents the challenges in the current immunization system and shows how information communication technologies can be used to enhance immunization coverage. The paper suggests a framework to capture the complex and dynamic nature of the immunization process, to enhance the understanding of the immunization health care problems and to generate insights that may increase the immunization coverage effectiveness.

1. Introduction

The role Information Communication Technologies (ICTs) play in improving the efficiencies and effectiveness of healthcare delivery has been well established in the more developed and industrialized parts of the world, however, the same is not true for developing countries in general. Traditional and new ICTs are being used to diffuse information to rural communities in developing countries (Gurstein, 2001). Developing countries lag behind in advances in information technologies over the internet, however, they are increasingly being used to increase the availability and quality of healthcare in remote areas, disseminate healthcare information to the public and provide knowledge to the healthcare professions (Musa, Meso and Mbarika, 2005).

The government of Uganda has designated ICT as a priority policy area and is committed to harnessing the ICT sector for national development (Scan-ICT Project, 2002). ICT role is generally low in the Ugandan healthcare environments, although most of the major hospitals and the medical schools use computers for administrative purposes, but only in limited ways. ICTs have greatly impacted the health sector and are increasingly being used to improve the administrative

efficiency of health systems. Service delivery in the health sector is still a challenge in many developing countries. Some of the issues that are faced by the health services include deficiencies in service delivery, growing gaps in facility and equipment upkeep, inequity of access to basic health services by the communities, inefficient allocation of scarce resources and lack of coordination among key stakeholders (Fraser and McGrath, 2000). The use of ICT technologies can increase the quality of health service delivery by providing reliable information and effective communication and efficient use of resources (Semwanga and Williams, 2006). The availability of information and communication techniques enables impoverished communities to access health care services which otherwise would be difficult under conventional healthcare systems.

2. Background

Preventable childhood diseases such as measles, polio and premature deaths still occur particularly in the developing countries due to low immunization coverage (WHO, 1999). In a study to evaluate new tendencies and strategies in international immunization, Martin and Marshall (2002) suggest that "*failure to immunize the world's children with life saving vaccines results in more than 3 million premature deaths annually*". According to Mbarika (2004), healthcare is one of the most fundamental needs for Sub-Saharan Africa. Various approaches have been applied to understand immunization coverage problems, however, there are still acknowledged deficiencies in these approaches and this has given rise to research efforts for alternative solutions including the need to adopt new technologies to address some of these problems. Primary healthcare has the role of monitoring of outbreaks and providing optimum continuous care for many diseases and is characterized by uncertainty, complexity, time delays and competitive stakeholder viewpoints. ICTs offer a platform for health education which plays a major role in the prevention of many diseases.

Definition of key Terms

To put this paper into context this section defines three key terms used in this research, namely immunization coverage, healthcare system, Information and technology Communications.

Immunization coverage can be defined as the proportion of a population that has been vaccinated against a given infection at a particular age (Edmunds *et al*.,2002). Infant immunization is done against childhood diseases such as: poliomyelitis, diphtheria, neonatal tetanus, measles, mumps, and congenital rubella. A target of at least 95% immunization coverage of children at 2 years of age is achieved in western countries, while about 60% is achieved in developing countries (WHO, 1999). Information is not always easy to obtain and some countries simply use immunization records or certificates held by schools or Health centers. If the number of children vaccinated in a district or country is not known, the number of doses administered, distributed or imported may be used to estimate the

number who received vaccine, however the problem is more difficult to solve for immunizations given in multiple doses.

The second key term used is ***healthcare system***, used to describe the organisation of human, physical and financial resources in the provision of health services to a given population. A healthcare system may include links between hospitals, home care agencies, long term care facilities and people (physicians, nurses, social workers, health care providers).

The third term, ***Information Communication Technologies*** is used to define tools that facilitate communication, processing and transmission of information and the sharing of knowledge by electronic means. This encompasses the full range of electronic digital and analog ICTs, from radio and television to telephones (fixed and mobile), computers, electronic-based media such as digital text and audio-video recording, and the Internet, but excludes the non-electronic technologies.

The rest of the paper is presented in these sections: Section three, presents compelling arguments why immunization system should adopt ICTs. Section four highlights some of the cases of applications of ICT in rural areas. Section five presents the research design pursued, while section six shows the key findings from data collection and other authors. Section seven presents the status of ICT, key challenges in healthcare, proposed ICT solution, benefits and challenges of ICT adoption and diffusion.

3. Why Immunization System should adopt ICT

Healthcare services in developing countries such as Uganda, in particular immunization services are provided through a decentralized system consisting of geographically spread health centres, regional hospitals which are categorized into health districts and health sub-districts with various roles as shown in figure 1 below:

Fig 1. The Ugandan Immunization System

As illustrated in figure 1, the development of health plans, policies and service delivery are channelled from the national level through the district, health sub-district, health centre right up to the community. The decentralized health care system makes the management and process of planning easier, however, this requires an effective feedback system, supervision, monitoring and reporting if the goals of the system are to be achieved.

Each health centre has its own immunization schedule and plans but it is desirable that the different health centres /hospitals offering immunization services work in a cooperative environment and be able to exchange data and information on service delivery. In order to improve the efficiency and effectiveness of immunization health services provided in such a distributed structure, it is vital that information is shared since it is a vital resource to the management of any organisation. Effective data collection and sharing of information can be enhanced through the application of information and communication technologies.

Health care services like any other business involve a lot of transactions such as importation and delivery of medicines; construction of hospitals and clinics; hiring and deploying staff; processing and payments of staff salaries. Other activities include the purchase and delivery of food for patients; collecting and analysing disease spread data in a country, and educating the community about good healthy living (Wasukira, Somerwel and Wendt, 2003). ICTs provide a competitive leverage to respond to the challenges of maintaining up-to-date immunization records (fully immunized and drop outs), tracking of facilities and vaccines, stock management, vehicle tracking and human resource management required by the immunization system.

4. Cases of ICT in Healthcare Delivery

The use of ICTs is rather limited in healthcare particularly in developing countries where healthcare systems are mainly used for storage and transportation of textual information using stand-alone computers. Some of the healthcare systems that have been developed include billing, financial systems, patient registration, computer based record systems and pharmacy systems. Most of lab equipment and radiology equipment are now computerized. Telemedicine which uses telecommunication and multimedia technologies is now increasingly used for remote consultation, diagnostics and examination of patients over the internet. As far as improving education in health is concerned, ICTs are being used for sharing documents, simulations, interactive environments and e-learning.

HealthNet one of the most widely implemented computer-based telecommunications systems in sub-Saharan Africa is being used in over 30 countries by around 10,000 healthcare workers to exchange ideas and provide medical solutions to various problems (Mbarika, 2004). HealthNet uses low earth orbit satellites and phonelines to provide email access system of local telecommunications sites used to provide low cost access to healthcare information in developing countries through a link to basic email (Kasozi and Nkuuhe, 2003). Users mainly physicians and medical workers connect to the network through local telephone nodes to access services such as physician collaborations (Mozambique, Tanzania, Uganda), data collection (Gambia), healthcare delivery (Ethiopia), research (Ghana), medical databases, consultation and referral scheduling, epidemic alerts and medical libraries.

Mozambique, a sub-Saharan Africa country launched its first TeleMedicine project in 1998. A TeleMedicine link connecting two central hospitals was built based on existing terrestrial and satellite communications system using low cost equipment for transmission, exchange and visualization of images and radiographs (ITU,1998). In Uganda hand-helds (EpiHandy) are being used by healthcare staff for communication (e-mail), studies and surveys, consultations and treatment guidelines (Kasozi and Nkuuhe, 2003).

SatelLife (Groves, 1996) uses low orbit communication satellites to link up doctors via the internet through "store and forward technology. SatelLife provides service to remote medical units through email and internet traffic as international telephone connections to capital cities in the developing world. Across Sub-Saharan Africa, the Internet is used to report daily cases of meningitis to monitor emerging epidemics. When threshold levels are reached, mass vaccination is required and the Internet is used to rapidly mobilize medical personnel and effectively coordinate laboratories and specialist services.

Nambaziira (2006) designed an online tool for ordering, distribution and monitoring of vaccines from the central stores to the various districts. Some of the functionalities included the capture and generation of reports for vaccine requisitions, supplies, issuances and disposals. Such a tool would be used in the monitoring of vaccines and that would reduce on the vaccine wastage thus minimizing the costs.

The above studies show that various technologies have been used to improve healthcare in remote areas although some of the challenges pertaining to healthcare are not adequately addressed. There is need to adopt systems that address some of these challenges and this can be done by undertaking a system such as provided by the systems dynamics methodology. In a study carried out to assess the application of information and communication technologies (ICT) in health information access and dissemination in Uganda, Omona and Ikoja (2006) suggest that there is need to support and promote ICT as the most effective tool for health information access and dissemination.

5. Research Methodology

In order to understand factors that influence immunization coverage and their relationships, survey research supported by semi-structured interviews were conducted to understand the intricate information flows, delays and other competitive challenges by use of ICTs. Data obtained from the study was analysed with SPSS statistical package. Influence diagrams representing the relationships between variables were developed using Vensim modelling software. Out of these key information and processes required for immunization coverage improvement were derived.

5.1 Research Questions

In order to assess how ICTs may be used to enhance the effectiveness and efficiency of the healthcare immunization system, several pertinent questions were considered:

1. What kind of data and information may be collected to understand the immunization coverage?
2. Which stakeholders provide data and use information for monitoring of the immunization system?
3. How should such an immunization system work?
4. What kind of ICTs may be used to deliver healthcare services?
5. How can these ICT tools and techniques be adopted to improve the management of the Ugandan immunization system?

It is these research questions that guided the research design. These experiences documented in the case studies and the cumulative knowledge were used to design a framework for enhancing immunization coverage.

5.2 Field Studies

Field studies were used to determine the full range of activities and challenges associated with immunization coverage and to examine the various acknowledged factors associated with the provision and utilization of immunization services were carried out. The study area Mukono District, lies in the Central region of Uganda comprises of four counties and has a good representation of both rural (83%) and urban population (17%) with a population density of 264 persons per sq. Km. Secondly, the people of Mukono district reside both on the islands (1 county) and the mainland (3 counties) and the population of Mukono consists of more than 18 tribes which would benefit the research by gathering cultural beliefs and opinions from the various tribes. The study was analytical; involving the various stakeholders who are important as far as the immunization system is concerned.

Mothers - In each county of the selected district, 200 mothers were interviewed. Multi-stage sampling method was used to select target sample size of 800 mothers. The sample size was determined as follows: Since many variables were being measured, a prevalence of 50 percent, which demands the largest sample size was used. At 95% confidence interval with the immunization coverage (p) of 70% and level of permissible error (e) as $e \leq 10\%$ the sample size n was determined by the following equation:

$$n = \frac{z^2 pq}{e^2} \qquad \text{.....Equation 1}$$

where $p=0.7$, $q=(1-0.7)=0.3$, $z = 1.96$ and $e= 0.1$

$$n = \frac{(1.96)2(0.7)(0.3)}{(0.1)2} = 80.67136 \qquad \text{.....Equation 2}$$

Considering a non-response rate of 20% results in 100 respondents. A design effect consideration resulted into 200 (100 x 2) respondents for each county thus making the number of respondents in the four counties equal to 800. In each county, the planned number of interviews was at least 200 mothers. The interviews of the mothers were carried out consecutively until the completed number of interviews which was 800. A structured questionnaire was used to interview the mothers.

Health Workers: Three (03) private and five government (05) health facilities selected by simple random sampling from the district. Those that were selected included one (01) government hospital and one (01) private hospital and the rest were health centers and dispensaries. At each sampled health unit, two people were interviewed, one vaccinator and one Officer-in-Charge of vaccines bring the total of those interviewed to sixteen (16).

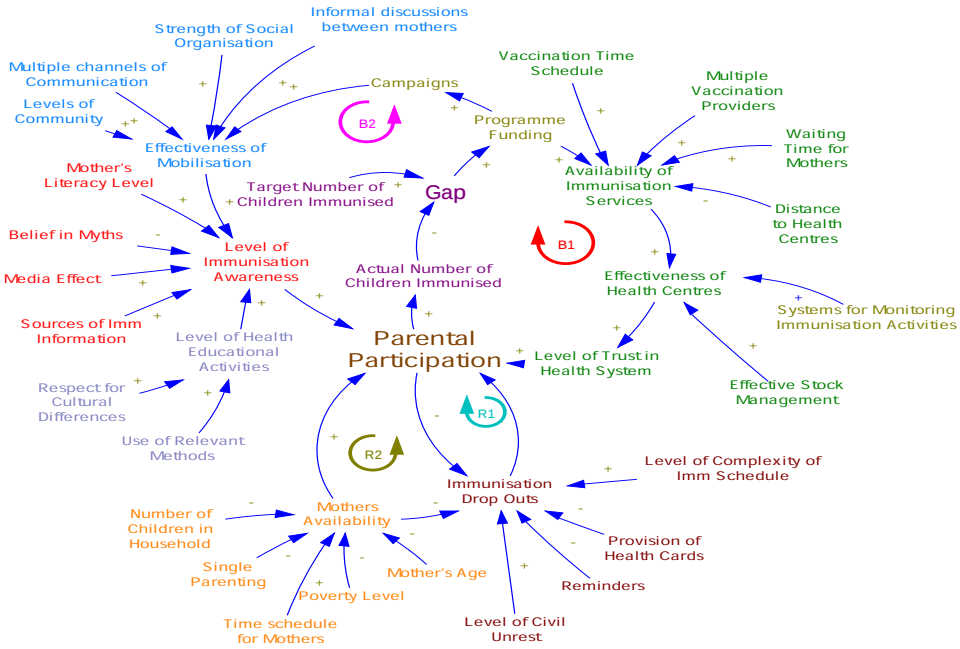
Officials: At the district level, several meetings with various officials from health services, administrative officials were held. Local community leaders, national officials as well as consultants with UNICEF were interviewed.

6. Factors Associated with immunization coverage

Findings from the field study as well as immunization studies of other researchers (Borooah, 2003; DISH, 2002; Drain *et al*., 2003; WHO, 1999) are represented in the influence diagram (refer to Figure 2). Figure 2, illustrates the intricate and complex relationships among factors affecting immunization coverage from a parental participation perspective and a number of feedback loops which may help to explain different immunization coverage levels (Rwashana and Williams, 2006). It is this feedback structure that gives rise to complexity, nonlinearity, and time delays in Immunization coverage. Figure 2 shows that immunization coverage can be enhanced through improvement of the following :-

- Immunization awareness and knowledge which can be improved through health education, effective mobilisation and campaigns, literacy levels, increased sources of information and use of relevant methods and content,
- Level of trust in the health system which is built by having effective health centres with effective vaccines, sufficient well trained health workers (less waiting time for mothers), health education services.
- Availability of immunization centres which are easily accessible to the community. Systems for reporting and monitoring immunization activities
- Tracking number of children being immunized and the drop outs
- Vaccine inventory control and monitoring.
- Reporting and monitoring of immunization activities.

Figure 2 : Influence Diagram for parental participation towards immunization



7. Status of ICT in Health

According to SCAN-ICT Project (2002), most governments health units have generally low ICT usage due to lack of basic ICT infrastructure. 80% of the Ugandan population is rural based with no electricity distribution thus constraining the diffusion of all forms of ICT. Internet connectivity and email usage in the health sector is still low compared to other sectors. Medical personnel most commonly use computers like accountants and secretaries. Until 1993, Uganda had a centralized health information system (HIS) which focussed on morbidity and mortality reporting, with data flowing only from individual health units to the district and national level. A health management information system (HMIS) that emphasizes use of information at the point of collection is currently in use (Gladwin, Dixon and Wilson, 2003).

Immunization records, at the health unit level are done on paper registers and tally sheets by the health workers. The tally sheets are forwarded to the health district level for entry into a computerized database Health Management Information System (HMIS). Districts that have a computer and resources to maintain it, are provided with an easy spreadsheet based system for compiling monthly and annual reports. The reports from the districts are delivered to the headquarters either by hand, fax or by email. The national office has a LAN at the headquarters to enable health offices gain access to the HMIS products such as the ministry of health website. Future expansion includes development of a WAN, to provide connectivity to the rest of the districts.

7.1 Challenges with Immunization Healthcare Delivery

Some of the issues associated with the current system used for recording and retrieving information include the following :-

1. Recording and capturing the data from the tally sheets into the computer is done at the district level which is very tedious, voluminous and often results into errors arising from poor handwriting thus causing a lot of delays in compilation.
2. It is difficult to trace the children who have not completed the immunization schedule status of each child since records are done using a one-ff tally sheets.
3. Long time-delays for mothers resulting in recording of manual systems.
4. Lack of up-to-date reporting leads in inaccurate forecasts and targets, occasional stock outs and vaccine wastage at the various immunization centres which results into missed opportunities by the children.
5. Vaccines are estimated based on the district population and this results in wastages and stock outs since some of the births are not registered at the district.
6. Staff capturing the data lack competence in using the data capturing tools which results in slow processing and a lot of errors.
7. Poor management and lack of storage space of the paper files.
8. Insufficient monitoring and supervision of health centres.

7.2 Proposed ICT Solution

The ICT infrastructure enables the sharing the ICT capabilities which provide services for other systems of the organization (Broadbenta *et al.*, 1999). For Broadbenta *et al.*, (1999), these capabilities require the complex combination of the technical infrastructure (cabling infrastructure, hardware platform, base software platform), ICT shared services (as communications services), ICT applications (as WEB services), the human operators and the managerial expertise to guarantee reliable services (see figure 1).

The paper suggests the following ICT innovations (Figure 3) towards the improvement of immunization coverage :-

Child: In the proposed system, a SMART card will be prepared for each child born. The card will have a chip card that can be swapped for record updates and the following details:

- Photograph of the child
- Child's name, year of birth and sex
- Father's name, address, telephone number
- Mother's name, address, telephone number
- Location : Local council, village, parish, county, district
- Dates of immunization of the various vaccines

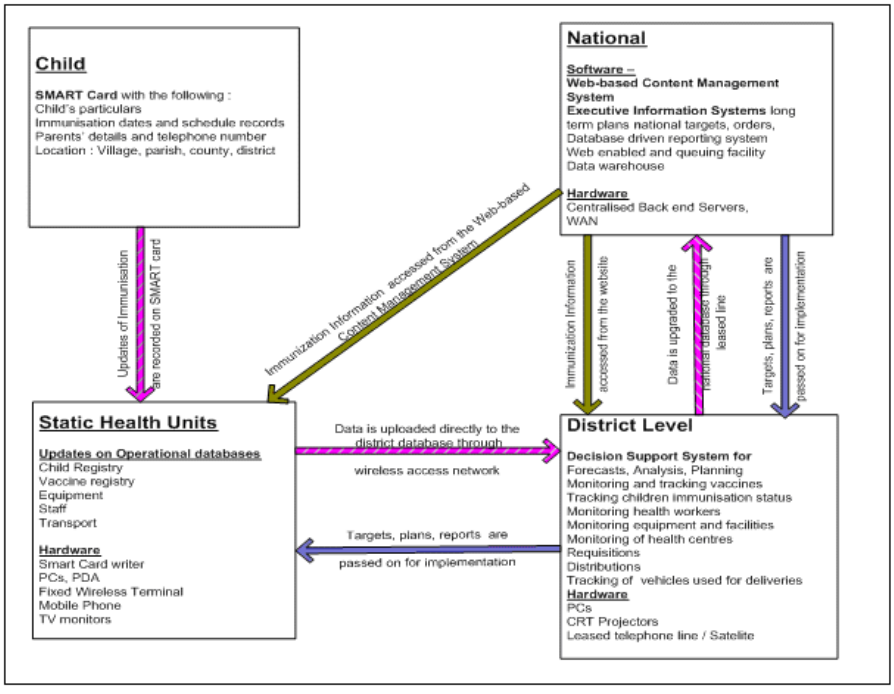
Health Units: The health units will have a SMART card writer which can update the immunization records. Stock and logistics updates and orders will be captured by using the handheld devices / Personal Digital Assistant (PDA). The

data is then uploaded onto the handheld device and sent to the district database through a fixed wireless terminal. Orders and updates of vaccines from the various health units is compiled to facilitate decision making processes at the district level. Reports on the performance of the various health units are generated and the data is then uploaded to the national database. Health workers will use the PDA to access immunization information (diseases, vaccines, side effects, immunization schedule, stock management) from the immunization content management system. PDA. Such information on a PDA when connected to the right video/ audio output drivers could be shown using a TV monitor thus facilitating training and health education in the community

District level: At the district level, data can be used to prepare forecasts, analysis and plans for the district. The reports generated can be used for monitoring and tracking of vaccines, children, health workers as well as preparing requisitions and facilitate distributions of vaccines to the various health sub districts and health units. The data from all the districts is compiled to obtain the national immunization performance. The data can then be used to generate national targets, imports of vaccines and supplies, management of staff and equipment as well as enhance decision making process for long term plans.

National level: An immunisation content management system that has all the information pertaining to immunisation (diseases, vaccines, schedules, side effects) and can be accessed by all districts and health units should be developed. A Health Management Information System linking all the health facilities at all community, district levels should be used for support and monitoring of immunization activities, planning. decision making, education and disease surveillance.

Fig 3: Proposed ICT Framework for the Immunization Services



7.3 Benefits Provided by ICT Adoption

Some of the benefits that would be provided by adoption of ICT services in the healthcare immunization system are :

1. Improved quality by reduction of variations in healthcare practices can significantly improve policy design effectiveness,
2. Improved sharing and dissemination of reliable information concerning vaccine administration, storage, side effects through online accessibility to the medical workers as well as the public improving their attitude towards immunization.
3. Timely reporting of disease outbreaks/epidemics will be made easier thus enabling timely interventions by the authorities.
4. Reduction in time delays for mothers due to reduced manual recording processes
5. Improved coordination and monitoring of immunization activities, facilities and equipment upkeep and allocation of scarce resources.
6. Improved overall administrative effectiveness of the immunization system which results in efficient use of resources both human and financial.

7.4 Challenges to adoption and diffusion of ICTs

Some of the challenges to adoption and diffusion of information technologies in remote areas for healthcare delivery include the following:-

a. Lack of telecommunication infrastructure and underdeveloped state of Internet Service Providers (ISP) which results in low internet penetration and insufficient bandwidth.
b. Lack of ICT awareness
c. Lack of relevant content which is essential to make ICT effective for the community thus presenting the need to repackage and develop local content.
d. Lack of skilled ICT personnel who are willing to live in remote areas, so there is need for capacity building and training
e. Cultural barriers include reluctance to adopt new technologies, language barriers and fear of SMART card technology which presents issues of confidentiality and privacy.
f. Political barriers such as unstable economic climate in the country.
g. Remote areas experience intermittent power supplies thus requiring other forms of power supply
h. High costs of installation and maintenance of ICTs
i. There is need to further enhance the enabling regulatory environment.

Conclusion and Future Studies

Simple, low cost techniques that are sustainable should be developed based on the following strategies for overcoming barriers to the successful integration of ICT into the delivery of immunization healthcare systems. First, the development use of standardised systems. Second, the government needs to provide political leadership to accelerate the adoption of electronic health systems and, third to create a public database that holds data at the community level, but is fed through the decision making structure to improve healthcare provision nationally and local level.

References

Barenzi, J. Makumbi, I.and Seruyange, I. (2000). *Immunization practice in Uganda.* A manual for operational level health workers. UNEPI/TRA.

Borooah, V.K., (2003). Gender bias among children in India in their diet and immunization against disease. *Social Science and Medicine Journal*, **58** :1719-1731.

Broadbent, M.,Weill,P.andSt.Clair,D.(1999),"The implications of information technology infrastructure for business process redesign, *MIS Quarterly*, **23**(2) :159-82.

Chanopas, A., Krairit, D. and Khang, D.B. (2006). Managing information technology infrastructure : a new flexibility framework. *Management Research News* **29**(10) : 632-651.

DISH II Project (2002). Childhood Immunization in Uganda : A Report of Qualitative Research, K2-Research Uganda Ltd.

Drain P.K., Ralaivao J.S., Rakotonandrasana A., Carnell M.A., (2003). Introducing auto-disable syringes to the national immunization programme in Madagascar. *Bulletin of WHO*, **81**(8).

Edmunds, W.J., Brisson, M., Melegaro, A. and Gay, N.J.(2002). The potential costeffectiveness of acellular pertussis booster vaccination in England and Wales. *Vaccine,* **20**:1316-30.

EPI (2000). *EPI plan of action (2000-2002).* Ministry of Health Uganda government.

Fraser, H.S.F. and McGrath, J.D. (2000). Information technology and telemedicine in sub-Saharan Africa. *British Medical Journal* **321**:465-466.

Gladwin, J., Dixon, R.A. and Wilson, T.D. (2003). Implementing a new health management information system. *Health Policy and Planning*. **18**(2).

Groves, T. (1996) SatelLife : getting relevant information to the developing world. *British Medical Journa*l **313**:1606-1609.

Gurstein, M. (2001). Community Informatics, Community networks and strategies for flexible networking In L. Keeble and B. Loader (eds). *Community Informatics : Shaping computer mediated social relations*, Routledge, London

Kasozi, M. and Nkuuhe, J. (2003). Uganda Chartered Healthnet promotes healthcare using Handheld devices. *I-network Uganda*. **2**:4

Martin, J.F. and Marshall, J. (2002). New tendencies and strategies in international immunization: GAVI and The Vaccine Fund. *Vaccine* **21**:587-592.

Mbarika, V.W.A. (2004). Is Telemedicine the Panacea for Sub-Saharan Africa's Medical Nightmare ? *Communication for the ACM,* **47** (7): 21-24.

Musa,P.F., Meso,P., and Mbarika,V.W. (2005). Toward sustainable adoption of technologies for human development in Sub-Saharan Africa: Precursors, diagnostics and prescriptions. *Communications of Associations for Information Systems*. 15 (33).

Nambaziira, S. (2006). An online tool for Monitoring and tracking vaccines and vaccine logistics utilization at district level in Uganda. Masters Thesis.

Omona and Ikoja-Odongo (2006). Application of information and communication technology (ICT) in health... *Journal of Librarianship and Information Science*; 38:45-55.

Rwashana and Williams (2006). An Evaluation of healthcare policy in immunization coverage in Uganda. *In the Proceedings of the 24th International System Dynamics Conference 23-27 July, 2006*, System Dynamics Society. ISBN 978-0-974-5329-5-0.

SATELLIFE PDA Project (2002).Testing the use of handheld computers for heathcare in Ghana, Uganda, and Kenya. Report published: 28 February 2003.

Unwim, T. (2004). Towards a framework for the use of ICT in Teacher Training in Africa. *Journal of Open and Distance Education in Open Learning in Less developed countries*.

Wasukira, E., Somerwel, F. and Wendt, D. (2003). ICT for development : ICT in Health Seminar. *I-network Uganda*. **2**(4) .

World Health Organization (1999). Measles. Progress towards global control and regional elimination. *Weekly Epidemiology*, 74(50): 429-434.

13
Geometrical Spatial Data Integration in Geo-Information Management

Ismail Wadembere, Patrick J. Ogao

Geospatial information systems provide users with both absolute and relative locations of objects and their associated attributes. Partly, and to achieve this, the users need to use different geospatial data sets from various sources. The problem is that, there will inevitable exists some geometrical mismatch between these datasets and thus making it difficult to fit the data sets exactly with each other. This paper develops a conceptual mechanism for geospatial data integration that can identify, compare, determine differences and adjust spatial geometries of objects for information management.

1 Introduction

One of the reasons that Individual users and organizations use GIS, is to exchange geospatial data as a means to location based decision-making. A considerable amount of preprocessing has to be done, before and after the data has been integrated.

Within this decade, we are seeing developments in technologies which are creating services which need geospatial data comparison and integration (Najar et al., 2006). This is so evident in: Google Earth; Microsoft's MapPoint.Net; O'Reilly's Where 2.0; Intergraph's reorganization around "SIM"; Oracle Locator and Spatial; ESRI's ArcGIS 9.x; US Census' MAF/TIGER integration; new platforms, new vendors, new users and the many conferences on mobile commerce and location-based services (Batty, 2004, Alperin, 2005) and varying application (Busgeeth and Rivett, 2004). These developments and changes are so diverse that they don't even seem related, but they are (Sonnen, 2005); that is, they take advantage of the vast geo-information available, which reflect the increased importance of location in information systems and services (Strader et al., 2004).

Increasingly, we are seeing technologies like the Internet, web services, image capture and communication, low cost spatial data collection tools, grid computing power, On-Line Transaction Processing (OLTP) application (Skog, 1998), (Batty, 2004), (Sharma, 2005), (Sohir, 2005) among others utilizing geo-spatial data. These technologies are positively influencing the need for spatial data integration.

The need for geospatial integration is further fuelled by Location Based Services and Location-Tracking Trends; for example in the United States, where they have moved to location-aware mobile phones which is being driven by the federal government's enhanced 911(E911) mandate (FCC, 2006). Also, researchers are experimenting with ideas such as leaving messages for other people attached

to certain places, all enabled by on-line location services (Sharpe, 2002). All these pinpoint to the emergent need for geospatial data integration to improve on sharing and information management.

Further needs can be seen in Enterprise Location Services and Software (ELSS); which is about the use of location-specific information within information systems. To really appreciate this, we need to understand that several technologies combine to make ELSS grow and these include the Service-Oriented Architectures (SOA), and the automated location determination technology like GPS, cellular networks and radio frequency (Sonnen, 2005). Also, spatially-enabled IT Infrastructure, Open Standards, Spatial Data infrastructures (SDI), etc are becoming ubiquitous which underscore the need for a location ready SOA platform (Sharma, 2005, Sohir, 2005). All these have provided avenues for sharing geo-information from different sources, and which also need to be integrated.

We should not forget that for many organizations, integration is still a major challenge in GIS projects. According to Aberdeen Group as quoted by (Batty, 2004), on average 40 percent of IT budgets is spent on integration, and as much as 70 percent in some cases.

As we move towards new approaches to data integration, there is need to relate it to information management. As we do that, we have always to remember that Information systems are always new and changing (Claver et al., 2000) and the issue of management is needed as from the very foundation of computer science in 1945 (Floridi and Sanders, 2002) it was information which needed to be managed in order to achieve goals as information has been considered as a 'utility' just like electric power (Malhotra, 2000).

The integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. The reason for integration is twofold: First, given a set of existing information systems, an integrated view can be created to facilitate information access and reuse through a single information access point. Second, given a certain information need, data from different complementing information systems is to be combined to gain a more comprehensive basis to satisfy the need.

This paper develops a conceptual Geometrical Spatial Integration Model (GSIM), which can identify, compare, determine differences and adjust spatial geometries of objects so that they fit exactly on each other for information management. It examines data integration approaches and limitations of existing geo-data integration approaches, the philosophical underpinning for geospatial data integration and developing GSIM, triangulation of methods for GSIM, specification and design of GSIM, future work and Conclusion.

2 Data Integration Approaches

The integration problem can be addressed on each of the presented system layers using the following general approaches (Ziegler and Dittrich, 2004, Jones, 2005, Bernard, et al, 2003, Foster, et al, 2003).

A. *Manual Integration*: Here, users directly interact with all relevant information systems and manually integrate selected data. The users have to deal with different user interfaces and query languages that requires detailed knowledge on location, logical data representation, and data semantics.

B. *Peer-to-peer* (P2P) integration is a decentralized approach to integration between distributed, autonomous peers where data can be mutually shared and integrated. P2P integration constitutes, depending on the provided integration functionality, either a uniform data access approach or a data access interface for subsequent manual or application-based integration.

C. *Common User Interface*: is were the user is supplied with a common user interface (e.g. web browser) that provides a uniform look and feel. Data from relevant information systems is still separately presented so that homogenization and integration of data has to be done by the users e.g. search engines

D. *Integration by Applications*: This approach uses integration applications that access various data sources and return integrated results to the user. This solution is practical for a small number of component systems as applications become increasingly fat as the number of system interfaces and data formats to homogenize and integrate grows. Examples include workflow management systems that allow to implement business processes where each single step is executed by a different application or user.

E. *Integration by Middleware*: Middleware provides reusable functionality that is generally used to solve dedicated aspects of the integration problem, e.g. SQL-middleware. While applications are relieved from implementing common integration functionality, integration efforts are still needed in applications. Additionally, different middleware tools usually have to be combined to build integrated systems.

F. *Uniform Data Access*: In this case, a logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available on this level. However, global provision of physically integrated data can be time-consuming since data access, homogenization, and integration have to be done at runtime. Examples include mediated query systems, portals for Internet/intranet, federated database systems, etc.

G. *Common Data Storage*: Here, physical data integration is performed by transferring data to a new data storage; local sources can either be retired or remain operational. In general, physical data integration provides fast data access.

In Summary, we see integration efforts under (i) using distributed models that are based on open systems technologies like OpenGIS based on extensible

markup language (XML) (Nebert, 2003) and different geo-information sharing infrastructures like Spatial Data infrastructures (SDI) (Sohir, 2005), (Musinguzi et al., 2004), Geospatial Data Clearinghouse, Proprietary protocols, GIS-network integration (Evans, 1997), (Nebert, 2003), etc. (ii) How it can be done between organizations/individuals (Evans, 1997), (Erdi and Sava, 2005). (iii) Then ways of storing geo-data with varying standards in same geodatabase using multiple representation (Kilpelainen, 1997), (Friis-Christensen et al., 2005) and attribute versioning (Bell, 2005).

However, these do not meet all the requirements, as some times, there is need for actual integration i.e. combining data with varying geometries. This requires to adjust the data sets so that they fit exactly on each other and can be used in analysis and modeling. Thus, in this paper we investigate conceptual understanding for Geometrical Spatial Data Integration Model (GSIM), which can identify, compare and determine differences between the geometries of geo-spatial datasets, then adjusting them so that they can be integrated before they are adapted/adopted in sharing avenues and used in spatial analysis and modeling.

3 Philosophical Underpinning for Geospatial Data Integration & GSIM

Philosophy is about observing, posing questions, investigating, analyzing, developing and evaluating human ideas (Wikipedia, 2006) with the overall purpose of coming out with techniques, theories and methods to employ and apply in the study of certain subject. When it comes to GSIM, philosophy is seen as a guide in proposing, designing, implementing and evaluating a model (Floridi, 2004).

From literature, (Floridi, 2004) related philosophy to cooking, as it is a matter not of attempting all at once but of careful and gradual preparation. Even the most astonishing results are always a matter of thoughtful choice and precise doses of the conceptual ingredients involved, of gradual, orderly and timely preparation and exact mixture. He argues that if we do not succeed in solving a problem, the reason may have arisen from our failure to recognize its complexity. It is also important to note that negative solutions are as satisfactory and useful as positive solutions. They help to clear the ground of pointless debates. This is further supported by (Peikoff, 2000), who relates philosophy to conceptual enquiry - critical reasoning, including analyzing the meaning of concepts, identifying logical connections between theories, evaluating arguments, and exposing fallacies. This is very vital in GSIM conceptualization so that it meets the needs of the users and we become certain that it will follow the paradigm, epistemology and ontology within information systems.

In (Hirschheim and Klein, 1989) Burrell and Morgan define paradigms as "meta-theoretical assumptions about the nature of the subject of study." Although, this differs from Kuhn's classic conception of paradigms which were defined as "universally recognized scientific achievements that for a time, provide model problems and solutions to a community of practitioners" (Bancroft, 2004); (Chen, 2002); (Hoyningen-Huene, 1993), (Kuhn, 1962). In GSIM, we use paradigm as the most fundamental set of assumptions adopted by a professional community that

allows its members to share similar perceptions and engage in commonly shared practices. Thus, it is an intellectual tool for problem designing and solving to shape society; hence it may have normative implications, as all design implies a solution of how things are to be instead of their present state. Each paradigm may comprise of various theories, models, methodologies, methods, and techniques, all designed to support the solving of problems, whether of particular or general type.

A paradigm consists of assumptions about knowledge (epistemology) and how to acquire it, and about the physical and social world (ontology) - such assumptions are shared by all scientific and professional communities. As developers, we must conduct inquiry as part of systems design and have to intervene into the social world as part of systems implementation, it is natural to distinguish between two types of related assumptions: those associated with the way in which system developers acquire knowledge needed to design the system (epistemological assumptions), and those that relate to their view of the social and technical world (ontological assumptions) (Hirschheim and Klein, 1989).

Two types of assumptions about knowledge (epistemological) and the world (ontological) are given by Burrell and Morgan to yield two dimensions: a subjectivist-objectivist dimension and an order-conflict dimension. In the former, the essence of the objectivist position "is to apply models and methods derived from the natural sciences to the study of human affairs. The objectivist treats the social world as if it were the natural world. In contrast, the subjectivist position denies the appropriateness of natural science methods for studying the social world and seeks to understand the basis of human life by delving into the depths of subjective experience of individuals (Hirschheim and Klein, 1989). In the order-conflict dimension, the order or integrationist view emphasizes a social world characterized by order, stability, integration, consensus, and functional coordination. The conflict view stresses change, conflict, disintegration, and coercion.

The dimensions when mapped onto one another yield four paradigms (Hirschheim and Klein, 1989) functionalism (objective-order); social relativism (subjective-order); radical structuralism (objective-conflict); and neohumanism (subjective-conflict). The functionalist paradigm is concerned with providing explanations of the status quo, social order, social integration, consensus, need satisfaction, and rational choice. It seeks to explain how the individual elements of a social system interact to form an integrated whole (Hirschheim and Klein, 1989)

As we deal with different approaches, we should remember that science, by nature, is tentative meaning that change is inevitable. Kuhn describes change as a scientific revolution that results in "the tradition-shattering complements to the tradition-bound activity of normal science" (Bancroft, 2004). His ideas that the scientific community is tethered to outdated notions that prevent them from moving forward to serve humankind are incredibly intuitive. It makes sense to view science as an ever-changing entity that must be approached with new perspective in order to avoid a stagnation of ideas.

Over the past two decades, there has been a growing number of Information System (IS) publications that advocate the use of alternative approaches to understand the organizational, behavioral and social consequences of IS planning, development, adoption and use (Lau, 1999). This has shown maturity of methodological pluralism in IS research through its diverse approaches that range from critical social theory, case study research, grounded theory, ethnography, action research to multi-method triangulation (Lau, 1999). Such inquiry on the use of multi-methods is explored further in this paper with emphasis to understand philosophical and theoretical underpinning for the development of GSIM.

4 Triangulation of Methods for GSIM

As Information System (IS) continues to become an established field, academicians have taken different approaches to do research in IS. There are many reasons for that, but the most profound include: -

a. In order, to have IS as an independent field, researchers (King and Lyytinen, 2005) have been faced with the question of which way to go and as result, they have taken different approaches.
b. By practitioners and business community wondering if IS research is relevant have also forced IS researchers to take different approaches (Benbasat and Zmud, 1999).
c. In order to achieve a relation between IS and other fields has forced IS researchers to use different approaches.
d. Different approaches have been used by researchers as it gives them the opportunity to be creative.
e. The question of whether existing research approaches given a basis to start new research has forced IS researchers to design new approaches.
f. For new researchers in IS with no or little knowledge about IS research approaches and paradigms have ended up creating new ones.
g. The editors of leading journals have also led to new and different approaches as IS researchers struggle to write their work according to journal specifications.
h. The rapid change in IT have led IS researchers to adopt or create news approaches which can cop with IT changes.
i. The changing needs of users like the use of location in many IS which leads to triangulation of methods to develop GSIM.

About the use of system approach: a system may be decomposed into components or subsystems that may in turn be systems that may be decomposed etc (Wangler and Backlund, 2005). Furthermore, any system may be a part of a greater system. Hence, it is a central task in systems development to come to grips with:- 1) the best way to decompose a system into components and subsystems, 2) how the components interact with each other, and 3) how the system relates

to and interacts with the superior system of which it is a part. In consequence, employing a partially systemic, partially reductionist view while working with information systems development, entails:- a) that you decompose the system into parts, that may be, as much as possible, independently developed (i.e. components with high cohesion and low coupling), and b) that you clearly identify the way the system interacts with the business processes and functions that it is meant to support.

As we design GSIM, the basic characteristics of IS which is handled should be true, up-to-date, standard, flexible, unrepeated, in desired form and sufficient to the needs to the user and their share ability (Erdi and Sava, 2005) within the IS, should be taken care of.

We should always remember that both diverse academic disciplines and diverse research communities contribute to IS research (Niehaves, 2005). It is against this background of diversity that we develop GSIM using multi-methods so that it can benefit from such a variety of approaches and contributions. In the process, it will help us to see IS-related phenomena from different points of view (Mingers, 2001), (Weber, 2004), thus GSIM benefiting at end. Among the issues that we put into consideration is to develop GSIM using different approaches basing on independent objects, which can function both as stand alone or integrated with others to accomplish a task – different objects can be used in combination or independently.

This has been done so that GSIM is not tied to one format (Bernard et al., 2003) and in case there are changes, then the specific object concerned is the one modified and individual objects are employed to accomplish particular task of GSIM. This is so as this research deals with integration of spatial vector data using the geometry primitives as the basis for integration. Emphasis is placed on developing GSIM that can be implemented using messaging, UML for object development and XML (especially GML) manipulation for identifying data components and trigging different objects.

The idea being to utilize the following methods: - spatial validation, spatial matching, spatial interpretation, spatial reverse engineering (Aronson, 1998) and GIS spatial analysis functions like overlay, merge, connectivity, neighborhood, operations.

5 Specifications and Design of GSIM

Assuming, a user has two geo-spatial data sets from different sources and s/he is interested in using the two data sets, thus s/he has to integrate the two data sets.

With that problem at hand and as we design a GSIM, there is need to understand that:-

a) geo-spatial data may contain bias like some features being over emphasized and some under represented

b) different data sets representing the same entities may cover different area extent

c) data sets may just be incomplete i.e. do not have all the required elements for certain analysis and modeling to be meaningful
d) sources contain closely related and overlapping data
e) data is stored in multiple data models and schemas
f) topology requirements like overlap, containment, connections and adjacency
g) data sources have differing query processing capabilities

Questions we asked ourselves in order to accomplish the problem and these form the basis for design of GSIM are:- a) Are two spatial data sets different? b) Is the different due to geometry? c) Which of the data sets is correct or is better? d) What parameters of geometry are different? e) Can the geometrical parameter be changed? f) How much is geometrical difference? g) How to adjust the geometrical parameters? h) How to build the spatial geometry? i) How to combine the data sets?

The sub-model (objects) which form GSIM are based on above questions: spatial scaling, spatial data matching, best spatial data set, spatial components identification, geometrical spatial difference, geometrical spatial adjustment, geometrical spatial builder, topology creation, and model development.

Spatial Geometrical Comparison

Spatial Scaling

Two data sets are checked to see if they occupy the same aerial extent, if not the data sets are adjusted so that they occupy the same area. This may involve change of scale arising due to data shrinkage or enlargement of a particular data set. If data sets representing the same area have different scales, then scale change has to be done so that the data sets cover the same area extent. The intension here is not to handle whole scale of the data within the object although scale change can take place so that the dataset cover the same area extent. The scale change, which is handled by the GSIM, affects individual geometries of the spatial elements for example changing shape of polygon say for a plot of land.

Spatial Data Matching

The main purpose for this object is to determine if two spatial data sets are different and if the difference is due to geometry. Here, we also determine which geometrical parameters are causing the difference. This involves spatial object matching to predict errors that are as result of not exact matching (due to spatial validation) and end up producing unwanted components for example silver lines/ polygons.

We employ the ideas of spatial data matching and spatial interpretation in understanding the spatial data sets, similar geographies and comparing spatial data sets at object level. The end product at this level is algorithm that can determine spatial geometry differences and identify which components of the geometry are

causing the mismatch. The main considerations are scale, resolution, and actual errors in one data set.

Best Spatial Data Set

This object compares the data sets with known data sets (especially base data set) or known controls with the purpose of finding correct data sets or assessing which data sets should be used as the basis for integration. If none then which one appears to be more correct and should be used as the basis for spatial adjustment. The end result in this stage is algorithm which compares data sets against set parameters and comes out with the most probably data set

Spatial Geometrical Adjustment

Spatial Component Identification

Since different components of geo-spatial data sets have to be handled differently during spatial adjustment; this object breaks down the different primary component of the data set into individual layers i.e. point layer, ploylines layer, and polygon layer.

Geometrical Spatial Difference

This is basically used for geometrical quantitative difference determination. It involves the process of extracting the geo-data's spatial geometry and determining the geometry relationships of components depending on the particular geo-data source.

Spatial validation is utilized to determine elements and is based on logical cartographic consistency such as overlap, containment, adjacency & connectivity, and other topological requirements like polygons have to be closed, one label for each polygon, no duplicate arcs, no overshoot arcs (dangles), etc. GIS spatial analysis functions like overlay, merge, connectivity, neighborhood, operations are also utilized to come up with an algorithm for geometrical quantitative spatial difference determination; which should be able to handle:-

- For line features, the whole line feature is picked at a time and comparison takes place in the following parameters: - length of the ploylines, location (using coordinates) of the nodes and vertices, overlap, no nodes, thickness, and continuity (connectivity).
- For the point features, an imaginary fine grid is over laid on the two data sets so that the values in each grid are compared. The grid size is according to the scale of the data set i.e. the grid square is determined by the scale at which the comparison is being done. The difference is determined and issues that are compared include precision, resolution, and actual data values. If the difference is zero, we assume that the two cells are the same and if the resultant is not zero (could be positive or negative), then the two are different.

- For polygons, whole polygons are compared, the shape, corner/edge coordinates, overlaps, label points, adjacency, containment, over size, under size, neighborhood, not closed, and no label.

Geometrical Alignment

After identifying the components causing the differences and determining how much of the difference exist; this object breaks down geometries of data to individual primitives. Then, examines the geometry elements and carries out adjustments according to the obtained quantity of the difference.

Spatial validation and Spatial Reverse Engineering (SRE) are employed here to help in disintegrating geo-data from any source into primitive objects/elements. This is done by putting into consideration the strength and goals of SRE (i) to analyze a data set, (ii) to identify the data's components, (iii) to identify the interrelationships of the data components, and (iv) to create representations of the data in another form or at a higher level of abstraction. SRE helps to return to a less complex or more primitive state or stage of the spatial data geometry. The resultant algorithm has to accomplish *ploylines, points and polygons correction* according to the geometrical spatial difference.

Geometrical Spatial Builder

This object looks at the different spatial primitives which have been adjusted and builds them into components which have spatial meaning and these are later combined during topology building.

Geometrical Spatial Integration Model

Topology Creation

As GIS data should have proper topology to enable users carry out modeling and analysis, this unit/object combines the spatial objects generated using geometrical spatial builder into proper polygon, ploylines and points and creates relationship between them to obtain meaningful spatial data set.

Combining components

This is the actual combining of data sets together after the above units/objects have played their roles. After developing the different units, the resultant - Geometrical Spatial Integration Model (GSIM) is achieved using an iterative process where the design rules for each possible sequence of the unit/object is taken into consideration and deployed according to given user needs/requirements. The resultant GSIM and also the different units/objects can be deployed using approaches like novel service retrieval (Klein and Bernstein, 2004) that captures service semantics using process models and applies a pattern-matching algorithm to find the services with the behavior the user wants.

Model Validation and Evaluation

To make sure the model can be utilized; GSIM is evaluated with help of topological requirements testing. This is done theoretically and experimentally using the following criteria: the maturity and capabilities of each object, whether or how well each object can be used in conjunction with others, and how the whole model can fill full geometrical integration requirements.

Begin with describing the goals of GSIM, then describe the functionality of the GSIM. Observe the interaction between the different objects. A goal-based evaluation is performed to decide if goals are met. Then spatial validation is employed to determine variables and is based on a) logical cartographic consistency such as overlap, containment, adjacency & connectivity, b) closed polygons, c) one label for each polygon, d) no duplicate arcs, e) no overshoot arcs (dangles). GIS spatial analysis functions like overlay, merge, connectivity, neighborhood, and operations are also utilized.

6 Future Work

The designed GSIM is going to implemented and tested so that is meets the need of the expanding geo-technologies and before it can be adapted in sharing avenues like SDI and used in spatial analysis and modeling.

7 Conclusion

The main function of geospatial information systems is to provide location specific information which is easy to understand and can be used in decision-making. In this paper, we have designed a conceptual Geometrical Spatial Integration Model that can be used to identify, compare, determine difference and adjust geometries of spatial data sets.

References

Alperin, J. (2005) Spatial Information Technology, Globalisation and International Development. http://www.fes.uwaterloo.ca/crs/plan654/.

Aronson, D. (1998) Systems thinking. *R \&D Innovator. (now Innovative Leader),* 6.

Bancroft, A. (2004) The Structure of Scientific Revolutions by Thomas S. Kuhn., Position Paper, ETEC 660/695.

Batty, P. (2004) Future Trends & the Spatial Industry. Geospatial Solutions http://www.geospatial-online.com/geospatialsolutions/article/articleDetail.jsp?id=101548.

Bell, C. A. (2005) Attribute-Level Versioning: A Relational Mechanism for Version Storage and Retrieval. *School of Engineering.* Richmond, Virginia (USA), Virginia Commonwealth University.

Bernard, L., Einspanier, U., Lutz, M. & Portele, C. (2003) Interoperability in GI Service Chains-The Way Forward. *6th AGILE Conference on Geographic Information Science*, 179-188.

Busgeeth, K. & Rivett, U. (2004) The use of a spatial information system in the management of HIV/AIDS in South Africa. *International Journal of Health Geographics* 3**,** 13.

Chen, D. (2002) An Empirical Test of Thomas Kuhn's Structure of Scientific Revolutions: How Does Science Progress?

Claver, E., Reyes, G. & Juan, L. (2000) Analysis of Research in Information Systems (1981-1997). *Information and Management,* 37**,** 181-195.

Erdi, A. & Sava, D. S. (2005) Institutional Policies on Geographical Information System (GIS) Studies in Turkey. *Pharaohs to Geoinformatics, FIGWorkingWeek 2005 and GSDI-8.* Cairo, Egypt.

Evans, D. J. (1997) Infrastructures for Sharing Geographic Information among Environmental agencies. Massachusetts Institute of Technology.

FCC (2006) Federal Communications Commission. http://www.fcc.gov/911/enhanced/.

Floridi, L. (2004) Philosophy. *The Philosophers' Magazine* http://www.philosophers.co.uk/index.htm.

Floridi, L. & Sanders, J. W. (2002) Mapping the Foundationalist debate in Computer Ethics. *Ethics and Information technology,* 4**,** 1-9.

Foster, H., Uchitel, S., Magee, J. & Kramer, J. (2003) Model-based verification of Web service compositions. *Automated Software Engineering, 2003. Proceedings. 18th IEEE International Conference on***,** 152-161.

Foster, I. (2005) Service-Oriented Science. American Association for the Advancement of Science.

Friis-christensen, A., Nytun, J. P., Jensen, C. S. & Skogan, D. (2005) A Conceptual Schema Language for the Management of Multiple Representations of Geographic Entities. *Transactions in GIS,* 9**,** 345-380.

Hirschheim, R. & Klein, H. K. (1989) Four Paradigms of Information Systems Development. *Communications of the ACM,* 32**,** 1199-1216.

Hoyningen-huene, P. (1993) Reconstructing Scientific Resolutions. Thomas S. Kuhn's Philosophy of Science (trans. Levine A T). University of Chicago Press, Chicago.

Kampshoff, S. (2006) Mathematical Models for Geometrical Integration. Geodetic Institute, RWTH Aachen Univeristy, Germany.

Kilpelainen, T. (1997) Multiple Representation and Generalization of Geo-databases for Topographic Maps. Finnish Geodetic Institute.

Kuhn, T. S. (1962) Historical Structure of Scientific Discovery. *Science, American Association of the Advancement of Science,* 136.

Lau, F. (1999) Toward a framework for action research in information systems studies. *Information Technology and People,* 12**,** 148-175.

Malhotra, Y. (2000) Knowledge Management for [E-] Business Performance. Information Strategy. *The Executives Journal,* 16**,** 5-16.

Mingers, J. (2001) Combining IS research methods: towards a pluralist methodology. *Information Systems Research,* 12**,** 240-259.

Musinguzi, M., Bax, G. & Tickodri-togboa, S. S. (2004) Opportunities and Challenges for SDI development in development countries - A Case study of Uganda. *Proc. 12th Int. Conf. on Geoinformatics - Geospatial Information Research: Bridging the Pacific and Atlantic.* University of Gvle, Sweden, Geoinformatics.

Najar, C., Rajabifard, A., Williamson, I. & Giger, C. (2006) A Framework for Comparing Spatial Data Infrastructures An Australian-Swiss Case Study. *GSDI-9 Conference Proceedings.* Santiago, Chile.

Nebert, D. (2003) *Issues of Data Standardization*, Global Spatial Data Infrastructure (GSDI).

Niehaves, B. (2005) Epistemological perspectives on multi-method information system Research" , . European Research Center for Information Systems, University of Mnster, Leonardo-Campus 3, 48140 Mnster, Germany.

Papazoglou, M. P. (2003) Service -Oriented Computing: Concepts, Characteristics and Directions. *Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE'03).*

Peikoff, L. (2000) Objectivism: The Philosophy of Ayn Rand. www.peikoff.com.

Sharma, J. (2005) Spatial Database Architecture - Influence of IT Trends on Spatial Information Management. *Directions Magazine.* Directions Magazine.

Sharpe, B. (2002) Trends in Information Technology. *The Appliance Studio Ltd.*

Skog, J. (1998) Trends in applying spatial Technologies. *13th ESRI European User Conference.* France.

Sohir, M. H. (2005) The Role of ESA in Building the Egyptian Spatial Data Infrastructure (ESDI) Towards the Electronic Government (E-gov.). Spatial Portals and e-Government. *From Pharaohs to Geoinformatics, FIG Working Week 2005 and GSDI-8* Cairo, Egypt.

Sonnen, D. (2005) Spatial Information Management - Then, Now, Next. *Direction Magazine* http://www.directionsmag.com.

Strader, T., Tarasewich, P. & Nickerson, R. C. (2004) The State of Wireless Information Systems and Mobile Commerce Research *Information Systems and e-Business Management,* 2**,** 287-292,.

Wangler, B. & Backlund, A. (2005) Information Systems Engineering: What is it? *CAiSE Workshops,* 2**,** 427-437.

Weber, R. (2004) The Rhetoric of Positivism versus Interpretivism. *MIS Quarterly,* 28**,** iii-xii.

Wikipedia (2006) Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Philosophy.

Ziegler, P. & Dittrich, K. R. (2004) Three decades of data integration - all problems solved? *Database Technology Research Group.* Winterthurerstrasse 190, CH-8057 Zurich, Switzerland, Department of Informatics, university of Zurich.

14
A Spatial Decision Support Tool for Landfill Site Selection: Case For Municipal Solid Waste Management

Nakakawa Agnes and Ogao P.J.

One of the problems faced worldwide is waste management. It involves several activities, which can be categorized into: collection, transportation and disposal of waste. Computerizing the processes involved in these activities can help to improve efficiency and effectiveness in waste management. However, the process of particular interest and environmental concern is the effective selection of sites for waste disposal (landfill sites). The current process of selecting landfills in Uganda is manual, costly and time consuming. This paper presents the design and development of a Spatial Decision Support Tool as a computer-based technology that can be used to solve the complex process of landfill site selection for municipal solid waste management in Kampala and the neighboring Wakiso districts. Several parameters required in landfill site selection such as an area's distance from: roads, rivers, lakes, wetlands, towns, gazetted land, and the soil type, topography, land cover of an area among others were considered. The tool was developed based on existing literature on landfill site selection, Spatial Decision Support Systems (SDSS), Geographical Information Systems (GIS) and Multi-Criteria Evaluation (MCE). The model used by the tool was designed using ArcView Spatial Analyst 2.0, the user interfaces were designed using Avenue Programming language and Visual Basic 6.0, and ArcView GIS 3.2a provided the Database Management System. The results of the tool were validated using Ground Truthing. The results of this study can be very helpful during the procurement process of landfill sites; that is, the concerned authorities can save time and costs associated with inspection and evaluation for bidders whose sites are located far outside the range of potentially identified areas.

Introduction

Currently Uganda's urban areas are characterized by "careless and indiscriminate open waste-space-dumping" (NEMA, 1998). Over the years effective management of solid waste has been a major problem in the city of Kampala and Uganda at large. Poor waste management contributes to poor environmental and unsanitary conditions that threaten public health and the quality of life of urban dwellers (NEMA, 1998; NEMA, 2002). This is because the location of a landfill is a major determinant of the extent to which the landfill will pose an environmental risk (EPA, 1998).

Careless waste disposal results in pollution of surface and ground water, impairing soil permeability, and blockage of drainage systems (USEPA, 2002). Leachate from such dumps mixes with rainwater, hence contaminating drinking

water and degrades the water resource (EPA, 1998; Vitalis and Manoliadis, 2002). According to USEPA (2002), solid waste management can be improved by "either operating a properly sited, designed, constructed, and managed landfill, or burning of waste in a controlled facility that converts waste to energy". However urban areas of Uganda lack landfills, what mostly exists are poorly sited dumpsites (due to limited technical capacity for suitable landfill siting), which lack management for proper operation (NEMA, 2001).

Landfill site selection has been a problem not only experienced in Uganda but worldwide. However, over the years several researchers have discovered ways of partially solving this problem by applying computer based techniques in the process of selecting suitable landfill sites. According to Gao *et al., (*2004), any decision making process that focuses on problems that are either dependant or influenced by geographical information is spatial decision making.

Spatial Decision Making (SDM) Site selection is a semi-structured process that involves consideration of several factors before an optimal solution is got (Eldrandaly *et al.,* 2003). GIS as a decision support tool, simplifies the search for suitable sites for a particular purpose because of its capability of spatial feature extraction and classification (Vitalis and Manoliadis, 2002). Site selection process is complex hence the need to integrate several decision support tools such as Expert Systems (ES), GIS, and MCE for an optimal selection (Eldrandaly *et al.,* 2003).

A Decision Support System (DSS) is a class of computerized information systems that support semi-structured decision-making activities through the support for: development and use of mathematical models, and data manipulation (Power, 2004). In addition, SDSS is an interactive, computer-based system that supports a user or group of users in achieving a higher effectiveness of decision making while solving a semi-structured spatial decision problem (Malczewski, 1997; Peterson, 1998). SDSS support execution, interpretation, visualization, and interactive analysis of spatial models over multiple scenarios encountered in site selection (Hemant *et al*., 1999). GIS is a computer-based technology and methodology for collecting, processing, managing, analyzing, modeling, and presenting geographic (spatial) data for a wide range of applications (Eldrandaly *et al.*, 2003; Power 2004).

The integration of both GIS and Multi-Criteria Decision Analysis (MCDA) techniques improves decision-making because it enhances an environment for transformation and combination of geographical data and stakeholders' preferences (Malczewski, 2006). In site selection problems, GIS perform deterministic Overlay and buffer operations while; MCDM methods evaluate alternatives based on the decision maker's subjective values and priorities (Eldrandaly *et al.,* 2003).

A spatial problem can have both spatial and non-spatial aspects, all of which must be considered in decision-making for optimal results (Gao *et al.,* 2004). There is need for integration of a spatial model (representing spatial aspects) and a non-spatial model (representing non-spatial aspects) of the problem (Gao *et al.,* 2004; Eldrandaly *et al*., 2003). A spatial model is a set of spatial processes that convert input data into an output map using a specific function such as buffer or overlay

(Gao *et al.,* 2004). In this study, vector and grid (raster) data were both used in spatial analysis.

Spatial Modeling is the process of manipulating and analyzing spatial data to generate useful information for solving complex problems (Haggett and Chorley, 1967). Spatial modeling is useful when finding relationships among geographic features and helps decision makers to logically address the spatial problem (Gao *et al.,* 2004).

Related work

Research has been done in the area of SDM in site selection (Vitalis and Manoliadis, 2002; Adinarayana, 2003; Eldrandaly *et al.,* 2003; Herzog, 1999; Gao *et al.,* 2004; Lourdes, 1996; Gaim, 2004; Ramos, 2004; Yaw *et al.,* 2006 among others). The review of literature identified a need for development of a spatial decision support tool for landfill site selection for municipal solid waste management for Kampala district. This is because the existing landfill site selection systems or tools and multi-task site selection projects developed in other countries (like; Western Macedonia-Greece (Vitalis and Manoliadis, 2002), Philippines (Lourdes, 1996), South Calfornia- Savannah River site (Ehler *et al.,* 1995), Malaysia (Gaim, 2004), Larimer county- Colorado USA (Herzog, 1999), among others) cannot be efficiently and effectively used for the case of Kampala- Uganda without any modifications, due to the following reasons:

(i) Though the landfill site selection criterion is almost similar all over the world, there are some constraints related to locality, which may cause criteria factors to conflict with each other (Michael (1991) as cited by Yagoub and Taher, 1998). In addition, waste hierarchy options for different countries differ because of different geography, culture, environment, urban structure, and planning system among other factors (Pitt, 2005).

(ii) Tools designed by other countries are not readily available for direct use or for modification in order to be used for the case of Uganda.

Site Selection Criteria for Landfills

Several countries (like Australia, Malaysia, Niger, North Dakota, Philippines, Uganda, and United States among others) have put in place rules to follow when selecting suitable sites for sanitary landfills. These guidelines act as the primary mechanism used to protect the host community and the environment at large. Below are the factors that several researchers (EPA, 1998; EPA, 2003; Malaysian Government, 1995 as quoted by Gaim, 2004; North Dakota Department of Health, 2002; KCC, 2000; Yaw *et al.,* 2006 among others) have used to determine the appropriateness of a site to be used as a sanitary landfill.

(i) ***Site Capacity:*** A site should be capable of providing at least 10 years of use in order to; minimize costs for site establishment, operation and closure.

(ii) ***Land cover:*** buffer zones should be provided between the landfill and sensitive areas or other land uses. For example; at least 100m from public

roads, at least 200m from industrial developments, at least 500m from urban residential or commercial area, at least 1000m from rural residential areas. For the case of Malaysia (Gaim, 2004), land use types such as grassland, forests and cultivated land were considered appropriate for dumping except marshland and swamp type. For this study, grassland and bush land areas were considered appropriate for a landfill site.

(iii) ***Airports:*** The distance between an airport and a landfill should be a minimum of 3km, unless there is a clear demonstration of bird control measures at the landfill.

(iv) ***Surface Water:*** The distance between a landfill and the nearest surface water should be a minimum of 100m, or 200m to minimize the risk of polluting water with leachate. However, North Dakota Department of Health (2002) uses a minimum distance of 60m to the nearest surface water, and Yaw *et al.*, (2006), a distance of 300m from any water body. For this study, at least 200m were considered appropriate for a landfill site.

(v) ***Groundwater:*** An extremely deep water table region is suitable so that underground water is not contaminated by the leachate of the waste. According to North Dakota Department of Health (2002), the bottom of disposal trench should be at least four feet above the water table.

(vi) ***Local Topography:*** landforms located in flat or undulating land, in a disused quarry are suitable for waste disposal. Major landfills must not be sited in hilly areas, those with ground slopes nominally greater than 10%. However, EPA (2003) recommends a slope less than 5 %, and North Dakota Department of Health (2002), 15% slope or less. For this study, 20% or less was considered appropriate for a landfill site.

(vii) ***Soils:*** soil should be of sufficiently low permeability to significantly slow the passage of leachate from the site. Thus, sites in clay-rich environments are preferable.

(viii) ***Climate:*** areas with heavy rainfall need extra care to avoid side effects of drainage and erosion, sites with prevailing winds require extra efforts to control litter and dust.

(ix) ***Unstable Areas:*** landfills must not be located within 100m of an unstable area.

(x) ***Infrastructure:*** Although landfills should have suitable transport access, with power and water available, they should not be located within 100m of any major highways, city streets or other transportation routes. Yaw *et al.,* (2006) recommends 300m. It would be more cost efficient for landfills not to be located so far away in order to avoid high transportation costs. For this study, at least 200m were considered appropriate for a landfill site.

(xi) ***Local Flora and Fauna:*** Sites that contain protected or endangered fauna and/or flora, or sensitive ecosystems are unsuitable for landfill facilities.

(xii) ***Distance from environmentally sensitive or protected areas:*** A landfill must not be located in close proximity to sensitive areas such as fish sanctuaries, mangrove areas and areas gazetted for special protection would be excluded. Therefore a 3,000m buffer is necessary to surround an environmentally sensitive area. EPA (1998) recommends a buffer of at least 500m. For this study, at least a 3000m buffer was considered appropriate for a landfill site.

(xiii) ***Distance from urban areas:*** Landfills should not be placed too close to high-density urban areas in order to mitigate conflicts relating to the Not in My Back Yard syndrome (NIMBY). This guards against health problems, noise complaints, odour complaints, decreased property values and mischief due to scavenging animals. Development of landfills should be prohibited within 3000m from village or rural settlements. EPA (2003) recommends at least 500m from an urban residential or commercial area. Yaw *et al.,* (2006), a distance of 4000 m from a town. For this study, at least a 3000m buffer was considered appropriate for a landfill site.

(xiv) ***Population:*** Gaim (2004) recommends that areas with a population density less than 200 were regarded as suitable for landfills.

In addition to the above guidelines, some countries have developed landfill site selection systems or tools that use the above guidelines to make the site selection process efficient and effective (for example; Western Macedonia-Greece (Vitalis and Manoliadis, 2002), Philippines (Lourdes, 1996), South Calfornia- Savannah River site (Ehler *et al.,* 1995), Malaysia (Gaim, 2004), Larimer county- Colorado USA (Herzog, 1999), among others).

However Uganda still has a manual and time-consuming landfill site selection procedure that uses the guidelines given in KCC (2000). Through advertising, Kampala City Council (KCC) expresses interest in purchasing land that satisfies the above criteria, and then willing bidders respond to the advertisement. Thereafter the procurement process involves several activities one of which is the inspection and evaluation of sites that bidders are offering. This process is tedious, costly and time consuming. Kampala district currently has one sanitary landfill site located at Mpererwe, which was acquired using the above process. Currently, there is no computer-based system to help KCC have an effective and efficient landfill site selection process.

Methodology

All data that was used in the study was collected using the following methods: Library Research Method, Interviews with the Landfill Site Contractor of KCC, and Observation method. The digital maps (or data sets) of the required parameters for Kampala and Wakiso districts were obtained from the GIS labs of National

Forestry Authority (NFA), KCC and Uganda Bureau of Statistics (UBOS).

The architecture and Model of the tool (Figure 1 and Figure 2 respectively) were designed by modifying the architecture and framework of a decision support system for industrial site selection (Figure 3 and Figure 4 respectively) designed by Eldrandaly *et al.,* (2003).

The landfill site selection process performed by this tool is divided into two phases as explained below:

(i) First phase (Spatial Modeling phase) involves the use of ArcView GIS Geoprocessing tools to create buffers, intersects and select sites that satisfy the landfill site selection criteria; for example, distance from water bodies, towns, roads, among others. Themes were categorized into geographical, surface water and social factors. Weighted Overlay process and Multi-Criteria Evaluation were performed on all the parameters in each category; and the same processes were again applied on all the 3 category output maps. The remaining areas (areas or sites satisfying all the constraints of all layers or parameters) were further evaluated in phase II.

(ii) Second phase involves further analysis and evaluation of the potential sites obtained in the phase 1. Additional factors used in this phase include; site capacity, population distribution, and other factors obtained as a result of ground Truthing. Factors considered after Ground Truthing include; government policy on an area, and cultural beliefs or norms of people living in that area. Finally the tool gives the sites recommended for municipal solid waste disposal in the study area. However, the tool allows the decision maker to experiment the impact of weights (preferences of any criteria) assigned to each parameter; which provides flexibility during decision-making.

Tool Evaluation was done using both quantitatively and qualitatively measures. Qualitative Evaluation ensures that the system meets the user's requirements whereas Quantitative evaluation ensures that the system performs the required tasks effectively within the constraints of time, CPU speed, data storage limits (Goodchild and Rizzo, 1987) as cited by Ehler *et al.,* (1995)).

Architecture of the Prototype Spatial Decision Support Tool

This tool is based on a 3-tier architecture and below is a description of each tier;

(i) Database Management Component: ArcView GIS 3.2a performs the function of storing and maintaining the digital maps of all the required parameters in the digital GIS database. All data was obtained in vector format, which was later converted to grid (raster) format for analysis purposes. ArcView GIS 3.2a was used because according to Connolly and Begg (2004), GIS database stores various types and large volumes of spatial and temporal information, derived from survey and satellite photographs; and the data requires operations (on distance, intersection and others) that SQL does not provide.

(ii) Model Base Management Component; contains the spatial model and the MCE analysis procedures of the tool. This component was developed as follows:

a) Spatial Analyst 2.0a ModelBuilder was used to design the model using routines and processes involved in spatial analysis. GIS routines for Clipping, Data Conversion, Buffering, Weighted Overlay, Data Classification, and Slope Analysis were all used.

b) MCE was done together with the Weighted Overlay process in order to include the decision maker's preference in the landfill site selection criteria.

(iii) User Interface Component; was designed using scripts written in Avenue (ArcView's Object Oriented Programming Language) and the Dialog Designer extension provided in ArcView 3.2a. The Interface is used to; enter the explicit spatial datasets, implicit spatial datasets (parameters), and non-spatial input like decision maker's preferences, perform model simulation, and display maps and reports of analysis results.

Results

The landfill site selection model (spatial model) used in this tool was developed basing on the information obtained from the review of the different criteria used in other countries, and also from the expert knowledge that was provided by the Solid Waste Engineer of KCC and other officials who were interviewed. The parameters used in the tool are both spatial (Soils, Land Cover/Land Use, Topography, Lakes, Rivers, Wetlands, Major Towns, Roads, and Protected Areas) and non-spatial factors (Capacity, population distribution, and government policy). The digital maps used were taken and prepared in 2000 by NFA, which creates digital maps after every 5 years because of the high costs involved in the creation of those maps and the time required for cleaning and validating the maps.

The weight assigned to each parameter (see Equation 1) illustrates the relative importance of that parameter in the site selection process and the decision maker's preference to a particular landfill site selection criterion. The presented weights in the model were obtained after several simulations were done. Spatial Modeling (Phase I) using the above information, yielded the Potential Landfill Sites Weighted Overlay Map, which shows areas that meet the landfill site selection criteria of only the spatial factors.

For phase II of site selection, Query Builder was used to select only the potential landfill sites from the entire overlay map. All the selected features were converted to a separate vector theme mainly because vector data represents geographic features with greater precision compared to grid data (ESRI, 2000). This resulted in 249 potential sites for landfills in Kampala and Wakiso Districts (with the current Mpererwe sanitary landfill site inclusive). Non-spatial factors (for example site capacity, population distribution of the potentially identified areas, government

policy on the potentially identified areas, and land ownership) were not included in the spatial model, but were used to evaluate the 249 potential sites obtained in phase I.

Qualitative Evaluation/Validation was done by Ground Truthing, which involved visiting the potentially identified sites. Sites in 3 areas (Mpererwe, Naluvule and Buloba) were visited. However the accuracy of the obtained results is subject to; the time of satellite imagery (the time when the satellite images were taken), the accuracy of the data (that was obtained from KCC, NFA, and UBOS), and the weights a user or decision maker assigns to parameters.

The results of this work were acceptable by the Solid Waste Engineer of KCC; in the sense that parishes where some of the potential sites (identified by this study) are located, were among the areas that bidders who responded to the "The Call for Bidders" by KCC were offering in response to the advertisement. Those areas include sites in Nangabo, Kyambogo, and Wakiso subcounties. Evaluation reduced the potential sites from 249 to 27 potential sites (see Figure 5 and Figure 6), with the current Mpererwe sanitary landfill site inclusive.

Quantitative Evaluation/Validation involved testing or running the tool on platforms that had varying specifications, and it was able to effectively run on platforms. The tool enables one to select landfill sites in less time compared to the time taken during the procurement process of landfill sites (especially the task of inspection and evaluation of sites offered by bidders to KCC) and debates or number of meetings held before any site is chosen.

Conclusions and Future Work

Locating suitable sites for any purpose (for example industrial, landfills, and road construction or infrastructure development among others) has always been a challenge due to NIMBY attitudes among the communities (Ramos, 2004; Yaw *et al.*, 2006). This implies that decision makers should aim at choosing sites that cause minimum conflicts since it's not possible to find sites that cause no opposition (Yaw *et al.*, 2006). This study only covers municipal solid waste disposal in Kampala and Wakiso districts, and the tool is limited to a desktop implementation because of limitations of high costs, complexity regulation and time. The design, operations and maintenance of a landfill are out of the scope of this work.

The findings of this work can be useful to planners and researchers since this work serves as a guide for further development and research in the following areas:

(i) Ground water levels of the potential sites were not measured because of time limitations, costs and technical equipment required. Performing Ground water level analysis will modify and improve the results.

(ii) Digital maps for the entire country (Uganda) can be used so as to select landfill sites countrywide. Further Ground Truthing can then be done, so that the output can be stored as a separate digital map for areas suitable for landfills. In this, such areas can be reserved for waste disposal and not compromised to other purposes or services.

(iii) Lastly, the framework and implementation of this tool can be modified to include several stakeholders in the landfill site selection decision-making process. That will lead to incorporation of preferences of several stakeholders (for example Parish leaders, NEMA, Ministry of lands, landowners and community among others) in the landfill site selection decision-making exercise, and therefore minimize conflict. Such analysis will eventually yield the most suitable sites for waste disposal in the entire country.

References

Ehler, G. Cowen, D., and Mackey, *H.* (1995). Design and Implementation of a Spatial Decision Support System for Site Selection. *ESRI User Conference Proceedings,* California May 22-26.

Eldrandaly, E., Neil, E., and Dan, S. (2003). A COM-based Spatial Decision Support System for Industrial Site Selection. *Journal of Geographic Information and Decision Analysis,* 7(2): 72 92.

EPA (1998). Guidelines for Major Solid Waste Landfill Depots in South Australia. http://www.epa.sa.gov.au/pdfs/swlandfill.pdf, Accessed March 2006.

EPA (2003). Guidelines for the Siting, Design and Management of Solid Waste Disposal Sites in the Northern Territory. http://www.nt.gov.au/nreta/environment/waste/codes/pdf, Accessed March 2006.

ESRI (2000). ModelBuilder for ArcView GIS Spatial Analyst 2. http://www.esri.com, Accessed April 2006.

Gaim, J.K. (2004). GIS as Decision Support Tool for Landfills Siting. *Proceedings of Map Asia 2004,* Beijing, China. August 26-29.

Gao, S., Sundaram, D., and Paynter, J. (2004). Flexible Support for Spatial Decision-Making. *Proceedings of the 37th Hawaii International Conference on System Sciences*, 03(3): 30064a.

Haggett, P., and Chorley, R. (1967). *Models in Geography.* London: Methuen.

Hemant, K.B., Suresh, S., and Craig, H. (1999). Beyond Spreadsheets: Software for Building Decision Support Systems. *IEEE Computer*, **32**(3): 31-39.

Herzog T., M. (1999). Suitability Analysis Decision Support System for Landfill Siting (and other purposes), *19th Annual ESRI International User Conference Proceedings,* San Diego, California.

KCC (2000). Environment Impact Statement for the Proposed Extension to Mpererwe Sanitary Landfill. KCC Library.

Lourdes, V.A. (1996). Solid waste disposal site selection using Image Processing and Geographic Information Systems (GIS) techniques, *Proceedings of the 17th Asian Conference on Remote Sensing, SriLanka November 4 – 8.*

Malczewski, J. (1997). Spatial Decision Support Systems, NCGIA Core Curriculum in GIScience.http://www.ncgia.ucsb.edu/giscc/units/u127/u127.html, Accessed December 2006.

NEMA. (1998). State of Environment Uganda Report for Uganda. Kampala, NEMA library. Uganda.

NEMA. (2001). State of Environment Uganda Report for Uganda. Kampala, NEMA library. Uganda.

NEMA. (2002). Quarterly News Letter, Kampala, NEMA library. Uganda.

North Dakota Department of Health - Division of Waste Management, (2002). Guideline 14 Emergency Waste Disposal Variance Notification: Dead Or Diseased Livestock. http://www.health.state.nd.us, Accessed May 2006.

1Pitt, M. (2005). Trends in shopping center waste management. *Facilities*, **23**(11/12): 522- 533.

Power, D. J. (2004). Free Decision Support Systems Glossary. http://DSSResources.COM/glossary, Accessed September 2005.

Ramos, R. P. (2004). Establishing Disposal Siting Mechanism Towards a Sustainable Industrial Waste Management in the Philippines. *NZSSES International Conference Proceedings and Presentations on Sustainability Engineering and Science.* New Zealand Society.

United States Environmental Protection Agency, (2002). Solid Waste and Emergency Response. http://www.epa.gov/globalwarming. Accessed March 2006

Vitalis, K., and Manoliadis, O. (2002). A two level multi-criteria DSS for Landfill Site Selection Environmental Protection and Ecology International. *Journal of Geographical Information Systems,* 06(1): 49-56.

Yagoub, M.M., and Taher, B. (1998). GIS Applications for Dumping Site Selection. *Proceedings of the 18th Annual ESRI International User Conference on Geographic Information System (GIS) Applications,* San Diego Convention Center, July 27-31.

Yaw, A.T. *et al.,* (2006). Use of Geo-Spatial Data for Sustainable Management of Solid Waste in Niamey, Niger. *Journal of Sustainable Development in Africa,* 08(1): 203-210.

Appendix

Fig 1: Architecture of the Prototype Spatial Decision Support Tool

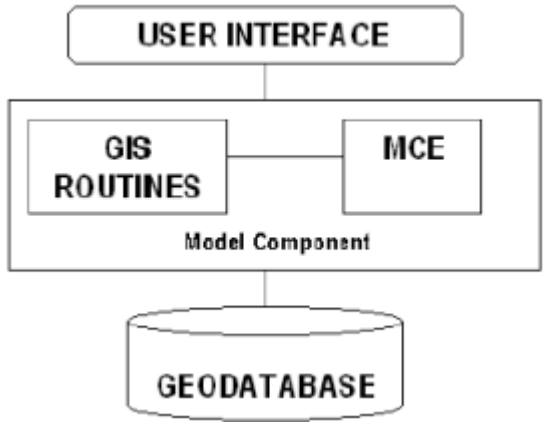


Fig 2: Conceptual Model of the Spatial Decision Support Tool

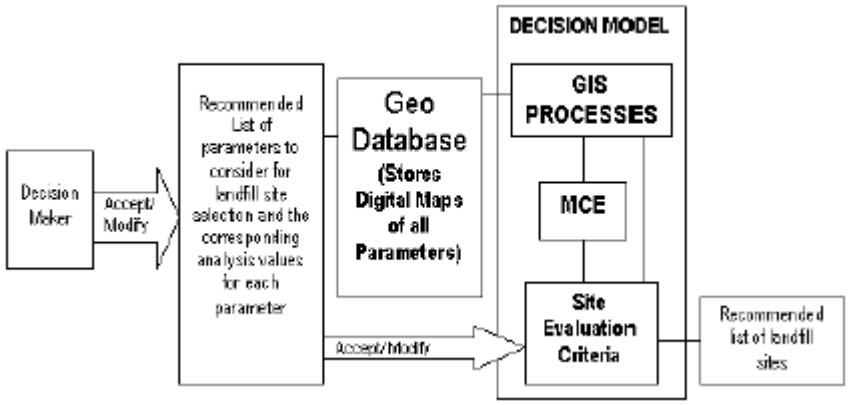
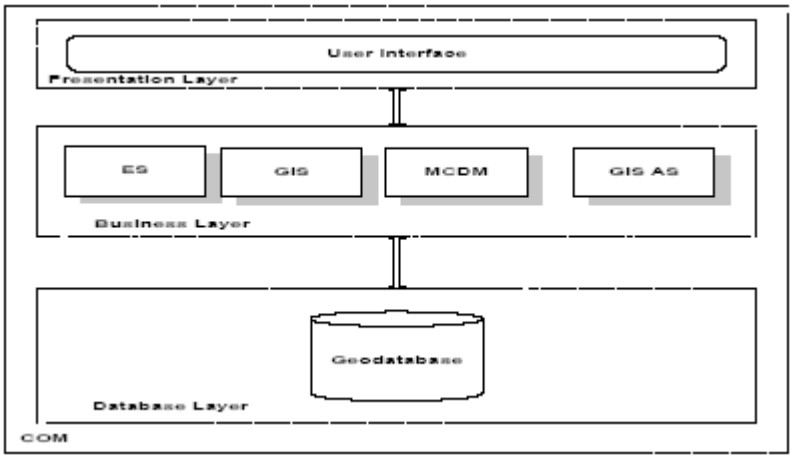


Fig 3: Three-Tier Architecture for the SDSS for Industrial Site Selection by Eldrandaly *et al.*, 2003



G = 0.2 (Slope) + 0.3 (Land cover) + 0.5 (Soils)
SW = 0.4 (Lakes) + 0.3 (Rivers) + 0.3 (Wetlands)
SF = 0.4 (Towns) + 0.3 (Protected Areas) + 0.3 (Roads)
Therefore;

Potential Landfill Sites Weighted Overlay Map = 0.4 G + 0.3 SW + 0.3 SF

Equation 1: Equation used in the Spatial Model

Fig 4: Framework for the SDSS for Industrial Site Selection by Eldrandaly *et al.*, 2003

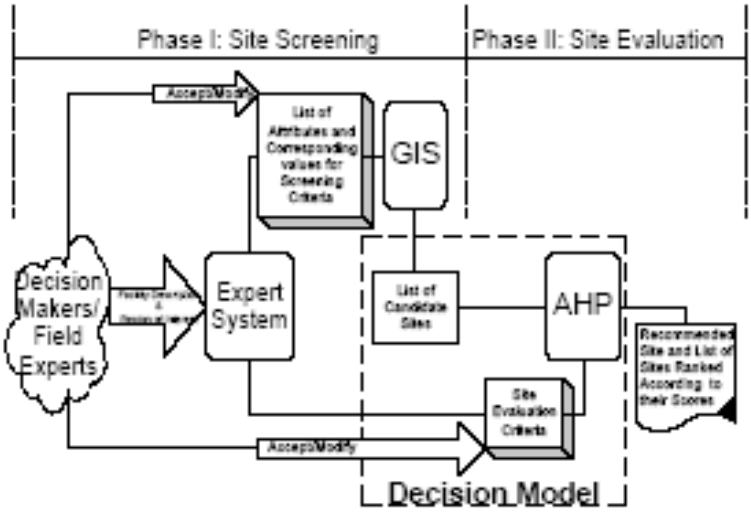


Fig 5: The 27 Potential Landfill Sites that are more than 30 Acres in Area

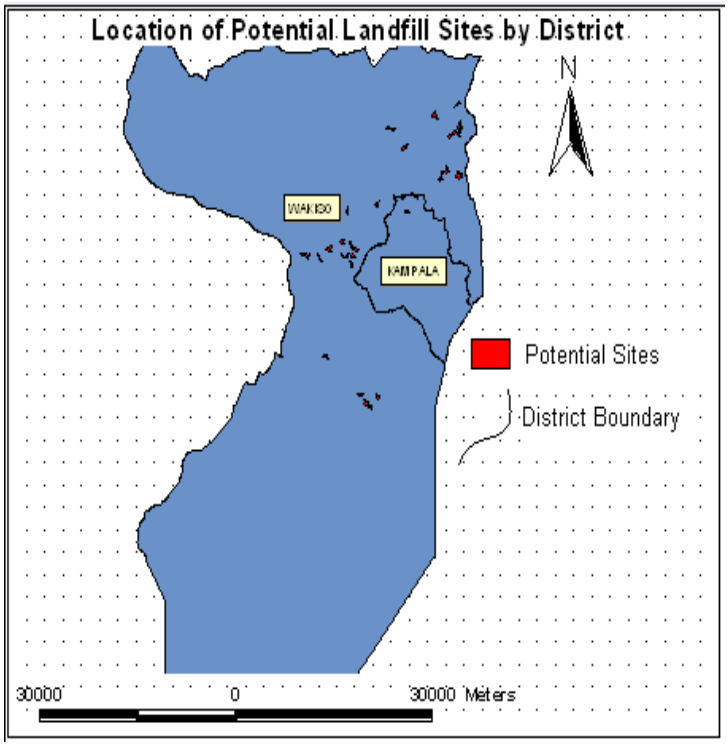
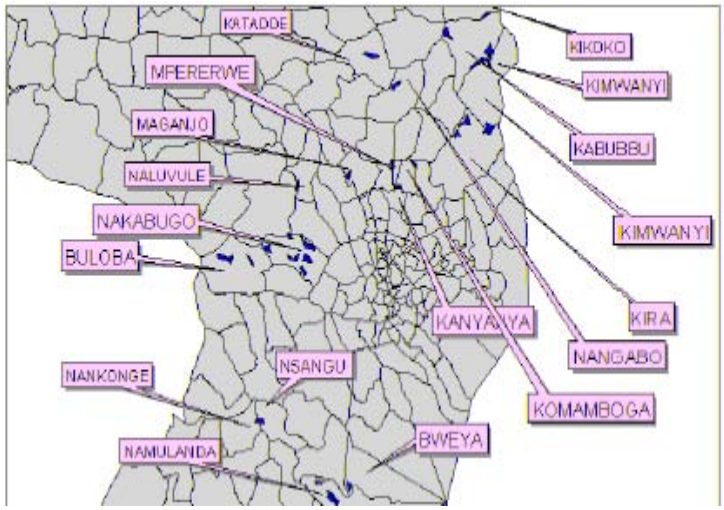


Fig 6: Parish Location of the 27 Potential Landfill Sites in Kampala and Wakiso Districts



15
Web Accessibility in Uganda: A study of Webmaster Perceptions

Rehema Baguma, Tom Wanyama, Patrick van Bommel and Patrick Ogao

Despite the fact that the proportion of people with disabilities in society has been increasing, many websites have remained inaccessible to them. In Uganda, a study on government agency websites established that 100% of the studied websites were not accessible to people with disabilities. While guidelines and tools for developing accessible websites exist in the public domain, research surrounding perceptions of IT workers about Web accessibility and how they interact with the guidelines at country level do not exist. In this paper, we examine the practice and perceptions of webmasters in Uganda on the accessibility of government websites and what can be done to improve the situation. This understanding is important to increase on the knowledge of why government websites in Uganda are not accessible and what the stakeholders can do to improve the situation.

Introduction

A considerable number of users of the Web have various types of disabilities such as vision, hearing, motor and cognitive impairments (Lazar et al., 2004). World wide, people with disabilities are estimated at 20% of the total world population (Shi, 2005). Web users with disabilities rely on assistive technologies to use the Web effectively. Such technologies include screen readers, voice recognition equipment, alternative pointing devices, alternate keyboards and refreshable braille displays (Lazaar et al., 2004). An accessible website is one that is sufficiently flexible to be used by all assistive technologies just as accessible buildings offer curb cuts, ramps, and elevators to allow people with disabilities to enter and navigate through the building with ease (Lazar et al., 2004).

Currently, there are a number of guidelines and tools Web designers and webmasters can use to make their websites accessible to people with disabilities. Such guidelines include the Web Content Accessibility guidelines (WCAG) developed by the World Wide Web Consortium (W3C), the US government's Section 508 Initiative, Americans with Disabilities Act (ADA), Australians with Disabilities Act and the National Institute on Ageing Guidelines (NIA). Similar guidelines exist in Canada, UK and Portugal. In addition to the guidelines, automated software tools that help in finding accessibility flaws in websites before the sites are publicly posted, are available. Such tools include bobby, ramp, infocus and a-prompt. More so new versions of web development tools such as dream weaver and front page include tools that assist developers with accessibility related issues (Lazaar et al., 2004).

Given that the guidelines and tools for realizing accessible websites are available, it would be expected for most websites to be accessible to people with disabilities. But the reverse is true even in countries where Web accessibility is a legal requirement (Lazar et al., 2004). World wide, a large percentage of websites (70–98%) are not accessible. In a study conducted by Nielsen (Nielsen, 2001), the usability of most current websites is on average three times higher for users without disabilities than for those who are blind or have low vision. In another research project published by Forrester Research (Huang, 2003), it was found that only one in four e-commerce sites surveyed met the minimum requirements provided by the Web Content Accessibility Guidelines (WCAG). Even in the public sector of the U.S., where Web accessibility is a legal mandate, a significant number of official websites still contain features that do not provide reasonable access to users with disabilities. Although Section 508 requires agencies to ensure that persons with disabilities have equal access to and use of federal e-government websites, widespread accessibility on e-government sites has not materialized since the 2001 compliance deadline (Jaeger, 2006). Studies of the accessibility of federal e-government sites have found low levels of accessibility, with usually less than one-third of sites being labelled accessible by these studies (Jaeger, 2006). In Taiwan, 83% of central government websites are not accessible to people with disabilities especially visual disabilities (Huang, 2003). In Uganda, a study on the accessibility of government websites revealed that only 14% of the studied websites provided some level of accessibility (7.1%). However, it was not clear whether the 7.1% conformance was out of a deliberate intention to make the sites accessible to people with disabilities or rather accidental or a result of other design considerations (Baguma et al., 2007).

Given that all the resources for making websites accessible are available, it is unclear why many websites have remained inaccessible. While guidelines for Web accessibility exist, research surrounding the effectiveness of those guidelines, how IT workers interact with those guidelines, and reasons for implementing accessibility, do not exist (Lazar et al., 2004). The people who decide whether a site will be built for accessibility or not are the Web developers and the clients. It is likely that if neither of these groups of people are aware of or passionate about Web accessibility, then a website will be built to be inaccessible (Lazar et al., 2004). However, the person that has the greatest influence on an existing web site is the webmaster (Lazar et al., 2004). The goal of this research was to examine why government websites in Uganda are not accessible to people with disabilities. The researchers created a survey to learn more about webmasters' perceptions and knowledge on Web accessibility. The results of this research are expected to increase on the knowledge about why government websites in Uganda are not accessible and how this situation can be improved.

The remainder of this paper is organized into five sections as follows: Web accessibility guidelines, related work, Web accessibility integration model, findings, discussion of the findings, conclusion and future work.

Web accessibility guidelines

To-date, a number of guidelines that Web designers and webmasters can follow to make their websites accessible to people with disabilities are available. A summary of the most prominent guidelines and their approach to realising Web accessibility follows:

The Web Content Accessibility Guidelines (WCAG): WCAG is a set of international Web Accessibility guidelines produced by the World Wide Web Consortium (W3C) for the design of accessible Web sites (Shi, 2005). The guidelines address two general themes- that is ensuring graceful transformation to accessible designs, and making content understandable and navigable. They are composed of fourteen specific guidelines, with each including the rationale behind the guideline and a list of checkpoint definitions. Each checkpoint is assigned a priority level – that is one, two or three based on the checkpoint's impact on accessibility.

The WCAG are intended for all Web content developers (page authors and site designers) and for developers of authoring tools such as HTML editors. They are recognized as the authority for designing and creating accessible web- sites, and have been used by several software developers to develop accessibility authoring and checking tools such as bobby (www.cast.org/bobby/) (Huang, 2003; Shi, 2005).

The Americans with Disabilities Act (ADA) of 1990: ADA includes several provisions that require employers to provide "reasonable accommodation" and mechanisms for "effective communication" to workers with disabilities. This law is applicable to the entire nation, not only to entities that receive federal funds. It was originally focused on areas such as employment, public accommodations, and telecommunication services. However, the subsequent growth of the Internet for communication in education, business, government and work settings has now broadened the scope to cover the Internet and the World Wide Web (Chiang et al., 2005).

Section 508 of the Rehabilitation Act

Section 508 of the Rehabilitation Act on the other hand defines the processes used by the U.S. federal government to procure electronic and information technology systems. One of the central aspects of the law is to ensure accessibility of electronic and information technology systems to people with disabilities who are federal employees or members of the general public (Huang, 2003). Section 508 has a lot of similarity with the Web Content Accessibility Guidelines (WCAG). This is possibly because it is based on the U.S. Access Board's Electronic and Information Technology Accessibility Standards, which are in turn based on the WCAG. This relationship has made it an important legal reference for Web accessibility in the U.S (Jaeger, 2006).

Both ADA and Section 508 make reference to WCAG as a more comprehensive Web accessibility resource developed by the W3C that helps designers make web pages as accessible as possible to the widest range of users, including users with

disabilities. In fact, Section 508 contains more or less the same guidelines as the WCAG but in a more compressed form. The major difference between the two sets of guidelines is that Section 508 is legally binding in the U.S. where as WCAG is an open non-legally binding Web Accessibility specification.

Australian Disability Discrimination Act (ADA): The Australian Disability Discrimination Act makes it unlawful for a service provider to discriminate against a person with disability by refusing to provide any service, which it provides to members of the public. A service provider is required to take reasonable steps to change a practice, which makes it unreasonably difficult for people with disabilities to make use of its services. According to the Australian Disability Discrimination Act 1992, inaccessible websites or pages are a sort of discrimination against people with disabilities and are thus illegal in Australia (Shi, 2005). ADA Australia is more like ADA US. They both existed before emergence of the Web technology, which they later integrated into the scope covered.

National Institute on Ageing (NIA) Guidelines for making senior friendly Websites: NIA guidelines were developed by the National Institute on Aging (NIA) in conjunction with the National Library of Medicine (NLM) to improve the usability of Web pages for older adults (Becker, 2004). These guidelines provide for the effective design of a Web page by taking into account font sizes, font types, colors, and styles; background images and colors; vertical scrolling; and text formats, among other design issues in order to make them accessible to people with ageing vision. However, the NIA guidelines only cover enhancing ccessibility for low vision Web users, a condition commonly suffered by ageing adults. Largely its recommendations are covered in the Web Content Accessibility Guidelines, although not as explicitly as in the NIA. Unlike other guidelines, it does not make reference to WCAG.

Accessibility Integration Model (WAIM): WAIM was created by Lazar and colleagues (Lazar et al., 2004) to help understand the problem of Web accessibility. It highlights the various influences on the accessibility or inaccessibility of a website. The purpose of the model is to guide researchers to investigate all the different angles of accessibility and to learn how to make sites more accessible. The Web Accessibility Integration Model has three categories of influences on Web accessibility namely: societal foundations, stakeholder perceptions, and Web development.

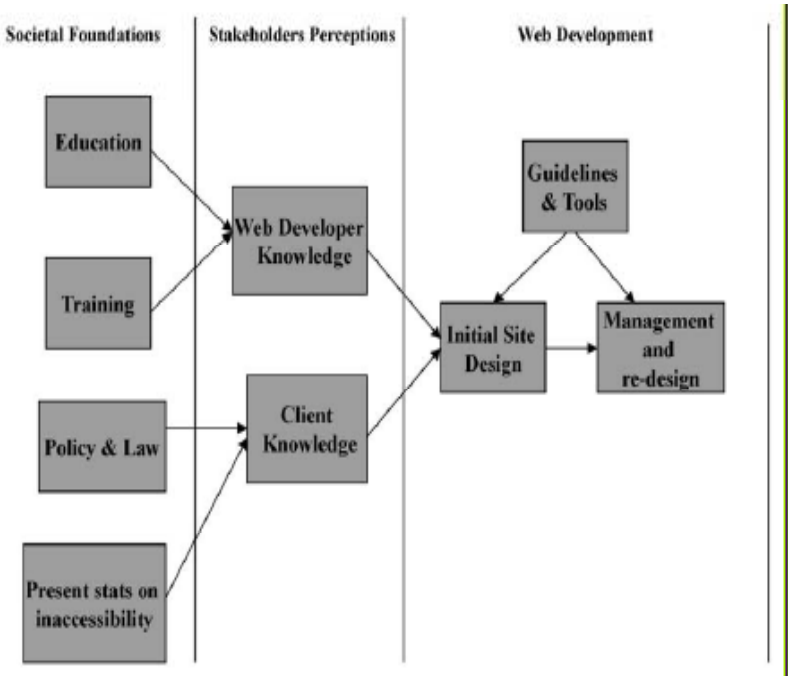
Societal foundations is concerned with how much web accessibility is valued in a particular society such as if accessibility is part of any national curriculum in Computer Science (CS), availability of training in accessibility and laws or policies that mandate Web accessibility. Stakeholders on the other hand are the people who decide whether a site will be built for accessibility or not. These include Web developers and the clients. Their perceptions are influenced by societal foundations such as education, government policy and statistics in the public media. The Web development part concerns the available guidelines and tools. Guidelines and tools such as WCAG help not only Web developers and webmasters with guidance, but

also help provide the current ''working definition'' for web accessibility. Good, well-written guidelines, and powerful software tools are likely to help improve levels of accessibility. Poorly written, confusing guidelines, and hard to use or unclear software tools are likely to keep sites from becoming accessible Lazar et al., 2004).

In Uganda, the influence of societal foundations towards making websites accessible is still lacking. Web accessibility is not part of any computing curriculum and there are no relevant government polices. The perceptions of stakeholders are not known. Although Web accessibility guidelines and tools such as WCAG and Bobby are available in the public domain, in Uganda, it is not known how much webmasters know about such guidelines and tools, if they are using them and why they are using them or why they are not using them.

Figure 1 presents a graphical representation of the three influences, the components within each influence and how the influences are related to each other.

Fig 1: Web Accessibility Integration Model adapted from Lazar et al., (2004)



Related work

Lazar and colleagues (lazar et al., 2004) in a study on the perceptions of webmasters found that 66% of webmasters had ever created accessible websites, 56% had their current websites accessible to people with visual disabilities, 27% had never created

an accessible website, and 0.5% were not sure. However, the geographical scope of this study was general and mainly covered webmasters who were known or in the reach of the authors. 45% of the webmasters were from USA, 24% from other countries and 41% did not indicate their countries of origin.

Other studies on Web accessibility have been focused on the state of accessibility of e-applications such as e-tourism, e-government, e-commerce and e-education (shi, 2005; Abanumy, 2005). So far, no such a study has been carried out on a specific country, later on Uganda. In this study, we examine the knowledge, Web design practices and perceptions of webmasters in Uganda in respect to Web accessibility. To achieve this, a survey was conducted on 30 webmasters of government agency websites. Results of the survey are presented in the next section.

Methodology

The objective of the study was to learn more about why webmasters in Uganda do not make their websites accessible and what can be done to improve the situation. To realise the set objective, a survey was created, with questions asking webmasters of government agency websites, their knowledge and perceptions of Web accessibility. After development, the survey was pre-tested for clarity, and posted to known webmasters' email addresses and potential mailing lists. The mailing lists chosen had either a national scope or wider scope but still with a substantial number of local subscribers that include webmasters. Webmasters that were known to the research team were also invited to participate. Some questionnaires were physically delivered to potential respondents and physically picked. The mailing lists used included I-Network (i-network@dgroups.org) (http://www.i-network.or.ug/) and Women of Uganda Network (WOUGNET) (wougnet-l@wougnet.org) (http://www.wougnet.org/). Guidelines for good Web survey usability were followed (Eysenbach and Wyatt, 2006). The survey comprised both close ended and open ended questions and a copy of the survey questionnaire is included as appendix A. The next section presents results of the survey.

Findings

This section presents results of the survey.

Participation: The survey questionnaire was administered to 30 webmasters of government agency websites. Out of the 30, 15 responded including 8 from ministries and 7 from parastatals. From each agency, one webmaster participated.

Table 1: Participation

Nature of Government Agency	Number	Percentage
Ministries	8	53%
Parastatals	7	47%

The first eleven questions were close ended and focused on current and future website accessibility and webmaster knowledge and experience with various software tools. The following subsections present results that were obtained:

Expertise of Webmasters: Webmasters were requested to indicate their level of expertise on which 73% regarded themselves as experts and the rest as intermediate. This could be attributed to the fact that staff in charge of administering websites are usually members of the IT department with formal IT training. Details are presented in table 2.

Table 2: Expertise of Webmasters

Level of IT Expertise	Number	Percentage
Expert	11	73%
Intermediate	4	27%
Novice	0	0%
Not sure	0	0%

Creation of accessible websites: The survey sought if a webmaster had ever created a website that was accessible to users with disabilities particularly visual impairments. 93% of the webmasters had never created such a website and 7% were not sure. Hence none of the webmasters surveyed had ever created an accessible website. Details of the responses are presented in table 3.

Table 3 Creation of accessible websites

Response	Number	Percentage
Yes	0	0%
No	14	93%
Not sure	1	7%

Familiarity with Web Accessibility Guidelines: The webmasters were asked if they were familiar with any of the accessibility guidelines by the Web Accessibility Initiative. Majority (67%) of the webmasters were not familiar with all the guidelines. Only 33% were familiar with the Web Content Accessibility Guidelines (WCAG). The high percentage of webmasters without knowledge of existing web accessibility guidelines could explain why all the webmasters had never created an accessible website. However given that even the few (33%) that are familiar with WCAG had never created an accessible website could mean that availability of the guidelines alone is not enough but rather a combination of many factors. Details of the responses are given in table 4.

Table 4 Familiarity with Web Accessibility Guidelines

Guideline	Number	Percentage
Web Content Accessibility Guidelines	5	33%
Authoring Tool Accessibility Guidelines	0	0%
User Agent Accessibility Guidelines	0	0%
Not familiar with any	8	67%

Familiarity with Web Accessibility Laws: In addition to the guidelines, webmasters were asked if they were familiar with any Web accessibility laws similar to the Web Content Accessibility Guidelines from other bodies or governments around the world. All the webmasters were not familiar with any laws on Web accessibility. Hence webmasters' knowledge of the global requirements on Web accessibility is very limited which is a negative influence on their perceptions.

Accessibility of websites: On whether the websites that webmasters were currently overseeing were accessible to users with visual impairments, all the webmasters indicated that their current websites were not accessible. This correlates to the low levels of awareness about existing Web accessibility guidelines and related laws.

Awareness about accessibility checking tools: They were further asked if they were aware that there are software tools that can check a website for its accessibility to people with visual impairments, and provide useful feedback. Only one webmaster was familiar with the availability of such software tools. This further explains why the government websites are not accessible. But given that the one webmaster aware of the tools had also not made his organisation's website accessible means that there are more reasons why government websites in Uganda are not accessible other than perceptions of webmasters alone.

Usage of free accessibility tools: The webmasters were also asked if they had ever used a free web-based accessibility tool, such as bobby. All the webmasters had never used any free accessibility tools. This conforms to the fact that none of the agencies surveyed had an accessible website.

Usage of screen readers to test websites: In addition, they were asked whether they had ever tested the accessibility of their websites using a screen reader. Only one had ever used a screen reader to test his website. Given that 100% of government websites are not accessible, the one webmaster that had ever used a screen reader possibly used it on another site other than the one he was currently managing or the testing could have had other motives other than making his agency's website accessible.

Plans to make websites accessible: It was also asked whether the webmasters' organisations had plans to make their websites accessible to users with visual impairments in the future. Majority (87%) indicated that their organizations had

no such plans, while 13% did not respond to this question. Results of this question confirm that Web accessibility is not yet perceived important for government websites in Uganda.

Because closed-ended questions could not reveal the complete story behind webmaster perceptions and actions, the second part of the survey was made open ended. The open ended questions sought elaborate views from webmasters about the challenges of making websites accessible to people with disabilities in Uganda, who they thought should be in charge of making websites accessible, what would influence them to make their websites accessible, whether they considered making their websites accessible during updates, and whether they considered ethics during planning and updating of their websites. The responses obtained provided a qualitative collection of webmaster perceptions and some overall trends when many responses were indicating similar ideas. These are given in the following sub sections.

Challenges of making websites accessible: This question sought webmasters' views on the challenges of making websites accessible for users with visual impairments in general and government websites in particular: The responses obtained included the following:

- The target audience for most websites are never clearly defined
- Uganda lacks relevant laws for enforcing Web accessibility. The existing laws are old and do not apply to the Web and Internet. In addition, the old laws are weak for example where as they mandate that buildings should be made accessible to people with disabilities, so far very few buildings have been built to meet this standard.
- Web designers are not aware of the existence of Web accessibility guidelines and tools e.g. WCAG and Bobby.
- There are no relevant policies and laws to educate and compel website owners to make their websites accessible.
- There are still a lot of challenges faced in the implementation of IT in organizations such as lack of enough funds, lack of skills among staff and the general population, high cost of acquisition and maintenance hence a number of organizations are yet to install adequate IT infrastructure and services.
- The country's IT sector is still in infant stages. The government is just beginning to realise the importance of a strong IT sector as evidenced by the recent creation of a ministry of ICT in June 2006.
- High levels of illiteracy: Uganda still has low levels of literacy. This is worse for people with disabilities where about 99% have not had a chance to access formal education

The above responses explain the results obtained earlier in the close-ended questions.

Who should be responsible for making a website accessible: Webmasters were asked who among a webmaster, a systems analyst, a programmer, a help desk manager and a disability compliance office should be responsible for making a website accessible and to give reasons for their choices. 47% felt webmasters, systems analysts and programmers should be responsible. The reason given was that accessibility needs to be considered at all stages of website development hence all people involved in the development process have a role to play. 33% perceived Web accessibility as a purely technical issue hence better managed by systems analysts. To them, systems analysts should be solely responsible for making websites accessible because they are the ones in charge of technical details. 20% were not sure. Part of the responses to this question are in line with the viewpoint of the web accessibility integration model proposed by Lazar and colleagues (See Lazar et al., 2004), while the other part confirmed the strong role of systems analysts/ engineers in making websites accessible but also the emphasis on systems analysts/ engineers as the only people responsible depicts the low level of understanding of the topic by this group of people.

Factors that would influence webmasters to make their websites accessible: The following responses were obtained:

- Laws requiring all government agencies to make their websites accessible
- Sensitization and awareness for government policy makers and IT staff about the importance of making websites accessible
- Training and sensitization for webmasters and Web developers on building and managing accessible websites
- The number of stakeholders with disabilities such as visual impairments
- Availability of funds to acquire relevant tools
- Knowledge and availability of easy to use tools for the design of such websites

Responses to this question are also in line with the influences of the web accessibility integration model (WAIM). Possibly the accessibility of government websites in Uganda could be improved by nurturing WAIM's influences.

Consideration of accessibility during website updates: All the webmasters reported that accessibility is never considered during their website updates. The reason for this behaviour is that agencies do not consider making their websites accessible a priority. This could be related to lack of efforts to change society perceptions about the value of making government websites accessible.

Consideration of ethics in planning and updating websites: Webmasters were asked to explain why or why not they considered ethics in planning and or updating websites. 47% acknowledged that they considered ethics in planning and updating their websites because ethics is very fundamental in the process of updating their websites. However they noted that they did not consider accessibility as part of the ethical aspects for website management. This can be attributed to the fact that there is still low value attached to making government websites accessible by the stakeholders.

Discussion of findings

The survey established that 100% of the government websites whose webmasters responded to the survey were not accessible to users with visual disabilities. However, a considerable level of webmasters (33%) were familiar with at least one Web accessibility guideline that is WCAG and 7% were aware about the availability of Web accessibility tools. This created a contradiction as to why 100% of the government websites are not accessible when 33% of their webmasters are aware about the availability of web accessibility guidelines. The inconsistency was later explained by the results of the open ended questions. From the responses to open-ended questions, it was established that Web accessibility is not yet perceived important for government websites in Uganda. So far, there is little if any efforts to change society perceptions about the value of making government websites accessible. Stakeholders attach low value to making government websites accessible to people with disabilities such as visual impairments. It was ascertained that this behavior is caused by the following factors: infant IT sector, low priority attached to Web accessibility, lack of awareness about the importance of Web accessibility, lack of training, lack of relevant policies, client ignorance, inadequate software tools, phobia for technical difficulties involved and high levels of illiteracy.

On who should be responsible for making websites accessible, webmasters, systems analyst/engineers and programmers were given as the central people but the role of systems analysts/engineers was emphasised as the pillar for making websites accessible. A number of possible actions that can improve the situation were proposed. The propositions include the following: enacting relevant legislation, sensitisation and training, publication of national disability statistics in the public media, funding and availability of easy to use tools.

Ultimately, the outcomes of the survey point to the need to cultivate the Ugandan environment with Web accessibility influences such as education, laws and training, changing stakeholder perceptions and availing Web accessibility development tools and guidelines. Efforts towards this are a worthwhile starting point towards making government websites accessible to people with disabilities such as visual impairments.

Conclusion and future work

Given that tools and guidelines are available to help in the development of websites that are accessible to people with disabilities, it is surprising that almost all websites in Uganda are still not accessible to them. This study was aimed at establishing why so many government websites in Uganda are not accessible to people with disabilities. Most webmasters surveyed supported the concept of web accessibility but cited roadblocks to achieving accessibility such as the infant IT sector, low priority attached to Web accessibility, lack of: awareness, training, and relevant policies, client ignorance, inadequate software tools, phobia for technical difficulties involved and high levels of illiteracy. While this study focused on webmasters (people who manage existing websites), in the future, web developers are another important group to study.

References

Abanumy, A., Al-Badi, A. and Mayhew, P. (2005). E-government website Accessibility: In-Depth Evaluation of Saudi Arabia and Oman. *The Electronic Journal of e-Government* Volume 3 Issue 3.

Becker S.A. (2004). E-government Visual Accessibility for Older Adults Users, *Social Science Computer review*, Vol. 22 No.1 Spring 2004 11-23.

Chiang M.F. et al., (2005). Computer and World Wide Web Accessibility by Visually Disabled Patients: Problems and Solutions, *Survey of Ophthalmology* Volume 50 Number 4 July – August 2005. Elsevier Ltd.

Eysenbach G and Wyatt J. (2006). Facilitating research (Chapter 6.3). *In: McKenzie B (ed.): Internet and Medicine* (3rd edition), pp. 211-225. Oxford University Press (in press).

Huang, C.J. (2003). Usability of E-Government Web Sites for People with Disabilities, *In Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, IEEE Computer Society, 2003.

Ivory, M., Mankoff, J., & Le, A. (2003). Using automated tools to improve web site usage by users with diverse abilities. *IT and Society*, 1(3), 195–236.

Jaeger P. T. (2006a). Assessing Section 508 compliance on federal e-government Web sites: A multi-method, user-centered evaluation of accessibility for persons with disabilities, *Government Information Quarterly* 23 (2006) 169–190, science Direct.

Jaeger P.T. (2006b). Beyond Section 508: The spectrum of Legal Requirements for Accessible E-government Web sites in the United States, *Journal of e-government Information* (2004) 518–533.

Lazar, J., Dudley-Sponaugle, A. and Greenidge, K.D. (2004). Improving Web accessibility: A study of Webmaster perceptions. *Computers in Human Behavior*, 20, 269–288.

Nielsen, J. (2001). Beyond Accessibility: Treating Users with Disabilities as People. *Jakob Nielsen's Alertbox* November 11, 2001. Accessed on 5th November 2006 from: http// www.useit.com /alertbox/20011111.html.

Shi, Y., (2005a). The accessibility of Queensland visitor information centres' websites, *Science Direct*, Elsevier Ltd

Shi, Y. (2006b). E-Government Web site Accessibility in Australia and China: Longitudinal Study, *Social Science Computer Review* Volume 24 Number 3, Sage Publications.

Takagi H. et al. (2005). Accessibility Designer: Visualizing Usability for the Blind. *ASSETS'04* October 18-20, 2004 Atlanta, Georgia, USA, ACM 2004.

Appendix a: survey questionaire for webmasters of government Agency Websites in Uganda By: Rehema Baguma, Faculty Of Computing & It, Makererere University.

(This survey is targeted at webmasters of government agency web sites to establish the state of accessibility of government Web sites to Web users with visual impairments in Uganda. It's part of broader research on E-government Web design framework for improving accessibility for users with disabilities).

Webmasters of government agency web sites, please forward filled questionnaire & any related questions to rbaguma@cit.mak.ac.ug.

Questions

1. Name of employing entity
2. How would you classify your computing experience?
   a. Expert
   b. Intermediate
   c. Novice
   d. Not Sure
3. Have you ever created a website that is accessible for users with visual impairments (blind, low vision and color blind?
   a. Yes
   b. No
   c. Not Sure
4. Are you familiar with any of the following accessibility guidelines from the Web Accessibility

   Initiative of the World Wide Web Consortium (W3C)? (Check all that apply):
   a. Web Content Accessibility Guidelines
   b. Authoring Tool Accessibility Guidelines
   c. User Agent Accessibility Guidelines
   d. Not familiar with any accessibility guidelines
5. Are you familiar with any laws similar to the Web Content Accessibility Guidelines from other bodies or governments around the world?
   a. Yes (if yes specify.....)
   b. No
   c. Not Sure
6. Is the website that you are currently overseeing accessible to users with visual impairments
   a. Yes
   b. No
   c. Not Sure
7. Are you aware that there are software tools that can check your website to see if it is accessible to people with visual impairments, and provide useful feedback?
   a. Yes
   b. No
   c. Not Sure

8. Have you ever used a free web-based accessibility tool, e.g., Bobby?
   a. Yes
   b. No
   c. Not Sure
9. Have you ever used a non-web-based accessibility tool, e.g., A-Prompt, INFOCUS, Page Screamer?
   a. Yes
   b. No
   c. Not Sure
10. Have you ever tested your website using a screen reader? (A screen reader reads the text out loud in computer-synthesized speech.)
    a. Yes
    b. No
    c. Not Sure
11. Does your organization have any plans to make your website accessible to users with visual impairments in the future?
    a. Yes
    b. No
    c. Not Sure
12. What do you think are the challenges of making a web site accessible for users with visual impairments? (Explain your answer).
    .....
    .....
    .....
    .....
    .....
13. What do you think are the challenges of making government web sites accessible for users with visual impairments in Uganda (Explain your answer).
    .....
    .....
    .....
    .....
14. Who do you think should be responsible for making a website accessible for users with visual impairments? (Check all that apply.)
    a. Webmaster
    b. Systems Analyst/Engineer
    c. Programmer
    d. Help Desk Manager

e. Disability Compliance Office

Explain your answer

.....
.....
.....
.....
.....

15. What factors would influence you to make your current Web site accessible for users with visual impairments (skip if web site already accessible)?

.....
.....
.....
.....
.....

16. When you make updates to your website, do you consider the factor of making the site accessible to all users? If yes explain how?

.....
.....
.....
.....

17. Do you consider ethics in planning and/or updating your current websites? Why or Why not?

.....
.....
.....
.....
.....

16
Knowledge Management Technologies and Organizational Business Processes: Integration for Business Delivery Performance in Sub Saharan Africa

Asifiwe Collins Gyavira Rubanju

To remain competitive, organizations must efficiently and effectively create, locate, capture, and share their organization's knowledge and expertise. This increasingly requires making the organization's knowledge explicit and recording it for easier distribution and reuse. The paper is based mainly on theory, literature review and borrows some practical examples of knowledge management theories in trying to understand business performance. Knowledge management together with mindful application of the business technology can be of extreme importance for performance in organizations. Human capital with effective use of the available technologies could play a more important role in growth and performance of any organization. However, questions of what technology to use and the process of its use are paramount in understanding its application in businesses. Note should be taken that massive investment in sophisticated technology is not a prerequisite to organizational performance.

Introduction

"People bring imagination and life to a transforming technology." -- *Business Week*, The Internet Age (Special Report), October 4, 1999, p. 108).

The concept of knowledge management is not new in information systems practice and research. However, with the increased complexity in the business environment, there is need to provide business executives and scholars with pragmatic understanding about integrating knowledge management strategy and technologies in business processes for successful performance.

There is not yet a common consensus on the concept of knowledge management; however, the shared theme is that increasingly, knowledge in the minds of organizational members is of greatest value as the organizational resource. *Knowledge Managemen*t comprises a range of practices used by organisations to identify, create, represent, and distribute *knowledge* for reuse, awareness and learning. Various governments and Organizations are showing a tremendous interest in implementing knowledge management processes and technologies, and are even beginning to adopt knowledge management as part of their overall business strategy in Africa. The new branch of management (KM) is important for achieving breakthrough business performance through the synergy of people,

processes, and technology; with its focus on management of change, uncertainty, and complexity. Although knowledge management is becoming widely accepted, few organizations today are fully capable of developing and leveraging critical organizational knowledge to improve their performance (Heibeler, 1996). There is a general agreement on the importance of organizational wealth of knowledge. As we transition from an era of information scarcity to information superfluity, there is need for re-focusing on human sense making processes underlying decisions, choices, and performance. In this new paradigm for increasingly uncertain and complex business environments, effective performance and outcomes must be a result of how human capital and technology are integrated in a business enterprise to leverage strategic opportunities and challenges.

Stopped: In the recent past, Knowledge Management (KM) has evolved into a mature reality from what was merely a plague on the "good idea" radar only a few years ago. Growing pervasiveness of KM in worldwide industries, organizations, and institutions marks a turning point event for what was called a fashion just a few years ago. KM has become embedded in the policy, strategy, and implementation processes of worldwide corporations, governments, and institutions. Doubling in size from 2001, the global KM market has been projected to reach US$8.8 billion in a few years to come. Likewise, (Malhotra, 2004a) notes that the market for KM business application capabilities is expected to grow to $148 billion by the next year. The broader application context of KM, which includes learning, education, and training industries, offers similarly sanguine forecasts.

One can see the impact of knowledge management everywhere but in the KM technology-performance statistics (Malhotra, 2003). This seems like a contradiction of sorts given the pervasive role of information and communication technologies in most KM applications. Some industry estimates have pegged the failure rate of technology implementations for business process reengineering efforts at 70 percent. Recent industry data suggest a similar failure rate of KM related technology implementations and related applications (Darrell et al., 2002). Significant failure rates persist despite tremendous improvements in sophistication of technologies and major gains in related price-performance ratios. These *failures* are as a result of performance problems that have been attributed to technology implementation being too costly and slow. Interestingly, just a few months ago, some research studies had found negative correlation between technological investments and business performance (Alinean, 2002; Hoffman, 2002).

Strassmann (1997), based upon multi-year macroeconomic analysis of hundreds of corporations, had emphasized that it is not computers but what people do with them that matters. He had further emphasized the role of users' motivation and commitment in IT performance. Relatively recent research on implementation of enterprise level KMS (Malhotra, 1998a; Malhotra and Galletta, 1999; Malhotra and Galletta, 2003; Malhotra and Galletta, n.d. a; Malhotra and Galletta, n.d. b) has found empirical support for such socio-psychological factors in determining IT and KMS performance. An earlier study by Forrester Research had similarly

determined that the top-performing companies in terms of revenue, return on assets, and cash-flow growth spend less on IT on average than other companies.

Started Knowledge Management: The Information Processing Paradigm; The information-processing view of knowledge management has been prevalent in information systems practice and research over the last few decades. This perspective originated in the era when business environment was less vacillating, the products and services and the corresponding core competencies had a long multi-year shelf life, and the organizational and industry boundaries were clearly demarcated over the foreseeable future. The relatively structured and predictable business and competitive environment rewarded firms' focus on economies of scale. Such economies of scale were often based on high level of efficiencies of scale in absence of impending threat of rapid obsolescence of product and service definitions as well as demarcations of existing organizational and industry boundaries.

For a period of time, the information-processing paradigm has been prevalent over three phases that include 1) Automation: increased efficiency of operations; 2) Rationalization of procedures: streamlining of procedures and eliminating obvious bottlenecks that are revealed by automation for enhanced efficiency of operations; and, 3). Re-engineering: radical redesign of business processes that depends upon information technology intensive radical redesign of workflows and work processes.

These have been characterized by technology intensive, optimization-driven, efficiency-seeking organizational change (Malhotra 1999c, 1999d, in press). The deployment of information technologies in all the three phases was based on a relatively predictable view of products and services as well as contributory organizational and industrial structures.

Based on the convergence-oriented view of information systems, the information processing view of knowledge management is often characterized by benchmarking and transfer of best practices (cf: Allee 1997, O'Dell and Grayson 1998). The key assumptions of the information-processing view are often based on the premise of the generalizability of issues across temporal and contextual frames of diverse organizations. Such interpretations have often assumed that adaptive functioning of the organization can be based on explicit knowledge of individuals archived in corporate databases and technology-based knowledge repositories (cf: Applegate et al., 1988):

> *"Information systems will maintain the corporate history, experience and expertise that long-term employees now hold. The information systems themselves -- not the people – can become the stable structure of the organization. People will be free to come and go, but the value of their experience will be incorporated in the systems that help them and their successors run the business."*

There is a simplistic storage of knowledge in Africa, a factor of the past, with simplistic assumptions on which individuals and firms have based themselves to form a rule-of –thumb and best practices for guiding their future actions.

As Malhotra (2006) stated, some of the recent past simplistic assumptions with their interpretations of knowledge management include but not limited to the following in table 1.

Table 1. Knowledge Management: The Information Processing Paradigm

The process of collecting organizing classifying and disseminating information throughout an organization so as to make it purposeful to those who need it. (*Midrange Systems*: Albert 1998)
Knowledge management IT concerns organizing and analyzing information in a company's computer databases so this knowledge can be readily shared throughout a company instead of languishing in the department where it was created inaccessible to other employees. (*CPA Journal* 1998)
Identification of categories of knowledge needed to support the overall business strategy assessment of current state of the firm's knowledge and transformation of the current knowledge base into a new and more powerful knowledge base by filling knowledge gaps. (*Computerworld*: Gopal & Gagnon 1995)
Knowledge management in general tries to organize and make available important know-how wherever and whenever it's needed. This includes processes procedures patents reference works formulas "best practices" forecasts and fixes. Technologically intranets groupware data warehouses networks bulletin boards videoconferencing are key tools for storing and distributing this intelligence. (*Computerworld*: Maglitta 1996)
Mapping knowledge and information resources both on-line and off-line; Training guiding and equipping users with knowledge access tools; Monitoring outside news and information. (*Computerworld*: Maglitta 1995)

In the above stated notion of knowledge management, there is much emphasis on machinery than the way people in organizations acquire, share and create new knowledge for the benefit of the organization. There is by far, considering the meaning of knowledge as "unproblematic, predefined, and prepackaged" (Boland 1987), less focus on human dimension of organizational knowledge creation in this whole notion of knowledge management.

There are primary contexts in which knowledge management will have on the organization's performance. One of these is Technology context (Zack, 1999), which addresses the existing information technology infrastructure and capabilities supporting the knowledge management architecture. While the adage is that knowledge management is 10% technology and 90% people, without the ability to collect, index, store, and distribute explicit knowledge electronically

and seamlessly to where needed when needed, the organizational capabilities and incentives will not be fully exploited.

Sophistication of KM technology implementations and its failures. The important question here is whether more sophisticated technologies often deliver. However, (Zack, 1999) states that the technology need not be complex or leading-edge to provide significant benefit. Its absence, however, would have prevented both from effectively managing their knowledge. It is true that Knowledge management promises much, but often delivers very little. Knowledge management is not simply a matter of installing new software or changing a small aspect of the business. KM is about knowledge sharing amongst people in the organization, technologies that help the sharing and mechanisms for bringing new knowledge into the organization (Birkinshaw, 2001).

In addition, Malhotra (2004b) notes that despite increasing sophistication of KM technologies, we are observing increasing failures of KM technology implementations. It is important to note that such failures result from the knowledge gaps between technology inputs, knowledge processes, and business performance. Malhotra (2004b) also highlights evidence that business organizations and companies spend less on technology and are not leaders in adoption of most hyped Real Time Enterprises (RTE) technologies succeed where others fail. The RTE enterprise which is considered the epitome of the agile adaptive and responsive enterprise capable of anticipating surprise; hence the attempt to reconcile its sense making and information processing capabilities is all the more interesting. However, the theoretical generalizations and their practical implications are relevant to IT and KM systems in most enterprises traversing through changing business environments.

Management and coordination of diverse technology architectures, data architectures, and system architectures poses obvious knowledge management challenges (Malhotra, 1996; Malhotra, 2001a; Malhotra, 2004b). Such challenges result from the need for integrating diverse technologies, computer programs, and data sources across internal business processes. These challenges are compounded manifold by the concurrent need for simultaneously adapting enterprise architectures to keep up with changes in the external business environment. For this to happen, changes in the existing technologies or their replacement with newer technologies must be done. Growing business enterprises often have too much (unprocessed) data and (processed) information and too many technologies. However, for most high-risk and high-return strategic decisions, timely information is often unavailable as more and more of such information is external in nature (Drucker, 1994; Malhotra, 1993; Terreberry, 1968; Emery and Trist, 1965).

As a result, most organizations have incomplete knowledge of explicit and tacit data, information, and decision models available within the enterprise. In other words, often they may not know if the available data, information, and decision models are indeed up to speed with the radical discontinuous changes in the business environment (Arthur, 1996; Malhotra, 2000a; Nadler and Shaw, 1995). Therefore,

incomplete and often outdated data, information, and decision models drive the realization of the strategic execution, but with diminishing effectiveness.

The mechanistic information-processing orientation of the technology push-model of KM generally does not encourage diverse interpretations of information or possibility of multiple responses to same information. This model has served the needs of business performance given more manageable volumes of information and lesser variety of systems within relatively certain business environment. However, with recent unprecedented growth in volumes of data and information in almost all sizes of organizations, the continuously evolving variety of technology architectures, and the radically changing business environment, this model has outlasted its utility. The limitations of the technology-push model are evident in the following depiction of IT architectures as described in Information Week by LeClaire and Cooper (2000):

> *The infrastructure issue is affecting all businesses . . . E-business is forcing companies to rearchitect all or part of their IT infrastructures – and to do it quickly. For better or worse, the classic timeline of total business-process reengineering – where consultants are brought in, models are drawn up, and plans are implemented gradually over months or years – just isn't fast enough to give companies the e-commerce-ready IT infrastructures they need . . . Many companies can't afford to go back to the drawing board and completely rearchitect critical systems such as order fulfillment and product databases from the bottom up because they greatly depend on existing infrastructure. More often, business-process reengineering is done reactively. Beyond its disruptive effect on business operations, most IT managers and executives don't feel there's enough time to take a holistic approach to the problem, so they attack tactical issues one-by-one. Many companies tackle a specific problem with a definitive solution rather than completely overhaul the workflow that spans from a customer query to online catalogs to order processing.*

Does Information technology substitute human interaction? The Internet revolution has caused some writers to make absurd claims about how the world of work will change. One popular article by Tom Malone and Robert Laubacher (1998) foresaw the emergence of the "e-lance economy" in which individuals would work as freelancers rather than members of firms. Another line of thinking talked about the "paperless office". These arguments are not plain right for the reason that people need social interaction- both for its own sake and because it provides a powerful vehicle for learning and sharing information and knowledge. The Social Life of Information (Brown and Duguid 2000) provides an excellent counterpoint to the argument that technology is going to change the way we work. It explains the importance of the social interaction between people that lies at the heart of knowledge management.

This simple insight has important implications for the management of knowledge. First, it helps to explain why most knowledge databases are so poorly used. And secondary, it cautions us that IT tools and "social" tools such as communities of practice are complementary. A recent article by Morten Hansen and colleagues (2000) suggested that firms should focus on either a "codification strategy" which involves putting the firms' knowledge unto IT databases, or on a "personalization strategy" which involves building strong social networks.

What drives performance in business? Knowledge Management programs are typically tied to organisational objectives and are intended to achieve specific outcomes, such as shared intelligence, improved performance, competitive advantage, or higher levels of innovation. The gap between IT and business performance has grown with the shifting focus of business technology strategists and executives. Over the past two decades, their emphasis has shifted from IT (Porter and Millar, 1985; Hammer 1990) to information (Evans and Wurster, 2002; Rayport and Sviokla, 1995; Hopper, 1990; Huber, 1993; Malhotra, 1995) to knowledge (Holsapple and Singh, 2001; Holsapple, 2002; Koenig and Srikantaiah, 2000a; Malhotra, 2004b; Malhotra, 2000b; Malhotra, 1998c) as the lever of competitive advantage. Many industry executives and most analysts have incorrectly presumed or pitched technology as the primary enabler of business performance (Collins, 2001; Schrage, 2002).

The findings from the research on best performing companies over a given period of time (Collins, 2001) presented in terms of the inputs-processing-outcomes framework used for contrasting the technology-push model with the strategy-pull model of KM implementation. Given latest advances in web services, the strategic framework of KM discussed here presents a viable alternative for delivering business performance as well as enterprise agility and adaptability (Strassmann, 2003). These findings, according to Collins 2001, were presented as Lessons learned from some of the most successful business enterprises that distinguished themselves by making the leap from ''good to great'' (Collins, 2001)

The technology-push model of Knowledge Management embraces the following; input driven paradigm of KM (IT-based systems developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application'' (Alavi and Leidner, 2001), process driven paradigm of KM (helping people share and put knowledge into action by creating access, context, infrastructure, and simultaneously reducing learning cycles'' (Massey et al., 2001) and outcomes driven paradigm of KM.

Some Lessons about outcomes paradigm: strategic execution, the primary enabler

(1) How a company reacts to technological change is a good indicator of its inner drive for greatness versus mediocrity. Great companies respond with thoughtfulness and creativity, driven by a compulsion to turn unrealized potential into results; mediocre companies react and lurch about, motivated by fear of being left behind

(2) Any decision about technology needs to fit directly with three key non-technological questions: What are you deeply passionate about? What can you be the best in the world at? What drives your economic engine? If a technology does not fit squarely within the execution of these three core business issues, the good-to-great companies ignore all hype and fear and just go about their business with a remarkable degree of equanimity

(3) The good-to-great companies understood that doing what you are good at will only make you good; focusing solely on what you can potentially do better than any other organization is the only path to greatness

Lessons about processing paradigm: how strategic execution drives technology utilization

(1) Thoughtless reliance on technology is a liability, not an asset. When used right – when linked to a simple, clear, and coherent concept rooted in deep understanding – technology is an essential driver in accelerating forward momentum. But when used wrongly – when grasped as an easy solution, without deep understanding of how it links to a clear and coherent concept – technology simply accelerates your own self-created demise

(2) No evidence was found that good-to-great companies had more or better information than the comparison companies. In fact both sets of companies had identical access to good information. The key, then, lies not in better information, but in turning information into information that cannot be ignored

(3) 80 percent of the good-to-great executives did not even mention technology as one of the top five factors in their transition from good-to-great. Certainly not because they ignored technology: they were technologically sophisticated and vastly superior to their comparisons

(4) A number of the good-to-great companies received extensive media coverage and awards for their pioneering use of technology. Yet the executives hardly talked about technology. It is as if the media articles and the executives were discussing two totally different sets of companies!

Lessons about technology inputs: how strategic execution drives technology deployment

(1) Technology-induced change is nothing new. The real question is not What is the role of technology? Rather, the real question is How do good-to-great organizations think differently about technology?

(2) It was never technology per se, but the pioneering application of carefully selected technologies. Every good-to-great company became a pioneer in the application of technology, but the technologies themselves varied greatly

(3) When used right, technology becomes an accelerator of momentum, not a creator of it. The good-to-great companies never began their transitions with

pioneering technology, for the simple reason that you cannot make good use of technology until you know which technologies are relevant

(4) You could have taken the exact same leading-edge technologies pioneered at the good-to-great companies and handed them to their direct comparisons for free, and the comparisons still would have failed to produce anywhere near the same results

Can knowledge management work? The technology evangelists, criticized by Stewart (2000), have endowed the KM technologies with intrinsic and infallible capability of getting the right information to the right person at the right time. Similar critiques (cf. Malhotra, 2000a; Hildebrand, 1999) have further unraveled and explained the ''myths'' associated such proclamations made by the technology evangelists. Specifically, it has been underscored that in wicked business environments (Churchman, 1971; Malhotra, 1997) characterized by radical discontinuous change (Malhotra, 2000a; Malhotra, 2002b), the deterministic and reductionist logic (Odom and Starns, 2003) of the evangelists does not hold. Incidentally, most high potential business opportunities and threats are often embedded within such environments (Arthur, 1996; Malhotra, 2000c; Malhotra, 2000d). Such environments are characterized by fundamental and ongoing changes in technologies as well as the strategic composition of market forces. What brings any increase in failure rates of KM technologies is often the rapid obsolescence given changing business needs and technology architectures. Skeptics of technology have observed that real knowledge is created and applied in the processes of socialization, externalization, combination, and internalization (Nonaka and Takeuchi, 1995) and outside the realm of KM technologies. Scholarly research on latest information systems and technologies, or lack thereof, has further contributed to the confusion between data management, information management, and knowledge management.

Hence, it is critical that a robust distinction between technology management and knowledge management should be based on theoretical arguments that have been tested empirically in the ''real world messes'' (Ackoff, 1979) and the ''world of re-everything'' (Arthur, 1996). We are observing diminishing credibility of information technologists (Anthes and Hoffman, 2003; Hoffman, 2003; Carr, 2003). A key reason for this is an urgent need for understanding how technologies, people, and processes together influence business performance (Murphy, 2003). Explicit focus on strategic execution as the driver of technology configurations in the strategy-pull KM framework reconciles many of the above problems. The evolving paradigm of technology architectures to on demand plug-and-play inter-enterprise business process networks (Levitt, 2001) is expected to facilitate future realization of KM value networks. Growing popularity of the web services architecture (based upon XML, UDDI, SOAP, WSDL) is expected to support the realization of real-time deployment of business performance driven systems based upon the proposed model (Kirkpatrick, 2003; Zetie, 2003; Murphy, 2003).

The technology-push model is attributable for the inputs – and processing – driven KM implementations with emphasis on pushing data, information,

and decisions. In contrast, the strategy-pull model recognizes that getting pre-programmed information to pre-determined persons at the pre-specified time may not by itself ensure business performance. Even if pre-programmed information does not become out-dated, the recipient's attention and engagement with that information is at least equally important. Equally important is the reflective capability of the recipient to determine if novel interpretation of the information is necessary or if consideration of novel responses is in order given external changes in the business environment. The technology-push model relies upon single-loop automated and unquestioned automatic and pre-programmed response to received stimulus. In contrast, the strategy-pull model has built in double-loop process that can enable a true sense-and-respond paradigm of KM. The focus of the technology-push model is on mechanistic information processing while the strategy-pull model facilitates organic sense making (Malhotra, 2001b). The distinctive models of knowledge management have been embedded in KM implementations of most organizations since KM became fashionable.

Therefore, what is actual enabler of business enterprise? The issues of technology deployment, technology utilization, and business performance need to be addressed together to ensure that technology can deliver upon the promise of business performance. Interestingly, most implementations of KM systems motivated by the technology-push model have inadvertently treated business performance as a residual: what remains after issues of technology deployment and utilization are addressed. This perhaps explains the current malaise of IT executives and IT management in not being able to connect with business performance needs (Hoffman, 2003).

Deployment of intranets, extranets, or, groupware cannot of itself deliver business performance. These technologies would need to be adopted and appropriated by the human users, integrated within their respective work-contexts, and effectively utilized while being driven by the performance outcomes of the enterprise. To deliver real-time response, business performance would need to drive the information needs and technology deployment needs. This is in congruence with the knowledge management logic of the top performing companies discussed earlier. These enterprises may not have created the buzz about the latest technologies. However, it is unquestionable that these best performing organizations harnessed organizational and inter-organizational knowledge embedded in business processes most effectively to deliver top-of-the-line results. The old model of technology deployment spanning months or often years often resulted in increasing misalignment with changing business needs. Interestingly, the proposed model turns the technology-push model on its head. The strategy-pull model treats business performance not as the residual but as the prime driver of information utilization as well as IT-deployment.

As noted before, there are three paradigms of KM which seem to be in contrast at a point. The inputs-driven paradigm considers information technology and KM as synonymous. The inputs-driven paradigm with its primary focuses on

technologies such as digital repositories, databases, intranets, and, groupware systems as been the mainstay of many KM implementation projects. Specific choices of technologies drive the KM equation with primary emphasis on getting the right information technologies in place. However, the availability of such technologies does not ensure that they positively influence business performance. For instance, installing a collaborative community platform may neither result in collaboration nor community (Barth, 2000; Charles, 2002; Verton, 2002)

The processing-driven paradigm of KM has its focus on best practices, training and learning programs, cultural change, collaboration, and virtual organizations. This paradigm considers KM primarily as means of processing information for various business activities. Most proponents of RTE belong to this paradigm given their credo of getting the right information to the right person at the right time. Specific focus is on the activities associated with information processing such as process redesign, workflow optimization, or automation of manual processes. Emphasis on processes ensures that relevant technologies are adopted and possibly utilized in service of the processes. However, technology is often depicted as an easy solution to achieve some type of information processing with tenuous if any link to strategic execution needed for business performance. Implementation failures and cost-and-time overruns that characterize many large-scale technology projects are directly attributable to this paradigm (Anthes and Hoffman, 2003; Strassmann, 2003). Often the missing link between technologies and business performance is attributable to choice of technologies intended to fix broken processes, business models, or organizational cultures.

The outcomes-driven paradigm of KM has its primary focus on business performance. Key emphasis is on strategic execution for driving selection and adaptation of processes and activities, and carefully selected technologies. For instance, if collaborative community activities do not contribute to the key customer value propositions or business value propositions of the enterprise, such activities are replaced with others that are more directly relevant to business performance (Malhotra, 2002a). If these activities are indeed relevant to business performance, then appropriate business models, processes, and culture are grown (Brooks, 1987) as a precursor to acceleration of their performance with the aid of KM technologies. Accordingly, emphasis on business performance outcomes as the key driver ensures that relevant processes and activities, as well as, related technologies are adopted, modified, rejected, replaced, or enhanced in service of business performance.

As discussed earlier, success in strategic execution of a business process or business model may be accelerated with carefully chosen technologies. However, in absence of good business processes and business model, even the most sophisticated technologies cannot ensure corporate survival.

Why do some businesses succeed (while others do not)? From related literature review, specific companies were chosen based on their visibility in the business technology press and popular media. The reviews of industry cases studies were

guided by our interest in understanding the link between investments in advanced technologies and resulting business performance.

Wal-Mart: RTE business model where technology matters less: Wal-Mart has emerged as a company that has set the benchmark of doing more with less. Wal-Mart did not build its competitive advantage by investing heavily or by investing in latest technologies (Schrage, 2002). A McKinsey Global Institute reports:

The technology that went into what Wal-Mart did was not brand new and not especially at the technological frontiers, but when it was combined with the firm's managerial and organizational innovations, the impact was huge.

Wal-Mart systematically and rigorously deployed its technologies with clear focus on its core value proposition of lowest prices for mass consumers. With that singular focus, it went about setting up its supply chains and inventory management systems to accelerate business performance. This was facilitated by the Real Time Enterprise (RTE) based on the hub-and-spoke model of truck routes. The new business model created the strong linkages with suppliers, which not only heavily subsidized the costs of technology investments but also pre-committed the partners to the success of the shared systems.

Dell: RTE business model that does more with less

Dell has developed and perfected its business model by developing strong ties with its customer base over the past two decades. It perfected its business model over several years before accelerating its business performance with the aid of carefully selected technologies. It has cultivated outstanding relationships with its virtual supply chain partners including outsourcing providers (such as Solectron) and technology vendors (such as HP, Sony, and EMC). Dell also benefits from technologies developed by its technology partners. It has been developing and extending the real time logic over the past several years first for selling and servicing desktop computers, and later to aggregation and distribution of value-added products and services servers, storage, networking, printers, switches, and handheld computers. According to a survey of 7,500 companies conducted by Alinean (2002), Dell is an economic IT spender. Dell is equally frugal in its R&D spending (1.5 percent of revenues), according to a recent Business Week report, despite its continuing forays into new products and services. Therefore, lessons learned from the world's greatest organizations, as noted in part, show that even simple technologies can generate great performance when empowered by smart minds of motivated and committed humans.

Conclusion

Technologists never evangelize without a disclaimer: "Technology is just an enabler." True enough -- and the disclaimer discloses part of the problem: Enabling what? One flaw in knowledge management is that it often neglects to ask what knowledge to manage and toward what end. Knowledge management activities are all over the map: Building databases, measuring intellectual capital, establishing

corporate libraries, building intranets, sharing best practices, installing groupware, leading training programs, leading cultural change, fostering collaboration, creating virtual organizations -- all of these are knowledge management, and every functional and staff leader can lay claim to it. But no one claims the big question: Why? (Tom Stewart in The Case Against Knowledge Management, Business 2.0, February 2002).

Just as managing a business depends on deciding what business you are in so knowledge management must begin by selecting the knowledge to be managed. It's no good assembling a library full of everything anybody could conceivably want to know about everything. The important and fundamental questions that must be answered in KM efforts include; What is the work group?, What does the group need to know?, Are you a standardizer or a customizer? Production strategies for which you mostly know what knowledge you need -- and for which the tasks are mostly well understood, the processes mostly routine, and the problems mostly familiar -- lend themselves to a knowledge management strategy of codification, automation, and librarianship. However, note should be taken of the danger that technology can be an enabler in the same sense that an alcoholic's spouse can be one.

Andrew Michuda, the chief executive of Sopheon, which provides knowledge management software and manages a network of thousands of technical experts and analysts, perfectly describes how knowledge management goes wrong: "KM hits a wall when it is generically applied. You need the richness of human interaction with the efficiencies of technology, focused on a knowledge-intensive business application. Knowledge management is much more effective if it is not a stand-alone button on somebody's PC but is integrated into a key business process."

References

Ackoff, R. (1979), ''The future of operations research is past'', Journal of the Operations Research Society, Vol. 30, p. 93.

Alavi, M. and Leidner, D. (2001), ''Review: knowledge management and knowledge management systems: conceptual foundations and research issues'', MIS Quarterly, Vol. 25 No. 1, pp. 107-36.

Alinean (2002), ''Alinean identifies why certain companies achieve higher ROI from IT investments'', available at: www.alinean.com

Allee, V. "Chevron Maps Key Processes and Transfers Best Practices," Knowledge Inc., April 1997.

Arthur, W. B., "Increasing Returns and the New World of Business," Harvard Business Review, July-August 1996, 74(4). Anthes, G.H. "A Step Beyond a Database," Computerworld, 25(9), 1991, p. 28.

Anthes, G.H. and Hoffman, T. (2003), ''Tarnished image'', Computerworld, May 12, pp. 37-40.

Applegate, L., Cash, J. & Mills D.Q. "Information Technology and Tomorrow's Manager," In McGowan, W.G. (Ed.), Revolution in Real Time: Managing Information Technology in the 1990s, pp. 33-48, Boston, MA, Harvard Business School Press, 1988.

Barth, S. (2000), ''KM horror stories'', Knowledge Management, Vol. 3 No. 10, pp. 36-40.

Brooks, F.P. Jr (1987), ''No silver bullet: essence and accidents of software engineering'', Computer, Vol. 20 No. 4, pp. 10-19.

Brown, J.S. and Duguid, P. (2000). The Social Life of Information, Cambridge, MA: Harvard Business School Press.

Boland, R.J. "The In-formation of Information Systems," In R.J. Boland and R. Hirschheim (Eds.), Critical Issues in Information Systems Research, pp. 363-379, Wiley, Chichester, 1987.

Carr, N. (2003), ''IT doesn't matter'', Harvard Business Review, Vol. 81 No. 5, pp. 41-9.

Charles, S.K. (2002), ''Knowledge management lessons from the document trenches'', Online, Vol. 26 No. 1, pp. 22-9.

Churchman, C.W. The Design of Inquiring Systems, Basic Books, New York, NY, 1971.

Collins, J. (2001), Good to Great: Why Some Companies Make the Leap and Others Don't, Harper-Business, New York, NY.

Darrell, R., Reichheld, F.F. and Schefter, P. (2002), ''Avoid the four perils of CRM'', Harvard Business Review, February, pp. 101-9.

Drucker, P. F., 'The Theory of Business,' Harvard Business Review, September-October 1994, pp 95-104.

Julian Birkinshaw, "*Why is Knowledge Management So Difficult?*," *Business Strategy Review* (vol. 12, no. 1, 2001).

Emery, F.E. and Trist, E.L. (1965), ''The causal texture of organizational environments'', Human Relations, Vol. 18, pp. 21-32.

Evans, P. and Wurster, T.S. (2002), Blown to Bits, Harvard Business School Press, Boston, MA.

Gill, T.G. "High-Tech Hidebound: Case Studies of Information Technologies that Inhibited Organizational Learning," Accounting, Management and Information Technologies, 5(1), 1995, pp. 41-60.

Hildebrand, C. (1999), ''Intellectual capitalism: does KM ¼ IT?'', CIO Magazine, September 15, available at: www.cio.com/archive/enterprise/091599_ic_content.html

Hoffman, T. (2002), '''Frugal' IT investors top best-performer list'', Computerworld, December 6, available at: www.computerworld.com/managementtopics/roi/story/0,10801,76468,00.html

Hoffman, T. (2003), ''Survey points to continuing friction between business, IT'', Computerworld, May 12, p. 10.

Holsapple, C.W. (2002), ''Knowledge and its attributes'', in Holsapple, C.W. (Ed.), Handbook on Knowledge Management 1: Knowledge Matters, Springer-Verlag, Heidelberg, pp. 165-88.

Holsapple, C.W. and Singh, M. (2001), ''The knowledge chain model: activities for competitiveness'', Expert Systems with Applications, Vol. 20 No. 1, pp. 77-98.

Hopper, M.D. (1990), ''Rattling SABRE – new ways to compete on information'', Harvard Business Review, May/June, pp. 118-25.

Huber, R.L. (1993), ''How Continental Bank outsourced its 'crown jewels''', Harvard Business Review, January/February, pp. 121-9.

Kirkpatrick, T.A. (2003), ''Complexity: how to stave off chaos'', CIO Insight, February 1, available at: www.cioinsight.com/print_article/0,3668,a ¼ 37126,00.asp

Koenig, M.D. and Srikantaiah, T.K. (2000a), ''The evolution of knowledge management'', in Srikantaiah, K. and Koenig, M.E.D. (Eds), Knowledge Management for the Information Professional, Information Today Inc., Medford, NJ, pp. 37-61.

LeClaire, J. and Cooper, L. (2000), ''Rapid-Fire IT Infrastructure'', Information Week, January 31, available at: www.informationweek.com/771/infrastruct.htm

Levitt, J. (2001), ''Plug-and-play redefined'', Information Week, April 2, available at: www.informationweek.com/831/web.htm

Malhotra, Y. (in press). "From Information Management to Knowledge Management: Beyond the 'Hi-Tech Hidebound' Systems," in K. Srikantaiah and M.E.D. Koenig (Eds.), Knowledge Management for the Information Professional, Information Today, Inc., Medford, NJ, 2000. URL: http://www.kmbook.com

Malhotra, Y. (1993), ''Role of information technology in managing organizational change and organizational interdependence'', BRINT Institute, LLC, New York, NY, available at: www.brint.com/ papers/change/

Malhotra, Y. (1995), ''IS productivity and outsourcing policy: a conceptual framework and empirical analysis'', Proceedings of Inaugural Americas Conference on Information Systems (Managerial Papers), Pittsburgh, PA, August 25-27, available at: www.brint.com/papers/outsourc/

Malhotra, Y. & Kirsch, L. "Personal Construct Analysis of Self-Control in IS Adoption: Empirical Evidence from Comparative Case Studies of IS Users & IS Champions," in the Proceedings of the First INFORMS Conference on Information Systems and Technology (Organizational Adoption & Learning Track), Washington D.C., May 5-8, 1996, pp. 105-114

Malhotra, Y. (1997), ''Knowledge management in inquiring organizations'', Proceedings of 3rd Americas Conference on Information Systems (Philosophy of Information Systems Mini-track), Indianapolis, IN, August 15-17, pp. 293-5, available at: www.kmnetwork.com/km.htm

Malhotra, Y. "Deciphering the Knowledge Management Hype", Journal for Quality & Participation, July/August 1998, pp. 58-60. URL: http://www.kmbook.com

Malhotra, Y. "Toward a Knowledge Ecology for Organizational White-Waters," Invited Keynote Presentation for the Knowledge Ecology Fair 98: Beyond Knowledge Management, Feb. 2 - 27, 1998a, accessible online at: http://www.brint.com/papers/ecology.htm.

Malhotra, Y. Role of Social Influence, Self Determination and Quality of Use in Information Technology Acceptance and Utilization: A Theoretical Framework and Empirical Field Study, Ph.D. thesis, July 1998c, Katz Graduate School of Business, University of Pittsburgh.

Malhotra, Y. "High-Tech Hidebound Cultures Disable Knowledge Management," in Knowledge Management (UK), February, 1999c.

Malhotra, Y. "Knowledge Management for Organizational White Waters: An Ecological Framework," in Knowledge Management (UK), March, 1999d.

Malhotra, Y. (2000a), ''From information management to knowledge management: beyond the 'hi-tech hidebound' systems'', in Srikantaiah, K. and Koenig, M.E.D. (Eds), Knowledge Management for the Information Professional, Information Today Inc., Medford, NJ, pp. 37-61, available at: www.brint.org/IMtoKM.pdf

Malhotra, Y. (2000b), ''Knowledge assets in the global economy: assessment of national intellectual capital'', Journal of Global Information Management, Vol. 8 No. 3, pp. 5-15, available at: www.kmnetwork.com/intellectualcapital.htm

Malhotra, Y. (2000c), ''Knowledge management and new organization forms: a framework for business model innovation'', Information Resources Management Journal, Vol. 13 No. 1, pp. 5-14, available at: www.brint.org/KMNewOrg.pdf

Malhotra, Y. (2000d), ''Knowledge management for e-business performance: advancing information strategy to 'internet time''', Information Strategy: The Executive's Journal, Vol. 16 No. 4, pp. 5-16, available at: www.brint.com/papers/kmebiz/kmebiz.html

Malhotra, Y. (2001a), ''Enabling next generation e-business architectures: balancing integration and flexibility for managing business transformation'', Intel e-Strategy White Paper, June, available at: www.brint.net/members/01060524/intelebusiness.pdf

Malhotra, Y. (2001b), ''Expert systems for knowledge management: crossing the chasm between information processing and sense making'', Expert Systems with Applications, Vol. 20 No. 1, pp. 7-16, available at: www.brint.org/expertsystems.pdf

Malhotra, Y. (2002a), ''Enabling knowledge exchanges for e-business communities'', Information Strategy: The Executive's Journal, Vol. 18 No. 3, pp. 26-31, available at: www.brint.org/ KnowledgeExchanges.pdf

Malhotra, Y. (2002b), ''Information ecology and knowledge management: toward knowledge ecology for hyperturbulent organizational environments'', Encyclopedia of Life Support Systems (EOLSS), UNESCO/Eolss Publishers, Oxford, available at: www.brint.org/ KMEcology.pdf

Malhotra, Y. (2003), ''Measuring national knowledge assets of a nation: knowledge systems for development (expert background paper)'', Expanding Public Space for the Development of the Knowledge Society: Report of the Ad Hoc Expert Group Meeting on Knowledge Systems for Development, 4-5 September, Department of Economic and Social Affairs Division for Public Administration and Development Management, United Nations, New York, pp. 68-126, available at: www.kmnetwork.com/ KnowledgeManagementMeasurementResearch.pdf; http://unpan1.un.org/ intradoc/ groups/public/documents/un/unpan011601.pdf; http://unpan1.un.org/intradoc/ groups/public/ documents/un/unpan014138.pdf

Malhotra, Y. (2004a), ''Desperately seeking self-determination: key to the new enterprise logic of customer relationships'', Proceedings of the Americas Conference on Information Systems (Process Automation and Management Track: Customer Relationship Management Mini-track), New York, NY, August 5-8.

Malhotra, Y. (2004b), ''Why knowledge management systems fail. Enablers and constraints of knowledge management in human enterprises'', in Koenig, M.E.D. and Srikantaiah, T.K. (Eds), Knowledge Management Lessons Learned: What Works and What Doesn't, Information Today Inc., Medford, NJ, pp. 87-112, available at: www.brint.org/ WhyKMSFail.htm

Malhotra, Y. and Galletta, D.F. (1999), ''Extending the technology acceptance model to account for social influence: theoretical bases and empirical validation'', Proceedings of the Hawaii International Conference on System Sciences (HICSS 32), pp. 6-19, available at: www.brint.org/technologyacceptance.pdf

Malhotra, Y. and Galletta, D.F. (2003), ''Role of commitment and motivation in knowledge management systems implementation: theory, conceptualization, and measurement of antecedents of success'', Proceedings of the Hawaii International Conference on Systems Sciences (HICSS 36), available at: www.brint.org/KMSuccess.pdf

Malhotra, Y. and Galletta, D.F. (n.d.a), ''A multidimensional commitment model of knowledge management systems acceptance and use'', Journal of Management Information Systems (in press).

Malhotra, Y. and Galletta, D.F. (n.d.b), ''If you build IT, and they come: building systems that users want to use'', Communications of the ACM (in press).

Massey, A.P., Montoya-Weiss, M.M. and Holcom, K. (2001), ''Re-engineering the customer relationship: leveraging knowledge assets at IBM'', Decision Support Systems, Vol. 32 No. 2, pp. 155-70.

Murphy, C. (2003), ''Tying it all together'', Information Week, March 17, available at: www.informationweek.com/shared/printableArticle.jhtml?articleID ¼ 8700225

Nadler, D.A. and Shaw, R.B. (1995), ''Change leadership: core competency for the twenty-first century'', in Nadler, D.A., Shaw, R.B. and Walton, A.E. (Eds), Discontinuous Change: Leading Organizational Transformation, Jossey-Bass, San Francisco, CA.

Nonaka, I. and Takeuchi, H. The Knowledge-Creating Company, Oxford University Press, New York, NY, 1995. Odom, C. and Starns, J. (2003), ''KM technologies assessment'', KM World, May, pp. 18-28.

O'Dell, C. and Grayson, C.J. "If Only We Knew What We Know: Identification And Transfer of Internal Best Practices," California Management Review, 40(3), Spring 1998, pp. 154-174.

Porter, M.E. and Millar, V.E. (1985), ''How information technology gives you competitive advantage'', Harvard Business Review, Vol. 63 No. 4, pp. 149-60.

Rayport, J.F. and Sviokla, J.J. (1995), ''Exploiting the virtual value chain'', Harvard Business Review, Vol. 73 No. 6, pp. 75-99.

Schrage, M. (2002), ''Wal-Mart trumps Moore's law'', Technology Review, Vol. 105 No. 2, p. 21.

Stewart, T.A. (2000), ''How Cisco and Alcoa make real time work'', Fortune, May 29.

Strassmann, P.A., The Squandered Computer: Evaluating the Business Alignment of Information Technologies, Information Economics Press, New Canaan, CT, 1997.

Strassmann, P. (2003), ''Enterprise software's end'', Computerworld, May 12, p. 35.

Terreberry, S. (1968), ''The evolution of organizational environments'', Administrative Science Quarterly, Vol. 12, pp. 590-613.

Verton, D. (2002), ''Insiders slam navy intranet'', Computerworld, May 27, pp. 1-16.

Zetie, C. (2003), ''Machine-to-machine integration: the next big thing?'', Information Week, April 14, available at: www.informationweek.com/story/showArticle.jhtml?articleID ¼ 8900042

17
Towards a Reusable Evaluation Framework for Ontology based biomedical Systems Integration

Gilbert Maiga

Evaluation of ontology based integrated biomedical systems is important for them to find wide adoption and reuse in distributed computing environments that facilitate information exchange and knowledge generation in biomedicine. The review reveals many approaches to information systems and ontology based evaluation with standards, none of which are generic enough for use in all situations. It also shows increased use and reliance on ontologies for biomedical integration systems to overcome the issues of semantic heterogeneity and bridging across levels of granularity in biomedical data. The wide acceptance and reuse of ontology based integration systems remains hampered by the lack of a general framework to assess these systems for quality. To address this requirement, a new flexible framework for evaluating ontology based biomedical integration systems is proposed. The proposed framework extends existing Information systems and ontology evaluation approaches. The framework is also informed by the theories of formal ontology, self organizing systems, summative and formative evaluation. It has the potential to relate ontology structure to user objectives in order to derive requirements for a flexible framework for evaluating ontology based integrated biological and clinical information systems in environments with changing user needs and increasing biomedical data.

1.0 Introduction

Biomedical integration systems bring together disparate sources of rapidly changing biological and clinical data into a transparent system to enable biologists and clinicians maximize the utilization of available information for knowledge acquisition. Ontology use has become an important approach to resolving semantic interoperability in heterogeneous information sources and creating reusable models for integrated biomedical systems. Ontology Integration is a type of ontology reuse process. It aims at building a resultant ontology by assembling, extending, specializing and adapting other ontologies (Pinto and Martins, 2000). The advantage of integration is that, from a set of small, modular, highly reusable ontologies, larger ontologies for specific purposes can be assembled (Pinto and Martins, 2000). Despite many attempts, the lack of a single satisfactory unifying ontology integration and evaluation approach to biomedical data remains (Davidson et al., 1995; Ding and Foo, 2002), creating an obstacle for ontology reuse and their wide adoption by industry (Alani and Brewster, 2006).

Ontology evaluation assesses a given ontology using a particular criterion of application to determine the best one for a given purpose (Brank et al., 2005). Evaluations are important for ontologies to be widely adopted for use in distributed computing environments where users need to assess and decide which one best fits their requirements (Pinto and Martins, 2000; Kalfoglou and Hu, 2006). Evaluation is also important for the success of ontology based knowledge technology (Gangemi et al., 2005). In ontology integration systems, evaluation is a criterion based technical judgment guiding the construction process and any refinement steps of both the integrated and resulting ontologies (Pinto and Martins, 2000). Ontologies are subjective knowledge artifacts that reflect particular interests of knowledge users as captured in the design which makes it difficult to select the right properties to use in ranking them since selection can depend on personal preferences and user requirements (Alani and Brewster, 2006), who also mention a need to investigate properties users tend to look for when judging the general quality or suitability of an ontology. Kalfoglou and Hu (2006) also suggest the need for a holistic evaluation strategy with a greater role and participation of user communities in the evaluation process.

Current ontology based approaches for integrating biological and clinical data remain independent non reusable models due to lack of a common evaluation framework with metrics for comparison to ensure quality (Kalfoglou and Schorlmer, (2003); Lambrix and Tan (2006). Comparing the effectiveness of such ontology based integration systems remains difficult and meaningless due to lack of a standard framework for evaluating them, and the use of tools that differ in function, input and outputs required (Natalya and Musen, 2002). Specialized criteria to analyze the resulting ontology following integration of knowledge and research to develop accurate evaluation metrics for ontology engineering is required (Pinto and Martins, 2000).

Existing ontology based evaluation models define standards for structuring and integrating knowledge using static relationships between concepts in a domain of discourse. The vast amounts of biological and clinical data require ontology integration systems that are able to capture and represent new structure and functions that emerge as data in the two domains increases. This study proposes a flexible reusable evaluation framework for integrated clinical and biological information that uses completeness to measure quality. General systems theory (GST) and that of self organization (from the perspective designing and building artificial systems) are adopted to explain the emergent properties of the system (Heylighen and Gershenson, 2003). The framework is also informed by the formative and summative evaluation frameworks. A Formal ontology theory perspective is adopted for structuring biological and clinical information. The novelty of this approach lies in the ability to relate ontology structure and user derived objectives in order to derive requirements for a flexible framework for evaluating ontology based integrated biological and clinical information systems in environments where biomedical of data is ever increasing and user needs changing.

This paper is organized into seven sections of introduction, information systems evaluation approaches, evaluation of ontologies, biomedical information systems integration and current work related to its evaluation, the proposed evaluation framework followed by the methodology and conclusions.

2.0 Information systems evaluation

Evaluation is used to assess and guide the result of IS development (Sun and Kantor, 2006). The evaluation cycle needs to be regularly conducted as part of the system development life cycle (Cronholm and Goldkuhl, 2003). The theory of evaluation, rooted in the twin ideas of accountability for rationale, and social enquiry for deriving the evaluation models is classified using an evaluation tree with methods, valuing and use branches (Alkin and Christie, 2004). Methods deal with knowledge construction, valuing establishes the vital role of the evaluator in valuing while use focuses on decision making (Alkin and Christie, 2004).

Evaluation approaches may be formal rational, interpretive or criteria based (Cronholm and Goldkuhl, 2003). Formal rational evaluations are largely quantitative processes concerned with the technical and economic aspects of the project. Interpretive approaches view IS as social systems with embedded information technology and aim to get a deeper understanding of what is to be evaluated, generate commitment and motivation. According to Cronholm and Goldkuhl (2003), evaluation should be performed depending on the evaluation context using three general strategies of goal based, goal free and criteria based evaluation. Goal based evaluations are formal rational and focus on intended services and outcomes of a program using quantitative or qualitative goals. Goal free evaluation is an interpretive approach preformed with limited evaluator involvement. Criteria based evaluation use selected general qualities for evaluation (Cronholm and Goldkuhl, 2003).

A framework to explain various approaches to information systems evaluation was developed by Ballantine et al., (2000). According to the framework IS evaluation is driven by purpose, process and people and is influenced by six factors of philosophy (technical and moral issues), power politics and cultural beliefs in the organization, the management style, the evaluator and resources (Ballantine et al., (2000). A matrix of these six factors gives an indication of the values and factors which underlie the use of evaluation approaches.

Avison et al., (1995) adopt a non technical contingency view in which evaluation is not an objective rational activity, but one dependent on the motives of people doing it making power and organizational issues important. They also identify impact analysis, effectiveness, economic, objectives, user satisfaction, usage, utility, standards, technical evaluation and process as approaches to IS evaluation. On the interpretive perspective to IS design and evaluation, Walsham (1993) argues that IS evaluation should consider the issues of content, social context and social process. Content considers issues of purpose for conducting the evaluation and associated factors. Social context considers the stakeholders in the situation, their needs and how to resolve conflict between those needs. Walsham (1993) also suggests that IS

evaluation is a multi-stage process occurring at several points, in different ways, during the product life-cycle and it is important to consider evaluation as a learning process for all involved.

Beynon-Davis et al (2004) propose a model for IS evaluation closely linked to the development process in which they distinguish between strategic, formative, summative and post mortem analysis. Strategic evaluation is conducted as part of the planning process during project selection and feasibility study. It attempts to establish the balance of predicted costs and benefits for an intended IS. Formative evaluation is a continuous, iterative informal process aimed at providing systematic feedback to designers and implementers, influencing the process of development and the final information system (Kumar, 1990; Walsham, 1993; Remenyi and Smith, 1999). Summative evaluation usually done at the end of the project is concerned with assessing the worth of a program outcome in light of initially specified success criteria (Walsham, 1993; Kumar, 1990). Post mortem analysis is conducted if the project has to undergo total, substantial or partial abandonment (Beynon-Davis et al., 2004).

3.0 Ontology evaluation

Ontology evaluation assesses a given ontology using a particular criterion of application in order to determine which of several ontologies best suits a given purpose (Brank et al., 2005). It is important if ontologies are to be widely adopted for use in distributed computing environments (Pinto and Martins, 2000; Kalfoglou and Hu, 2006). It is important for the success of ontology based knowledge technology (Gangemi et al., 2005). In integration systems, evaluation is a criterion based technical judgment guiding the construction process and any refinement steps of both the integrated and resulting ontologies (Pinto and Martins, 2000). Ontologies may be assessed by user ratings and reviews, meeting requirements of certain evaluation tests or for general ontological properties (Alani and Brewster, 2006). Evaluation can be conducted during design and development and prior to use (Kalfoglou and Hu, 2006). Approaches to ontology evaluation are identified according to type, purpose and level of evaluation, or they qualitative, quantitative, formal and philosophical.

According to ontology type and purpose, approaches are categorized by: comparison to a golden standard, the results of using the ontology in an application, comparisons with a source of data about the domain to be covered by the ontology, human evaluation to assess how well the ontology meets a set of predefined criteria, standards and requirements (Brank et al., 2005). Secondly, depending on the level at which an ontology is evaluated, approaches are grouped into: lexical, vocabulary, or data layer that focus on concepts, instances, facts included in the ontology, and the vocabulary used to represent or identify these concepts, hierarchy or taxonomy based on the "is-a" relationship, other semantic relations besides is-a relations including measures of precision and recall, context or application level, Syntactic level, structure and architecture (Brank et al., 2005). Multicriteria is a third approach to ontology evaluation. For each criterion, the ontology is

evaluated and given a numerical score. An overall score for the ontology is then computed as a weighted sum of its per-criterion scores. Multicriteria approaches support a combination of criteria from most of the levels (Brank et al., 2005).

Qualitative approaches to OE (Guarino, 1998; Gomez-Perez, 1994; Hovy, 2001) compare ontology to a gold standard but do not offer quantitative metrics for ontology evaluation. Brewster et al., (2004) propose methods to evaluate the congruence of an ontology with a given corpus in order to determine how appropriate it is for the representation of the knowledge of the domain represented by the texts. They also argue for the need to establish objective measures for ontology creation and the need for ontology ranking techniques as the number of ontologies available for reuse is continues growing.

Baker et al (2005) discuss philosophical, domain dependent and domain independent criteria for ontology evaluation. Philosophical Ontologists are concerned with evaluation for correctness of the conceptualization of knowledge in an ontology (Smith et al., 2004; Baker et al., 2005). Domain dependent evaluation is based on content and application (Baker et. al., 2005). Ontometric is an example of a tool for quantitative domain content evaluations of goal-based characteristics (Lozano-Tello and Gomez-Perez, 2004). Ontology Web Language constructs, formal and description logic are domain independent evaluations (Baker et al., 2005). In OWL construct evaluation, ontologies are ranked using a metric computed as a ratio of the frequency of class features (concepts) to properties and modifier features (roles and attributes) in its OWL sub language constructs.

In formal evaluation, the taxonomical structures of ontologies are compared with ideal predefined ones to detect inconsistencies. The OntoClean methodology relying on the notion of rigidity, unity, identity and dependence to evaluate whether specified constraints are violated within the ontology is a case of domain independent formal evaluation (Guarino and Welty, 2002). It is used during development for the formal evaluation of properties defined in an ontology using a predefined ideal taxonomical structure of metaproperties. In description logics evaluation, queries are used to interrogate the ontologies to reveal their level of complexity, suggesting their maturity and suitability to support knowledge discovery. Metrics used are: the classification hierarchies of ontologies, depth in ontologies, numbers of concepts, roles, instances, average number of child concepts and multiple inheritances (Haarslev et al., 2004).

Functional, usability and structural measures have been used to define a theoretical framework for ontology evaluation (Gangemi et al, 2005). Functional evaluation focuses on measuring how well an ontology serves its purpose (function). Usability evaluation is concerned with metadata and annotations of the ontology. Structural evaluation focuses on the structural properties of the ontology as a graph.

Issues with ontology evaluation

Ontologies are subjective knowledge artifacts in terms of time, place and cultural environment reflecting the particular interests of knowledge users as captured in

the design; this makes it difficult to pinpoint the right selection of parameters or structural properties to use in ranking ontologies since selection can depend on personal preferences of use requirements (Alani and Brewster, 2006). The authors point to the need for user based experiments to find out what properties users tend to look out for when judging the general quality or suitability of an ontology. This view is also expressed by Kalfoglou and Hu (2006) who suggest the need for a holistic evaluation strategy with a greater role and participation of user communities in the evaluation process.

4.0 Biomedical information systems integration

Biomedical integration systems bring together disparate sources of biological and clinical information into one coherent and transparent system to enable biologists and clinicians discover interesting relationships between database objects to formulate and test hypothesis in order to generate knowledge (Hongzhan et al., 2004; Hernandez and Subbarao, 2002). Existing data management technology is challenged by lack of stability, evolving nature, diversity and implicit scientific context that characterize biological data (Davidson et al., 2004). New tools are required to relate genetic and clinical data (Martin-Sanchez et al., 2004). Ontology use has become an important approach for biomedical databanks integration.

Ontology Integration approaches in Biomedicine

Philosophical Ontology is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality. Broadly, it refers to the study of what might exist (Smith, 2003). Information systems ontology is an agreement about a shared, formal, explicit and partial account of a conceptualization (Gruber, 1995; Ushold and Gruniinger, 1996). It makes possible sharing and reusing of knowledge, supports interoperability between systems and allows inference to be done over them. Computer ontologies facilitate semantic interoperability in heterogeneous information sources (Jarrar et al., 2002), enabling data exchange between different models described in standard formats (Hucka, 2003). Ontology integration (OI) is the composition of many ontologies to build new ones whose respective vocabularies are usually not interpreted in the same domain of discourse (Kalfoglou and Schorlermmer, 2003). OI is a type of ontology reuse process that aims at building an ontology, by assembling, extending, specializing and adapting, other ontologies which become parts of the resulting ontology (Pinto and Martins, 2000). Its advantage is that, from a set of small, modular, highly reusable ontologies, large ontologies for specific purposes can more easily be assembled (Pinto and Martins, 2000).

Integrating biomedical information systems remains challenging largely due to problems of semantic heterogeneity and differences in levels of granularity of the data sources, and OI is a possible approach to overcome this problem. Ontology-based attempts to bridge the gap between clinical and biological information in order to achieve interoperability include ONTOFUSION (Perez-Rey et al., 2006), SEMEDA (Kohler et al., 2003), ASLER (Yugyung et al., 2006), Kumar et

al., (2006) on the colon carcinoma, and that of (Sioutos et al., 2006) on cancer. The ONTOFUSION tool provides semantic level integration of genomic and clinical databases using a multiagent architecture, based on two processes: mapping and unification (Perez-Rey et al., 2006). A limitation of this approach is that the unified ontologies are too generic and the imperfections in domain ontologies are propagated into the virtual schemas (Kumar et al., 2006). It also does not bridge across levels of granularity for biomedical data. Using principles of the basic formal ontology, Kumar et al., (2006) describe a framework for integrating medical and biological information in order to draw inferences across various levels of granularity using the three sub ontologies of the Gene Ontology. Yugyung et al., (2006) describe a methodology for medical ontology integration using an incremental approach of semantic enrichment, refinement and integration (ALSER) that depends on measures of similarity between ontology models. A terminology and description logic based framework for integrating molecular and clinical cancer-related information has been described by Sioutos et al (2006). However, the approach is specific to cancer related integration issues and integrating with external sources remains a largely unresolved issue (Sioutos et al., 2006).

Evaluation of ontology integration systems - Related Work

Little work exists on evaluating the results of integrating biomedical ontologies. Pinto and Martins (2000) mention criteria for evaluating both the integrated ontology and the resulting ontology. They recommend both technical and user assessment of the candidate ontologies by domain experts and ontologists respectively using specialized criteria oriented to integration, and the selection of candidate ontologies using strict (hard) and desirable (soft) requirements. Using strict or desirable requirements, as a metric provides flexibility, as they can be adapted to integration processes that take into account particular features during the choice of one ontology.

Evaluation of the resultant ontology can be done according other criteria used for any ontology. The resultant ontology should be evaluated (verification and validation) and meet assessment criteria including completeness, conciseness, consistency, expandability and robustness (Gomez-Perez and Pazos 1995). Any ontology with an adequate design should have clarity, coherence, and extendibility, minimal encoding bias, minimal ontological commitment (Gruber, 1995). Pinto and Martins, 2000) point to the need for the resulting ontology to be consistent, non ambiguous and have both an adequate and appropriate level of detail.

Despite many approaches to biomedical data integration, there is lack of a single satisfactory strategy This lack of a single unifying approach to ontology integration is underscored by Ding Foo (2002) who point out the need to improve ontology integration using a structured approach that addresses issues of verification and consistency conditions for ontologies to be merged, the parameters to consider and the integration of ontologies with diverse relations. This lack of a unifying evaluation framework for integrated systems remains an obstacle for ontology reuse and may hinder their adoption by industry and the wider web community

(Alani and Brewster, 2006). The authors point to the need for ontology ranking techniques as the number of ontologies available for reuse continues to grow.

5.0 The proposed evaluation framework

Theoretical orientation

General systems theory (GST) and that of self organization (SOS) are adopted to explain the dynamic and emergent properties of biomedical integration systems. This study is also informed by the formative and summative evaluation frameworks. General systems theory is "elements in standing relationship, the joining and integrating of the web of relationships creates emergent properties of the whole that are not found in any analysis of the parts" (Von Bertalanffy, 1962). GST explains structure and properties of systems in terms of relationships from which new properties of wholes emerge. Some properties-of-the-whole cannot be found among those of elements, and the corresponding behavior of the whole cannot be explained in terms of the behavior of the parts.

Self organization (SO) is a process in which the internal organization of a system increases in complexity without being guided or managed by an outside source and displays emergent properties. The emergent properties do not exist if the lower level is removed (Gershenson, 2006). In SOS the environment is unpredictable and the elements interact to achieve dynamically a global function or behavior (Gershenson, 2006). Engineered systems cope with the dynamic or unpredictable environment by adaptation, anticipation, robustness or a mixture of these features (Gershenson, 2006). Self-organizing systems, rather than being a type of systems, are a perspective for studying, understanding, designing, controlling, and building systems; the crucial factor being the observer, who has to describe the process at an appropriate level and aspects, and to define the purpose of the system; self-organization can therefore be everywhere, it just needs to be observed (Heylighen and Gershenson, 2003). Organization is seen as structure that has a purpose. The observer has to focus their viewpoint, set the purpose of the system to see the attractor as an organized state at the right level and in the right aspect in order to observe self-organization - a perspective that can be used for designing, building, and controlling artificial systems. A key characteristic of an artificial self-organizing system is that structure and function of the system emerge from interactions between the elements (Heylighen and Gershenson (2003). It is this perspective of self organizing systems that is adopted for this study to explain the evaluation of ontology based biomedical integration systems

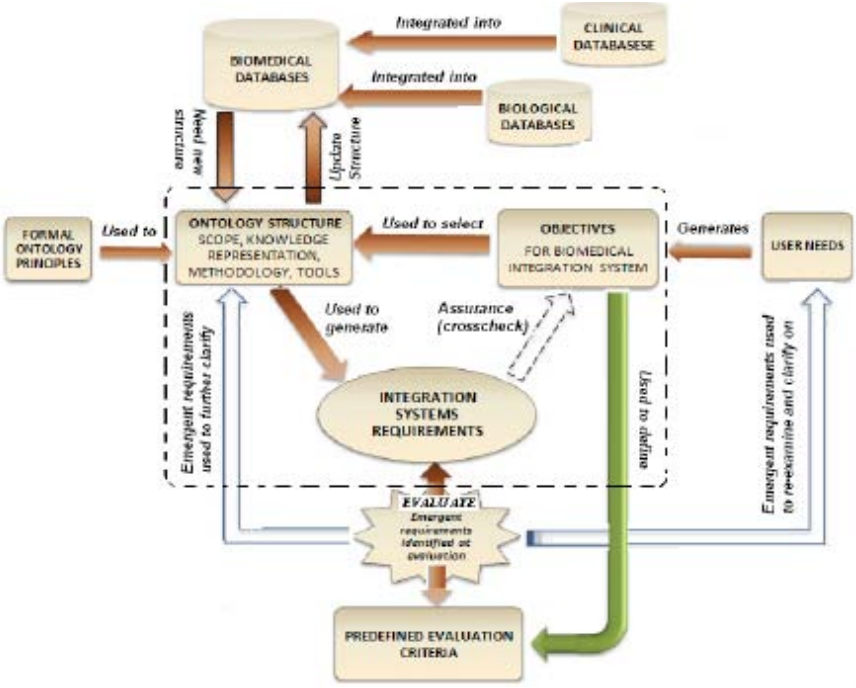
Current ontology based evaluation models define standards for structuring and integrating knowledge using static relationships between concepts in a domain of discourse. The vast amounts of biological and clinical data require ontology integration structures that are able to capture and represent the dynamic nature of relationships that emerge as data in the two domains increases. Integrating such clinical and biological information leads to biomedical systems with new ontology structure and functions. The objectives for building biomedical integration systems

also change with changing user needs. Change in ontology structure and user led objectives drive change in requirements for biomedical integration systems. New requirements for biomedical integration systems continuously emerge (emergent requirements). An effective evaluation process based on user needs, objectives and ontology structure has to account for and assess such dynamic relationships and emergent requirements. The proposed evaluation framework determines system quality for the biomedical integration system using effectiveness as a metric.

Conceptual framework

The framework assesses the suitability of a biomedical integration system against specified evaluation criteria in an environment of non static ontology structure and changing user needs. It seeks to determine the effectiveness of a biomedical integration system using completeness as a measure of system quality. It attempts to explain and answer key questions about (1) what do users need biological and clinical integration systems for? What are the limitations of existing systems in addressing these needs? What criteria help users to articulate these limitations? (2) For biomedical integration systems, what should be evaluated, for what purpose and which criteria should be used? (3) What is system quality in biomedical integration system? What requirements should a quality integration system conform to? What criteria help users to articulate these requirements? (4) What are the ontology structural requirements for a reusable quality biomedical integration system? How do they offer support for user needs and objectives? What criteria should be used to articulate such requirements? (5) What is the relationship between ontology based structural and user defined requirements?

The model



Dimensions of the model

The framework has dimensions of formal ontology, ontology structure, user needs, objectives, biomedical databases, integration system requirements and evaluation criteria. Clinical and biological (genomic) databases act as sources of data for the biomedical databases. Formal ontology is proposed in this model to structure such data in order to create knowledge and thereafter to update the biomedical databases. Formal ontology is used to define structural properties for a good quality integration system including the methodologies, development language and software. These have also been used in the OntoMetric evaluation method that enables selection of the most appropriate ontology for a project by comparing the importance of project objectives and the characteristics of ontologies in order to justify decisions (Lozano-Tello and Gomez, 2004).

The users and stakeholders of biomedical integration systems are biologists, medical professionals, health care workers and information systems specialists. Clinicians need to relate clinical data with insights from genetic research in order to advance knowledge about diagnosis and treatment of disease. Scientists need to test hypothesis by use of both existing clinical and genomic data. Information systems specialists need tools to structure, compare and manipulate the data. The various user needs are decomposed into objectives. The objectives as seen from the perspectives of the stakeholders are used to inform criteria for evaluating integration systems and guide the generation of ontology based biomedical integration requirements. Requirements for a quality biomedical integration system are generated by mapping those ontology based structures, functions and characteristics that help to meet the specified objectives. These are selected on the basis of ability to enable the system to conform to operational objectives. Once derived, they can be verified against the objectives

System quality, a key factor for evaluating this model is defined as "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or an implied need" (ISO 91, pp. 16). Quality is also viewed as "conformance to requirements" (Crosby, 1979, pp. 15; Garvin, 1984). This is the perspective from which quality is defined for this model. A need can be decomposed into a set of requirements upon features and characteristics of a product or service, and if all these requirements are conformed to, then that need has been satisfied. The evaluation criterion seeks to establish how well the biomedical integration system conforms to requirements. This is completeness for the system, and can be expressed as:

$$\text{Completeness} = \frac{\text{All Requirements met}}{\text{Total system requirements defined}}$$

In the model, this relationship expresses how well the ontology based structures satisfy the objective for biomedical system integration. Following evaluation, a proportion of total requirements defined are not met. These are here referred to

as emergent requirements (ER) and require re-examination of user needs and /or ontology structure leading to a redefinition of the integration requirements.

ER = (Total system requirements defined - All Requirements met)

Steps for using the framework

The objectives as generated from user needs. These guide the selection of those ontology based properties that help to meet objectives. It is from these properties that integration system requirements are derived. The proposed framework has guidelines to help modelers of ontology based integrated biological and clinical information systems to build systems that meet requirements for users. They also enable modelers to meet system requirements but are not an absolute solution on their own. The steps to be followed when using this framework for evaluating ontology based biomedical integration systems are:

1. Identification and prioritization of user needs.
2. Derive objectives from user needs.
3. Generate evaluation criteria from user needs and objectives
4. Describe ontology quality characteristics for the integration system.
5. Compare and match (map) objectives to ontology quality descriptions. Rate them.
6. Generate system requirements from quality descriptions.
7. Compare requirements against the predefined evaluation criteria.
8. Generate emergent requirements. Compare them to user needs and quality descriptions to refine requirements

The novelty of this framework lies in the ability to relate ontology structure and user derived objectives in order to derive requirements for evaluating ontology based integrated biological and clinical information systems in environments where the amounts of data are ever increasing and user needs are changing.

6.0 Proposed methodology

The planned approach uses both qualitative and quantitative research methods. Qualitative methods are used to get a full understanding of user needs and quality descriptions for biomedical integration systems. Following the development of the evaluation tool, a summative evaluation of a biomedical information system involving users will be undertaken to validate the framework.

From a review of existing literature and document analysis, it will be possible to identify and describe common structures, functions, inputs, outputs that enhance quality in ontology based integrated biomedical information systems. Questionnaires and in depth interviews with stakeholders (research scientists, molecular biologists, clinicians and health care workers in research institutes, health departments and hospitals) are to be used to collect data on user needs relevant for biomedical integration systems. Objectives for integration shall be synthesized from these needs and mapped (matched) to the quality descriptions

for the ontology based integrate systems. Requirements for the integrated system are to be derived from the quality descriptions.

A java based tool shall be built to validate the model with the stakeholders. UML is to be used to create a model for the tool development based on the requirements derived from the quality descriptions. Theoretical test data and selected biomedical integration systems are to be applied in the evaluation of the tool by users.

7.0 Conclusions

The paper identifies the challenges faced in evaluating ontology based biomedical integration systems. Evaluation is shown to be an important aspect of developing reusable ontology based biomedical integration systems for distributed computing environments. The paper provides the research background and approach to work in progress that aims to develop a reusable framework for evaluating integrated biomedical information systems. The theoretical, conceptual model and methodology towards such a framework are outlined. Steps for the utility of such a framework are also given.

References

Alani, H. and Brewster C. (2006). Metrics for Ranking Ontologies. In Proceedings of the 4th International Conference on Evaluation of Ontologies for the Web. EON Workshop Edinburgh International Conference Center. Edinburgh UK. May 22nd 2006. http://km.aifb.uni-karlsruhe.de/ws/eon 2006

Alkin, M.C. & Christie, C.A. (2004). An evaluation theory tree. In M.C. Alkin (Ed.), Evaluation Roots. Thousand Oaks, CA: Sage. Chapter 2. Available: www.sagepub.com/upm-data/5074_Alkin_Chapter_2.pdf.

Avison, D., Horton, J., Powell, P., Nandhakumar, J. (1995). Incorporating Evaluation in the Information Systems Development Process. Proceedings of the third evaluation of Information Technology Conference.Henly, UK.

Baker, C. O., Warren, R. H., Haarslev, V. (2004). Ontology Evaluation. Available: http://junobeach.cs.uwaterloo.ca. Accessed: 30th April 2007.

Ballantine, J., Levy, M., Martin, M., Munro, I., Powell, P., (2000). An Ethical Perspective on Information Systems Evaluation. International Journal of Agile Management Systems, 2(3) 233- 241.

Beynon-Davis, P., Owens, I., Williams, D.M. (2004). Information systems Evaluation and the Information Systems Development Process. The Journal of Enterprise Information Management. Volume 17 · Number 4 · 2004 · pp. 276–282. Emerald Group Publishing Limited

Brank, J., Grobelnik, M. and Mladenic, D. (2005). A Survey of Ontology Evaluation Techniques. At the conference on Data Mining and Data Warehouses (SiKDD).

Brank, J., Grobelnik, M. and D, Mladenic. (2006). Gold Standard Based Ontology Evaluation Using Instance Assignment. In Proceedings of the 4th International Conference on Evaluation of Ontologies for the Web. EON Workshop Edinburgh International Conference Center. Edinburgh UK. May 22nd 2006. http://km.aifb.uni-karlsruhe.de/ws/eon2006

Brewster,C., Alani, H., Dasmahapatra, S. Wilka, Y. (2004) Data Driven Ontology Evaluation. Proceedings of the Language Resources and Evaluation Conference (LREC 2004), Lisbon, Portugal.

Cronholm, S. and G. Goldkuhl (2003). Strategies for Information Systems Evaluation: Six Generic Types. Electronic Journal of Information Systems Evaluation 6(2): 65 -74.

Crosby, P.B., Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill, 1988.

Davidson, S., Overton, C. and P. Bunan. (1995). Challenges in Integrating Biological Data Sources. Journal of Computational Biology. 2(4).

Ding, Y. Foo, S. (2002) Ontology research and development, Part 2. A Review of Ontology mapping and evolving. Journal of Information Science, 28, 375 - 388.

Finne, H., Levin, M. and Nilssen, T, (1995). Trailing Research: a model for useful program evaluation, Evaluation, Vol 1,No 1.

Gangemi, A., Catenacci, C., Ciaramita, M. and Lehmann, J. (2005) A theoretical framework for ontology evaluation and validation. In Proceedings of SWAP 2005 .

Garvin, D.A., What does 'Product Quality' Really Mean?, Sloan Management Review, fall, 1984.

GERSHENSON, C. (2006). A General Methodology for Designing Self-Organizing Systems. ACM Journal Name, Vol. V, No. N, M 20YY.

Gomez-Perez, A. (1994). Some ideas and Examples to Evaluate Ontologies. Technical Report KSL-94-65. Knowledge System Laboratory. Stanford University. Also in Proceedings of the 11 th Conference on Artificial Intelligence for Applications, CAIA94.

Gruber, T. (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer studies, 43 (5/6):907–928.

Guarino, N. Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM: Vol. 45, No. 2.

Gomez-Perez, A. (1994). Some ideas and Examples to Evaluate Ontologies. Technical Report KSL-94-65. Knowledge System Laboratory. Stanford University. Also in Proceedings of the 11 th Conference on Artificial Intelligence for Applications, CAIA94.

Gomez-Perez and Pazos (1995).

Haarslev, V., Moeller, R. Wessel, M. (2004). Querying the Semantic Web with Racer + nRQL. In Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL'04). Ulm, Germany.

Hammer J. and Schneider, M. (2003). Genomics Algebra: A New, Integrating Data Model, Language, and Tool Processing and Querying Genomic Information. In Proceedings of the 2003 CIDR Conference.

Hernandez H. and Subbarao K. (2002). Integration of Biological Sources: Current Systems and Challenges Ahead.

Heylighen, F. and Gershenson, C. (2003). The meaning of self-organization in computing. IEEE Intelligent Systems, 72–75. http://pcp.vub.ac.be/Papers/IEEE.Self-organization. pdf.

Hongzhan, H. and Cathy, H., Anastasia, N., Zhangzhi H., Winona, C. and Barker, B. (2004). The iProClass integrated database for protein functional analysis, Computational Biology and Chistry, \bf 28(2004):87–96.

Hovy, Eduard,2001. Comparing sets of semantic relations in ontologies. In Rebecca Green,Carol A Bean, and Sung Hyon Myaeng(eds.), Semantics of Relationships, chapter 6.Dordrecht, NL:Kluwer.

Hucka, M. (2003). The System Biology Markup Language (SBML): A medium for representation and exchange of biochical network models. Bioinformatics, 19 (4):524-531.

Jarrar, M. and Meersman, R., (2002). Formal Ontology Engineering in the DOGMA Approach, Proc. of the Internat. Conf. on Ontologies, Databases and Applications of Semantics (ODBase 02), LNCS 2519, Springer Verlag pp. 1238-1254.

Kalfoglou, Y. and Schorlmer, M. (2003) Ontology mapping: the state of the art. The Knowledge Engineering Review, 18(1): 1–31.

Kalfoglou, Y. and Schorlmer, M. (2006). Issues with Evaluating and Using Publicly Available Ontologies

Kohler, J., Philipi, S. Lange, M. (2003) SEMEDA: ontology based semantic integration of biological databases. Bioinformatics, 19, 2420 -2427.

Kumar, K.(1990) Post Implentation Evaluation of Computer based Information Systems(CBIS): current practices, Communications of ACM, Vol 33, No 2, pp 203-12

Kumar, A., Linayip, Y., Smith, B. Grenon, P. (2006) Bridging the gap between medical and bioinformatics: An ontological case study in colon carcinoma. Computers in Biology and Medicine 36, 694–711

Lambrix, P. Tan, H. (2006) SAMBO—A system for aligning and merging biomedical ontologies Web Santics: Science, Services and Agents on the World Wide Web, 4 196 - 206

Lozano-Tello, A; Gomez-Perez, A (2004). ONTOMETRIC: A Method to Choose the Appropriate Ontology". Journal of Database Management. Special Issue on Ontological analysis, Evaluation.

Martin-Sanchez, I. Iakovidis, S. Norager, V. Maojo, P. de Groen, J. Van der Lei, T. Jones, K. Abraham-Fuchs, R.Apweiler,A. Babic, R. Baud,V. Breton, P. Cinquin, P. Doupi, M. Dugas, R. Eils, R. Engelbrecht, P. Ghazal, P. Jehenson, C.Kulikowski, K. Lampe, G. De Moor, S. Orphanoudakis, N. Rossing, B. Sarachan, A. Sousa, G. Spekowius, G. Thireos, G.Zahlmann, J. Zvarova, I. Hermosilla, F.J. Vicente, Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care,\t J. Biomed. Inform. \bf 37 (1) (2004) 30–42.

Natalya F. N. and Mark A. Musen. (2002). Evaluating Ontology-Mapping Tools: Requirements and Experience. Available: http: //citeseer.nj.nec.com. Accessed 6 may 2006.

Perez-Rey, D., Maojo, V., Garcia-Resal, M., Alonso-Calvo, R., Billhardt, H., Martin-Sanchez, F. Sousa, A. (2006). ONTOFUSION: Ontology-based integration of genomic and clinical databases. Computers in Biology and Medicine, 36, 712-730.

Pinto, H. S. Martins, J. P. (2000) Reusing Ontologies. In Proceedings of AAAI 2000 Spring Symposium Series, Workshop on Bringing Knowledge to Business Processes, AAAI Press.

Smith, B. (2003). Ontology. Blackwell Guide to the Philosophy of Computing and information, Oxford: Blackwell, 55- 166.

Smith, B., Köhler, J. Kumar, A. (2004). On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. E. Rahm (ed.), Database Integration in the Life Sciences (DILS 2004). Berlin: Springer.

Sioutos, N., Coronado, S., Haber, M., Hartel, F., Shaiu, W. Wright, W. C. (2006) NCI thesaurus: A santic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics, In Press.

Sun, Y. Kantor P.B (2006) Cross-Evaluation: A New Model for Information System Evaluation. Journal of the American Society for Information Science and Technology, 57, 614–628.

Ushold, M. and Gruninger M. (1996). Ontologies: Principles, methods and applications, in The Knowledge Engineering Review, 11(2): 93–55.

Von Bertalanffy, Ludwig (1962). General system theory - A Critical Review. General Systems 7, 1-20.Information Systems, May/June 2003. [In press]

Wache, S., Vogele, T. U., Visser, H., Stuckenschmidt, G., Schuster, H. and Hubner, S. (2001). Ontology-Based Integration of Information: A Survey of Existing Approaches.pp1-10.

Walsham, G. (1993). Intepreting Information systems in organizations, Wiley, Chichester.

Yugyung, L., Supekar, K. Geller, J. (2006) Ontology Integration: Experience with medical Terminologies. Computers in Biology and Medicine, 36, 893 - 919.

18
Organisational Implementation of ICT: Findings from NGOs in the United Kingdom and Lessons for Developing Countries

Dr Geoffrey Ocen

Summary. The UK and Uganda Governments are encouraging the voluntary/NGO sector to improve its infrastructure and deliver public services. This paper offers a new Technology Adoption Model (TAM) for dealing with the issues governing ICT adoption and the factors driving wider diffusion in voluntary and community sector organisations in the UK. The paper reports on website adoption process in 5 small and medium-sized voluntary sector organisations and identify Technology, Organisational and People (TOP) imperatives that provide new conceptual framework for facilitating organisational implementation of ICT. It suggests that it is helpful to classify organisations using a two-dimensional classification based on TOP schematic diagrams. This allows organisations to identify a vision, plan and implement effective ICT adoption. NGOs in Uganda and other developing countries face complex and particular problems that create barriers to ICT adoption and diffusion. The paper concludes by outlining cautious application of the TAM framework to such NGOs and other organisations.

Introduction

The UK and Uganda Governments are encouraging the voluntary/NGO sector to improve its infrastructure and deliver public services. Wealth creation and the fight against multiple deprivation and poverty can not be tackled by economic measures alone. It is recognised that voluntary/NGO sector have much to offer in terms of ensuring wider access to services. For example, it is the intention of the Uganda Government to promote its partnership with the sector and to enhance and up-grade the capacity of NGOs in order to enable the sector to participate effectively in service delivery, and raise the pace of development in rural areas and of poverty eradication nationwide (Lawson D, McKay A, and Okidi J, 2005). There is urgent need to increase the knowledge base with regards to the operation of the organisations in the NGO sector with a view to increasing the capacity of the organisations. Two of the most significant forces shaping organisations are globalisation and the continued, rapid and radical changes taking place in Information and Communication Technologies (ICT). The Australian Apec Study Centre (2002) reports that the sociologist, Anthony Giddens, defines globalisation as a decoupling of space and time, emphasising that

with instantaneous communications, knowledge and culture can be shared around the world simultaneously. It is the advances in ICT that has made instantaneous communications a reality. Accordingly, exploitation of ICT will assist with achievement of organisation effectiveness. To date, the extant literature has centred on the technology take up amongst businesses and larger voluntary organisations. This paper aims to begin to address this deficit. It provides an overview of a new framework for organisational implementation of IT by voluntary/NGO organisations and reports on the application of the model in 5 small and medium sized voluntary sector organisations in the UK. It concludes with a cautious expansion of the model to NGOs in Uganda and other developing countries.

In this paper, the hypothesis is summarised as follows:

The take up of ICT by voluntary/NGO sector can be supported and facilitated by identifying inhibiting factors and pursuing a programme of ***technology*** *adoption,* ***organisational*** ***c****hange and* ***people*** *development to achieve effective organisational implementation of ICT*

The UK Voluntary Sector Context

According to The UK Voluntary Sector Almanac (2006), key statistics for 2003/04, the latest data available, indicate that the sector: has an income of £26.3 billion, derives 38% of its income from statutory sources, has an operating expenditure of £24.9 billion, total assets of £66.8 billion and a paid workforce of at least 608,000. With regards to the future, the increasing expectations being placed upon voluntary and community organisations of all shapes and sizes. In particular, the greater emphasis on delivery of public services and the drive to increase active citizenship are leading to a larger, more visible sector (ibid).

It is generally acknowledged that funders are still largely interested in capital equipment rather than investing in strategy and running costs. Whilst equipment is fine, the real value of ICT is in the application of the tools in terms of making it work more effectively (Davey 2005). Sustainability and competitiveness of the sector will, in large part, depend on effective adoption of ICT and developing a culture of innovation and change.

The Home Office Active Community Unit (2003) reports that main barriers affecting ICT take up include: A lack of strategic understanding of ICT at senior management and trustee level, many organisations do not have ICT strategy, few sources of ICT advice and support and lack of affordable technical support. In addition, a preliminary audit from the 5 case studies identified the following additional barriers: Lack of funding which is most acute in smallest organisations, lack of staff appreciation or high staff resistance and lack of internal "Change Champions."

The Uganda NGO Sector Context

Uganda is selected to provide the context for a typical developing country. It has a national vision for the eradication of poverty focused on increased and sustained

economic growth, promoting good governance and security, increasing incomes for the poor, and improving the quality of life for all. The NGO sector is an important partner in its specific focus on basic education, public information and community-level services.

Willets (1995) states that the term, NGO or "non-governmental organisation" came into currency in 1945 because of the need for the UN to differentiate in its Charter between participation rights for intergovernmental specialised agencies and those for international private organisations. There has been an explosive growth in the number of NGOs, especially in the service delivery sectors, such as, health, education, micro-finance, roads, water and sanitation, and agriculture (Lawson D, McKay A, and Okidi J, 2005). About 3,500 NGOs have registered with the NGO Registration Board in 2000/01.

Using The World Bank's convention cited in Duke University (2001), there are two main categories of NGOs: i) operational NGOs - whose primary purpose is the design and implementation of development-related projects, and; ii) advocacy NGOs - whose primary purpose is to defend or promote a specific cause. The World Bank further classifies operational NGOs into three main groups: i) community-based organizations (CBOs) - which serve a specific population in a narrow geographic area; ii) national organizations - which operate in individual developing countries, and; iii) international organizations – which carry out operations in more than one developing country. The NGO Forum (2003) reports that with the exception of traditional faith-based organisations, the NGO sector in Uganda is still in its infancy, and most NGO are small. Many remain unspecialized and unfocused. Many consider themselves holistic and favour capacity building, advocacy and lobbying to direct service delivery, particularly so at national level. The Forum believes that as the sector matures, and professionalism increases, there is likely to be greater concentration. In general, advocacy NGOs carry out much the same functions, but with a different balance between them. In practice, operational NGOs often move into advocacy mode and advocacy NGOs sometimes run projects to meet short-term needs.

Ugandan NGOs receive funding mainly from international NGOs and donors. The average NGO generates about 2.5% of its funding from members and individual donation (ibid).

The nature and quality of individual NGOs vary greatly and generalisations about the sector is difficult to make. The Uganda Government and the World Bank are conducting a study of NGOs as service providers in Uganda. The objectives of the study are to describe the work of development NGOs, assess their effectiveness and efficiency, analyze resource flows to and from NGOs and incentives within the organisations, and understand what factors motivate NGOs and their staff (Lawson D, McKay A, and Okidi J, 2005). Duke University (2001) reports The World Bank as recognising that specific strengths of the NGO sector in developing countries include strong grassroots links, field-based development expertise, ability to innovate and adapt, participatory methodologies, and cost-

effectiveness. Commonly cited weaknesses of the sector in developing countries include: limited financial and management expertise, limited institutional capacity, low levels of self-sustainability; and poor internal and external communication.

Attempts to address ICT take up problem include the Uganda-based NGO Forum who are providing CSOs computers and accessories, training, technical support and maintenance, through the Computers-for-Development Programme. The computers will increase the use of modern ICTs and contribute to reduction of digital divide between the NGOs in Uganda and the global community.

Korten (1990) provides a useful contextual framework to study NGOs. He identifies three stages or generations of NGO evolution. First, the typical development NGO focuses on relief and welfare by delivering relief services directly to beneficiaries. NGOs in the second generation are engaged in small-scale, self-reliant local development by building the capacities of local communities to meet their needs through self reliant local action. In the third generation of "sustainable systems development" NGOs try to advance changes in policies and institutions at a local, national and international level; they move away from their operational service providing role towards a catalytic role. The NGO is starting to develop from a relief NGO to a development NGO. The majority of indigenous development NGOs in Uganda are at their infancy stages and therefore at Korten's first or second stage. It is therefore quite appropriate that in order to facilitate the evolution of these NGOs to the second or third stage, an emphasis on exploitation of ICT and attendant organisational change is timely.

Organisation effectiveness is an important concept in evaluating the benefits of ICT adoption. Many authors (Magalhaes, 1999; Walton, 1988) suggest that the following three measures can be used to determine organisational effectiveness at a very broad level: reduced costs as a result of automation, better management of information, more suitable positioning in the competitive market and transformation which encapsulates the benefits accrued from previous stages as well from new management structures and process innovation enabled by new technologies.

Organisations and Technology: Although NGOs are not-for-profit organisations, they still behave as organisations. In order to understand the ICT take up concept, it is important to have good insight into what organisations are and how they behave. Organisation theory offers many different, sometimes conflicting, views of how the phenomenon 'organisation' can be considered. A common assertion is that organisations are highly complex entities dealing with a great number of relevant issues with regard to their creation, existence, functionality and transformation. Major trends in organisational behaviour have been identified by many authors including Skipton (1983) and Mullins (2005). The main organisational theories categorised under Classical, Human Relations, Systems and Contingency are considered to be beyond the scope of this paper. However, the contingency approach is considered as a 21st century paradigm because it is flexible and takes account of organisational circumstances such as the

NGO and grass-roots partnerships. Above authors suggest that the basic concept of contingency approach assists managers to understand complex situations and take appropriate actions.

According to Mullins (2005), the main approaches that have dominated ICT implementation theory include: 1) Technological determinism: This is a technology driven approach that focuses mainly on the application of available technologies to organisational set ups through the use of appropriate methodological tools. 2) Socio-technical interactionism: This is a bottom-up approach that deals with the interaction between structures of the technology and the social structures of the organisation and with the emergent effects arising from such interaction. 3) Socio-economic Shaping of Technology (SST) - Mullins (2005) also states that the focus of SST is upon the ways in which technology is shaped by (rather than itself shaping) the economic, technical, political and social circumstances in which it is designed, developed and utilised. 4) Social Construction of Technology (SCT) views technology as emerging out of non rational-linear process of invention, design, development and innovation. 5) Actor Network Analysis (ANA) is an approach for describing and explaining social, organisational, scientific and technological structures, processes and events. It assumes that all the components of such structures form a network of relations that can be mapped and described in the same terms or vocabulary. 6) Organisational Imperative: This is a strategic top-down approach that creates links between business objectives, business strategy and ICT strategies and implementation.

With the exception of Technological determinism, all the approaches recognise some *interaction* between structures of the technology and the social structures of the organisation and with the emergent effects arising from such interaction (Ciborra, Patriotta and Erlicher, 1995). It is the *interaction* that is of great interest in this paper. By improving the interaction between technology structures (technology imperatives) and social structures (organisational and people imperatives), the technology adoption process can be enhanced. In other words, getting the right Technology, Organisational and People (TOP) imperatives will facilitate ICT implementation. I use the term imperative to denote change drivers or desirable characteristics that facilitate organisational implementation of ICT.

The Model

It is recognised that that none of the above approaches tackles ICT implementation exclusively. As shown in Figure 1, the proposed Technology Adoption Model (TAM) is based on 'Technology, Organisational and People (TOP) Imperatives'. The model is largely driven by the *organisational* imperative in terms of utilising senior management appreciation, appointing them as change champions and aligning technology to business objectives and service delivery. Wider diffusion and use of the website in terms of ongoing updates will depend to a large extent on bottom-up participation from the *people* or staff. The *technology* imperatives are significant in the identification and determination of simple and effective artefacts to aid diffusion.

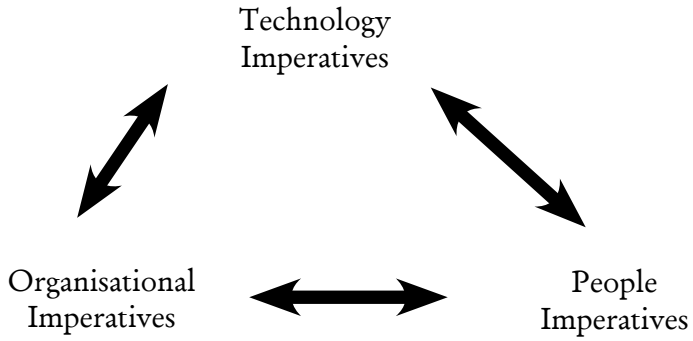
There are three aspects of the imperatives or change drivers identified namely technology, people and organisational. The technology aspect of the model takes an audit of and sets out a plan of how to implement ICT artefacts including specifications, content, architecture, installation, usage and monitoring of the effectiveness of the technology.

The people aspect of the model undertakes ICT skills audit, identifies change champions and implements essential learning for key staff and volunteers and evaluation of change to identify and promote relevant continuing professional development.

The organisational aspect of the model carries out initial organisational ICT strategic review and implements positive alignment of ICT to organisational goals, organisational culture, team structure, attitudes and relationships, budgetary and funding issues, and overcoming resistance to change. Indicators of change such as increased use of emails and accessing information via web are also identified.

With reference to the proposed TOP imperatives and TAM, several relationships can be discerned which ground the model within the socio-technical interactions. The social and technical interaction nature means that some of the TOP features outlined in later in this paper can be categorised under more than one imperative. As an example, some of the people imperatives can arguably be included and reported under organisational imperatives. Learning activities could be such an example.

Figure 1 The TOP-based Technology Adoption Model (TAM)



Methodology

Organisation website was selected as the ICT technology for adoption. The TAM framework as it relates to website adoption was applied and evaluated in two cycles. In the first cycle, an in-depth case study of DTO was undertaken over two year period in order to assess the usability of the model. In the second cycle, the model was applied over six month period to four carefully selected small and medium

sized voluntary organisations in London to assess its 'generalisability'. Eisenhardt (1989) specifically supports the use of case study methodology by arguing that it is particularly well suited to researching new or inadequately researched areas. ICT implementation within the voluntary sector is an area that is hardly researched. Bell (1983) suggests that case study can be appropriately conducted to identify processes that affect a particular situation.

In the UK, ICT has provided a great leap in the way the commercial sector does business. Regrettably, the voluntary and community sector, particularly the Small and Medium Voluntary Organisations, risk being left behind (Tagish 1999). In this research, multiple cases of the pilot (cycle 1) and four (cycle 2) small and medium sized voluntary organisations were selected. Yin (1993) argues for the use of several case studies because they allow for cross-case analysis to be used to build richer theory. This is supported by Hedges (1985) who suggests the use of four to six groups in establishing a reasonable minimum for the predictable replication of a research being undertaken. Peel (2006) also selects five SME case studies in his organisational culture study. As shown in Table 1, the following criteria were used to select the pilot and four case studies: 1) Size: Small and medium sized organisations which are considered as priority organisations. 2) Type of service: representative organisations providing services to typical client groups targeted by the voluntary sector - socially excluded people, older and younger, women and men. 3) Location: based in east London for ease of access to the author. 4) Level of ICT use: Organisations with low or poor use of ICT. To protect sensitive and confidential information, anonymity of the participating organisation is achieved by using unrelated acronyms.

Table 1 Overview of case studies

Organisation	Type of Service	Size (No of staff/ volunteers)	No of PCs
DTO	Community and Economic Development	25	Over 80 all linked to Internet
CO	Childcare provision	16	10, 1 PC for Internet access
TO	Training & Employment	12	30 all linked to Internet
DAO	Community Empowerment and Advocacy	2	14, 1 PC for Internet access
WCO	Training and Community Empowerment	2	3, 1 PC for Internet access

Key staff from the participating organisations completed an exit questionnaire to provide useful assessment of the effectiveness of the model with regards to supporting the website creation and adoption process.

Key Findings from the Case Studies

None of the organisation had a website prior to the intervention. DTO and TO developed strong presence of all three imperatives. DTO with four local area networks created and maintained a popular and well used interactive website. TO with over 30 computers within its Local Area Network also implemented a well used site. Each member of staff has been allocated a computer. In both cases, there exists high level of use of computers to deliver business objectives which include ICT training. DTO and TO are therefore good examples of technologically mature organisations. Sprague and McNurlin (1993) explain the concept of technologically mature organisations as being those in which management is comfortable managing the use of ICT and employees are comfortable using the technology. Furthermore, technologically mature organisations are the ones most likely to take advantage of the new uses of ICT.

From observations and completed exit questionnaires, the participating organisations were able to achieve one or more of the following main characteristics of organisations with high *technology* imperatives are:

- The artefacts (hardware and software) required to implement the ICT/ website are in place
- As a minimum all employees and volunteers should be able to access a computer with internet access on demand.
- Ability to own a network/web server, though not essential, is a distinct advantage for website management and updates
- Specifically with regards to website, access to remote web server to enable regular updates preferably at no additional costs

From observations and completed exit questionnaires, the participating organisations achieved one or more of the following main characteristics of organisations with high *organisational* imperatives are:

- CEOs and Board appreciate strategic role of ICT in improving service delivery and ability to drive through relevant organisational changes
- An ICT strategy
- Learning organisation where training is planned and structured
- Review of individual ICT use is part of staff appraisal system
- ICT budget managed organisationally
- Effective marketing strategy includes use of ICT media such as website

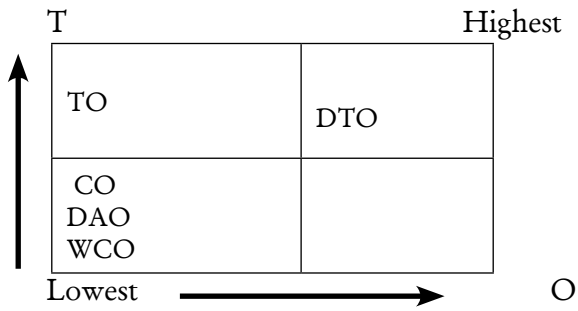
From observations and completed exit questionnaires, the participating organisations were able to achieve one or more of the following main characteristics of organisations with high *people* imperatives are:

- Staff and employees with ICT skills that can enable them to use ICT applications such as Word, emails and accessing websites
- ICT enlightened staff and volunteers who appreciate the benefits of ICT/ website and motivated to embrace change
- Managers willing to act as change champions
- Staff or volunteers with ICT technical expertise such as routine maintenance to provide technical advice, support and training

It is suggested that in order to facilitate the understanding of voluntary organisation in particular and organisations in general, it is helpful to classify organisations using a two-dimensional classification based on TOP schematic diagrams below. Once the status of an organisation has been determined, the challenge is therefore to develop and implement an ICT strategy that moves an organisation to the appropriate top right quadrant.

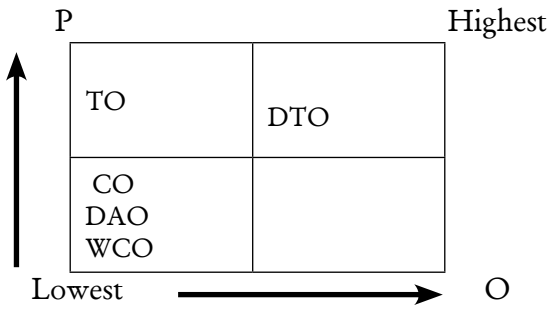
The Technology–Organisational (T-O) diagram (Figure 2) shows the status of an organisation with respect to having a high or low Technology or Organisational Imperatives. In Figure 2, DTO displays the highest T-O imperatives, TO has high T imperative but low O imperatives. CO, DAO and WCO have low T-O imperatives.

Figure 2 T-O Diagram



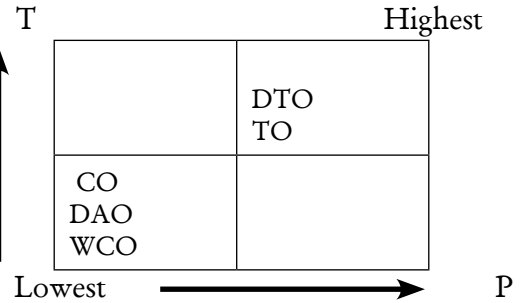
The People–Organisational (P-O) diagram shows the status of an organisation with respect to having a high or low People or Organisational Imperatives. In Figure 3, DTO displays the highest P-O imperatives, TO has high P imperative but low O imperatives. CO, DAO and WCO have low P-O imperatives.

Figure 3 P-O Diagram



The Technology–People (T-P) diagram shows the status of an organisation with respect to having a high or low Technology or People Imperatives. In Figure 4, DTO and TO display the high T-O imperatives, whereas CO, DAO and WCO have low T-P imperatives.

Figure 4 T-P Diagram



The importance of the classifications is that once it is undertaken, the barriers that need breaking down and the actions that must be taken become clearer. The aim is to move an organisation from the low quarter to the high quarter (top right) in the two dimensional diagram. Using this model, for the first time, voluntary/NGO can easily identify a vision and can recognise change required to realise effective ICT implementation.

Exit questionnaires completed by change champions and observations that resistance to change generally arose due to staff feeling ill-prepared, not motivated and competing demands on their time. The best ways of overcoming resistance to change included training the individuals or groups. For example, at CO training was conducted as necessary on Thursdays. Use of senior change champions was also helpful in motivating staff and in some cases change champions were also effective in prioritising the website tasks in the work programmes of the employees. A 'carrot and stick' approach enabled the change process to progress in a timely and planned manner. The approach involves incorporating participation in the technology process in the staff appraisal and monitoring framework and rewarding (non-monetary) staff who achieve set targets.

What lessons for the NGO sector in Uganda?

In the search for a Technology Adoption Model (TAM) that can be applied more widely to NGOs and indeed to most Small and Medium sized enterprises (SMEs), It is suggested that a cautious expansion of the proposed TAM model which was developed in social contexts and settings of voluntary organisations in relatively deprived neighbourhoods in the UK can be made. It is recognised that case study outcomes are perceived as local in nature with the results having relevance within the context of the particular situation. However, a conceptual generalisation can be attempted if its limitations are recognised. A numbers of authors such as Newby (1997) have extended emergent outcomes of such research to wider situations. Extension can be achieved if nominated outcomes can be backed up by literature

that supports such generalisation. In the end, NGOs in Uganda are also voluntary organisations albeit those that disproportionately face the barriers and obstacles identified. A practical application is that the suggested TOP classification can be applied to all organisations. The challenge for managers and change makers will then be to facilitate ICT implementation by improving appropriate TOP features using the TOP schematic classification. An effective strategy for NGO managers is to address inhibiting factors arising out of TOP factors. Key TOP issues that must be addressed in an appropriate organisational ICT strategy are outlined below.

Technology: *Technology* imperatives must be planned for. The artefacts (hardware and software) required to implement the technology must be identified. Accessibility to the technology should be addressed. For example, steps should be taken to ensure members of staff can access and use the technology with ease and efficiency. Access is relatively easy at the larger NGOs but might be quite difficult at the smaller organisations. The concept of ICT access is also closely linked to ICT diffusion. The concept of diffusion as suggested by Sullivan (1985) describes the level of deployment or decentralisation of ICT throughout the organisation. The more accessible the technology, the more decentralised and potentially utilised it is. The research also indicated that a progressive approach is an effective way to implement and increase ICT diffusion. For example, in the case of DTO, the website and its use/sophistication evolved over the years from a simple publishing medium to an interactive and visually attractive site. It is argued that a progressive adoption correlates well with the development of the organisational goals and services. In the five case studies, the sophistication of the website was largely determined by the complexity and capacity of the organisation.

Organisational: Strategies to achieve *organisational* imperatives must also be planned for. The vision and role to be played by the Chief Executive and committee members to initiate and drive through organisational changes must be stated. Steps must be taken to achieve a learning organisation where training is planned and structured. Like at DTO and TO, ICT use by individual members of staff must be incorporated into the staff appraisal process. Financial resources that are managed organisationally must be identified or developed. Finally, an effective marketing strategy must be developed to promote internal and external communication.

People: The various tasks required to implement the technology strategy should be allocated to individual or teams of *people* (staff). For example at DTO, a technology adoption committee was set up to oversee the implementation of the strategy. Another good lesson is that the role of an internal change champion is critical to the successful implementation and diffusion of ICT. The support of a change champion was secured within each participating organisation. Besides internal staff, the roles of external staff or consultants are very important and must be considered. There are contributions and limitations of consultants or 'outsiders' and those appointed must be mindful of embedded organisational culture and

values. Staff training programmes are required to ensure that members of staff and volunteers are provided with opportunities to improve or acquire relevant ICT skills (see learning organisation above).

Donors: Increasingly, there are a number of funding organisations who are seeking to provide capital funding to NGOs to acquire ICT artefacts. This paper shows that technology investment must be accompanied by a structured training programme for the staff and volunteers of the organisation. A clear lesson is that funding organisations must secure not only initial senior management approval for the investment programme but also a training programme to build the capacity of the organisation to ensure effective use of the technology. This situation is particularly necessary for NGOs with low organisational and people imperatives.

This paper contributes generally to guidance on policy and practice of organisational implementation of ICT within the voluntary/NGO sector. The success of a specific case of implementation will depend on a particular diagnosis of the organisational setting where the system will be deployed using the proposed TAM framework. As Markus (1983) also states, a structured understanding of the social setting can help to identify problems or specific needs that have to be addressed in implementing information system development. Different NGOs will be at different state of preparedness and capacity. Therefore, a one-size-fit-all approach will not work. This can be viewed as strength rather than a weakness because different NGOs are specialist providers requiring varying degree of ICT maturity.

In summary, the challenge for technology implementation policy is to classify NGOs according to suggested TOP schematic framework and to improve the relevant aspects of the organisation in order to achieve the TOP characteristics outlined. Government and other agencies can therefore support NGOs to achieve the characteristics associated with high TOP imperatives.

Further Research. Direct applications of the TOP-based model in Uganda and/or other developing countries and extension into other areas of ICT to test the Technology Adoption Model (TAM) are possible areas of further research. Such validations of the model would be beneficial.

References

APEC Centre (2002). What is globalisation? Available at: http://www.globalisationguide.org/01.html> Accessed 10 June 2004.

Bell, J. (1993). Doing your research project. Milton Keynes: Open University

Ciborra, C. U., Patriotta, G. and Erlicher, L. (1996). Disassembling Frames on the Assembly Line: the theory and practice of the new division of learning in advanced Manufacturing. Paper presented at the Information Technology and Changes in Organizational Work, Cambridge, UK.

Davey, S. N. (2005). The Importance of ICT in the Voluntary and Community Sector. Available at. <http://www.egovmonitor.com/node/2404>. Accessed 28 April 2006.

Duke University (2001). Categorizing NGOs. Available at: <http://docs.lib.duke.edu/igo/guides/ngo/define.htm> Accessed 12 November 2006.

Eisenhardt, K. M. (1989). Building theories from case studies, *Academy of Management Review*, **14** (4), pp. 532-550.

Home Office, Active Community Unit. (2003). Voluntary and Community sector Infrastructure – A Consultation Document. Available at: <http://www.homeoffice.gov.uk/documents/2003-cons-voluntary-community>, Accessed 5 January 2004.

Korten, D. (1990). Getting to the 21st century: voluntary action and the global agenda. West Hartford, CT: Kumarian Press.

Lawson D, McKay A, and Okidi J (2005). Poverty persistence and transitions in Uganda: a combined qualitative and quantitive analysis. Available at: <http://www.gprg.org/themes/t5-govrn-norms-outcms/inst-dev-ugan.htm>. Accessed 28 December 2006.

Magalhaes, R. (1999). The Organisational Implementation of Information Systems: towards a new theory. PhD Thesis. London School of Economics.

Markus, M (1983). Power politics and MIS implementation. Communication of the ICM 26 (6), pp. 430-444.

Mullins, L. J. (2005). Management and Organisational Behaviour. 7th ed. Prentice Hall.

The UK Voluntary Sector Almanac (2006): The State of the Sector. Available at: <http://www.ncvo-vol.org.uk/research/index.asp?id=2380&fID=158>. Accessed 18 April 2006.

Newby, M. J. (1997). Educational Action Research: The Death of Meaning? or, The Practitioner's response to Utopian Discourse. Educational Research, 39 (1), pp. 77-86.

NGO Forum (2003). Uganda at a glance. Available at: <http://www.ngoforum.or.ug/about/uganda_glance.htm> Accessed 25 November 2006.

Peel, D. (2006). An analysis of the impact of SME Organisational Culture on Coaching and Mentoring. International Journal of Evidence Based Coaching and Mentoring, 4 (1), pp. 9-18.

Skipton, M. D. (1983). Management and the Organisation. Management Research News, 5 (3), pp. 9-15.

Sprague, R.H. and McNurlin, B. C. (1998). Information Systems Management in practice. Upper Saddle River: N. J. Prentice -Hall.

Tagish, (1999). The Information Society – Meeting the Social and Economic Challenge (1999). Available at:

<http.//www.tagish.co.uk/tagish/pubs/is1/society2.htm> Accessed 3 November 2000.

Walton, R. (1988). Up and Running. integrating information technology and the organization. Boston, MA: Harvard Business School Press.

Willetts, P. (ed.) (1996). The Conscience of the World. The Influence of Non-Governmental Organizations in the UN System, Washington: Brookings Institution and London: Christopher Hurst.

Yin, R.K. (1994). Case Study Research. Thousand Oaks, Calif: Sage.

19
Complexity and Risk in IS Projects: A System Dynamics Approach

Paul Ssemaluulu and Ddembe Williams

In spite of ongoing research on IS risks and the increased sophistication of the tools and techniques developed, IS risks continue to be a challenge to IS professionals and managers. Increased complexity leads to increased risks. When we are confronted with a complex system, our knowledge and understanding of how different components work and interact, and accordingly how the system as a whole works, will always be incomplete. While many researchers have dwelt on project management techniques, it is apparent that we cannot have all the answers in advance since we cannot foretell the future. Due to the increasing complexity of IS solutions it is seen that existing information system development methodologies do not tackle this adequately. The primary purpose of this paper is to highlight how System Dynamics which employs systems thinking can be used to deal with the study of organizations (companies, public institutions, and other human organizations) as complex systems of human activity, with plurality of interest and values. It also shows how System Dynamics models can help companies to manage the risks and uncertainties related to complex IS projects. This paper partly describes some variables in an ongoing research where we aim to use the system dynamics methodology to create a better understanding of the link between information quality and customer satisfaction. We critically look at two variables that we deem important in the search for this relationship. These are complexity and risk in IS projects.

Introduction

The concept of risk is highly visible in any development effort and the best way to deal with it is to contain it. This can best be done by carrying out risk management. Risk management entails identifying risks, analysing exposure to the risks in the development effort and execution of the risk management plan.

There are a number of risks such as the following: Cost overruns, Cancelled projects, High maintenance costs, False productivity claims, Low quality, Missed schedules, and Low user satisfaction.

Inspite of ongoing research on IS risks and the increased sophistication of the tools and techniques developed, IS risk continues to be a challenge to IS professionals and managers. The major driver of risk appears to be the exponential growth in IS complexity and use of IS solutions. Our society is becoming more complex through the use of more complex technologies and organisational forms. This brings about more unpredictability giving rise to systems that are becoming more unpredictable and more unmanageable (Beck, et al., 1994).

Although great advances have been made in implementing information systems, problems still remain (Kenneth and Schneider, 2002). The following are some of the problems facing the development of IS:

- The implementation of IS has often been fraught with uncertainty (Alter and Giznberg, 1978) and have always faced cost and time overruns (Zmud, 1980).
- Resourceful employees (including many young employees) burn out and suffer serious psychological scars as a result of managing projects. Consequently, many of them change jobs and lose the courage they need for project management and the company loses valuable resources (Amtoft and Vestergaard, 2002).
- There is a culture of project management in many organisations that sees it as a sign of weakness and poor management to ask questions or openly acknowledge that you do not have all the right answers (Amtoft and Vestergaard, 2002).
- Traditional proffessional knowledge is not well suited to coping with complex and unique situations. Problem solving as encountered in mathematics and physics brings forward a narrow, technical rationality, emphasizing a rationalist framework for intepreting knowledge. The related problem-solving strategies are too limited in scope (Klabbers, 1996). This is because organisations are information systems within which information is used for decision-making and business process support.

Software development metholodologies attempt to reduce risk by gathering information and using structured processes. It is assumed then that following a good methodology and identifying risk factors, failure could be avoided (Kenneth and Schneider, 2002). However, persistent software failures attest to the fact that there are risks that cannot be overcome by traditional approaches. Therefore, the purpose of this paper is to highlight how system dynamics can be used to create a better understanding of the development and implementation effort of Information systems. System Dynamics models can help companies to manage the risks and uncertainities related to complex IS projects. System Dynamics is concerned with creating models or representations of real world systems of all kinds and studying their dynamics or behavior. The purpose in applying System dynamics is to facilitate understanding of the relationship between the behavior of the system over time and its underlying structure and strategic policies or decision rules (Caulfield and Maj, 2002).

System dynamics has been demonstrated to be an effective analytical tool in a wide variety of situations, both academic and practical and may be a good way to help us understand information systems development and implementation (Williams, 2004). Systems dynamics models are widely used in project management including large scale projects in shipbuilding (Sterman, 1992). System Dynamics involves simulation which is a dynamic representation of reality. During the course

of simulation, the model mimics important elements of what is being simulated. The model is used as a vehicle for experimentation in a "trial and error" way to demonstrate the likely effects of various policies. Those policies which produce the best result in the model will be implemented in real life (Williams, 2004). In such situations, simulation can be an effective, powerful and universal approach to problem solving of systems that would be too complex for mathematical analysis. System dynamics involves interpreting real life systems into computer simulation models that allow one to see how the structure and decision-making policies in a system create its behavior (Forrester, 1999). Simulation allows us to experience the long-term side effects of decisions in just a few minutes.

Challenges of is projects

Despite improved methods for system development and implementation, a number of challenges still exist as discussed in the subsections that follow below:

Requirements Volatility

A lot of emphasis has been placed in the information systems literature on developing complete requirements. Therefore, project managers often believe that gathering complete and consistent requirements can specify a system well enough that risks can be avoided. Unfortunately, correct and complete requirements are difficult for users to specify in systems because of the complexities of systems and limitations in human information processing capabilities (Kenneth and Schneider, 2002). Relying on complete requirement analysis may actually contribute to failure because of overconfidence and because of ignoring risks since requirements change over time (Williams, 2002).

IS Complexity

A complex system is an entity which is coherent in some recognizable way but whose elements, interactions and dynamics generate structures admitting surprise and novelty which cannot be defined in advance (Batty and Torrens, 2001).

Although significant numbers of IS projects are routinely completed successfully, a recent study on the state of IS in the UK carried out by Oxford University and Computer weekly reported that a mere 16% of IS projects were considered successful (Sauer and Cuthbertson, 2003). This is attributed to the increasing complexity of IS solutions and that existing information system development methodologies do not tackle this adequately. This is because such methodologies were developed at a time when IT complexity was at a much lower level, that these methodologies have not scaled regarding complexity (Sauer and Cuthbertson, 2003). In addition to this, new methodologies addressing the growth in complexity have not been developed. IS complexity has grown as the number of components and their integration has increased. This means that the complexity of IS development and use continues to grow substantially. Schneberger and McLean (2003) define complexity as dependent on a system's number of different types

of components, its number of types of links and its speed of change. Increased complexity leads to increased risks. When we are confronted with a complex system, our knowledge and understanding of how different components work and interact, and accordingly how the system as a whole works, will always be incomplete. The components may act and interact in ways we cannot fully predict. Such unpredicatable behavior may cause the complex system as a whole to behave in totally unpredictable ways. This brings about the concept of feedback, which is of course very important.

Many IS projects are designed to improve the operation of business activities that are dynamic, complex, non-linear systems which cannot be readily understood by using static modeling approaches. The dynamic systems are characterized by interactions of closed chains (or feedback loops) that, when combined, define the structure of the system and hence how it behaves over time (Kennedy, 2001). This affects correctness of output and makes it difficult to estimate the exact expenditures and therefore benefits (Marquez and Blanchar, 2004). What has become clear is that people and processes have a greater effect on project outcome than technology (Sabherwal et al., 2005)

Risk

Research in IS shows that risk management is one of the most neglected aspects of project management (Sauer and Cuthbertson, 2003). Risk management involves the definition of hazards that could threaten progress such that earlier problems can be identified, the greater the chance that they be corrected or compesated for with minimal disruption to the project (Sauer and Cuthbertson, 2002). Predicting the future is always a difficult feat especially in today's complex ever-changing world. Unanticipated opportunities and threats can result in catastrophic failures (Vitale, 1986). Projects are said to be successful if they reach their targets of scope, quality, time and cost. However, a project may satisfy these goals but fail because business needs may change between project conception and implementation. A bank may suffer a system failure during an upgrade and have hundreds of thousands of transactions worth billions of dollars being held in suspense (Boyd, 2002). This risk can be managed by using System Dynamics which generates insights into how the whole development process can be achieved, without having to build the real system first. This enables the project manager to stand back and reflect on the project as a whole. Managers and policy makers operate within complex organizations that are riddled with interdependencies, delays, and nonlinearities. Such dynamic complexity challenges decision makers to learn about the underlying causal relationships to make effective decisions and thus minimize risk (Connoly, 1999).

Uncertainty

Pich et al (2002) state that uncertainty is directly related to information adequacy. This means that the better the information quality that the manager receives, the

less the uncertainty. As a direct result of uncertainty, project failures are numerous in practice, there are budget and schedule overruns, compromised performance, and missed opportunities, (Morris and Hugh 1987, Tatikonda and Rosenthal 2000).

Visualization

IS project outcomes are effectively invisible. This visualization problem is a source of many IS project failures. Senior managers may ask for functions that are overambitious, or even impossible to deliver, without having any sense of the level of complexity entailed in meeting their request (Sauer and Cuthbertson, 2003). Dynamic Synthesis Methodology helps in bringing out this visualization in form of models and eventually in the Simulation experiments over time. Using simulation, risks are easily visualized by carrying out sensitivity analysis on the variables used in the modeling process. This helps everyone involved to properly understand the inherent risks at an early stage before implementation. Thus in the extreme, the project would not take off. This helps the Manager get a birds' eye view of the whole project scope before it is implemented.

Causal Loop Diagram

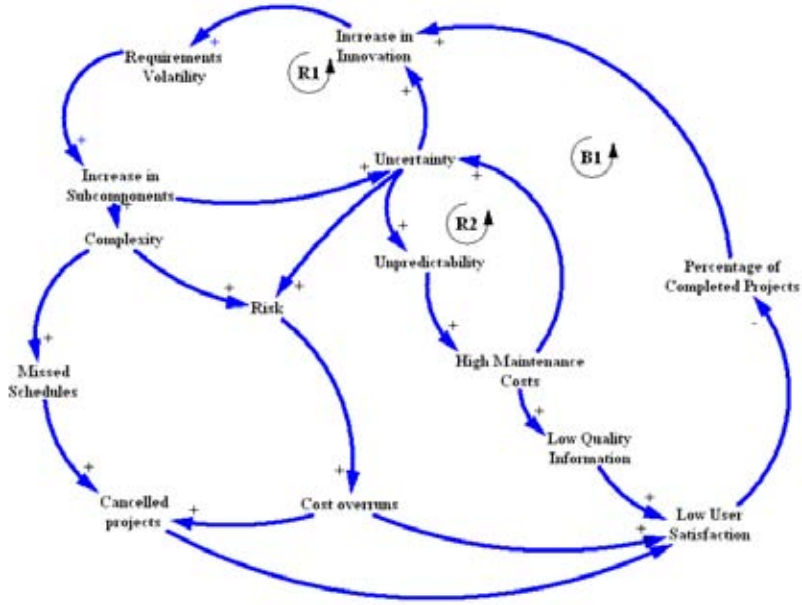
From the literature, we identify the following factors that affect the challenges of developing IS that are identified in the previous section.

- Increase in innovation
- Requirements volatility
- Increase in subcomponents
- Uncertainty
- Unpredictability
- Missed Schedules
- Risk
- Cost Overruns
- High Maintenance Costs
- Low Information quality
- Low User Satisfaction
- Cancelled Projects

It will be realised that factors such as Risk appear as factors since through feedback, risk increases through feedback.

Based on our study of the interaction of these factors, we develop a Causal Loop Diagram (CLD) shown in Fig.1 below:

Fig 1: Causal Loop Diagram of Identified Factors



Based on the CLD presented above, one can deduce that innovations increase over time because of pressures from industry and customers. This in turn leads to an increase in requirements volatility. The requirements volatility leads to an increase in the number of subcomponents in the system and hence contributing to complexity, risk and uncertainty. In like manner this drives the maintenance costs up, cost overruns and may also lead to missed schedules as well as cancelled projects. Uncertainty can lead to low quality information and hence low user/ customer satisfaction and may lower the percentage of completed projects. This in turn feeds back to the system, giving rise to an increase in innovations. The whole process then plays itself out again.

Conclusion

IS risks caused by uncertainty, and complexity leading to cost overruns and low user satisfaction continue to be a challenge to IS professionals and managers. System dynamics can be used to highlight the challenges and create better understanding in order to improve IS project outcomes before they are implemented.

Future work

Future areas of research will involve a field study in form of a quantitative study in two leading Telecommunications firms. To test this case study, another firm will be used for the validation of this study. After this, a system dynamics model will be built to test the dynamics of information quality on customer satisfaction in IS projects.

References

Alter, S.. and Ginzberg, M. (1978). Managing Uncertainty in MIS Implementation. Sloan Management Review. 20 (1):23-31.

Amtoft, A. and Vestergaard, A. (2002). Managing Complexity: Perspectives on Global (Project)Management Competencies. Report of Organisational Psychologists.

Batty, M. and Torrens, P.M. (2001). Modeling complexity: The Limits to Prediction. Center for Advanced Spatial Analysis. Working Paper No. 36.

Beck, U; Giddens, A, and Lash, S. (1994). Reflexive Modernization, Polity Press.

Boyd, T. (2002). McFarlane to Study Executive Role in ANZ system Fiasco. Australian Financial Review. Sydney.

Caulfield, C.G. and Maj, S.P. (2002). A Case for System Dynamics. Global Journal of Engineering Education. 6(1).

Connoly, D.J. (1999). Understanding Information Technology Investment Decision-making in the Context of Hotel Global Distribution Systems: A Multiple Case Study. PhD Dissertation-Virginia State University. November.

Dvir, D; Lipovetsky, S; Shenhar, A. and Tishler, A. (1998). In Search of Project Classification: A Non-universal Approach to Project Success Factors. Research Policy. 27:915-935.

Forrester, J.W. (1999). System Dynamics: The Foundation Under System Thinking. Sloan School of Management. Massachusetts Institute of Technology.

Harkema, S. (1999). Reflections on the Consequences of the Application of Complexity Theory for New Product Introductions. Report of Nyenrode Institute, University of Nyenrode, The Netherlands.

Klabbers, J.H.G. (1996). Problem Framing Through Gaming: Learning to Manage Complexity, Uncertainty, and Value Adjustment. Simulation and Gaming. 27(1):74-92.

Kennedy, M. (2001). The role of System Dynamics Models in improving the Information Systems Investment Appraisal in respect of Process Improvement Projects.Proceedings of Nineteenth International System Dynamics Conference. Atlanta, Georgia, USA.

Kenneth, R.W. and Schneider, H. (2002). The Role of Motivation and Risk Behavior in Software Development Success. Information Research 7 (3). April.

Markus,M.L. and Tanis,C. (2000). The Enterprise Systems Experience-From Adoption to Success, In Framing the Domains of IT Research:Glimpsing the Future Through the Past.

Marquez, A.C., and Blanchar, C. (2004). A Decision support System for Evaluating Operations Investments in High-Technology Systems. Decision Support Systems. DESCUP-11036.

McFarlan, F.W. (1981). Portifolio Approach to Information Systems. Harvard Business Review. 59(5):146.

Morris, P. W. G., G. H. Hugh. 1987. The Anatomy of Major Projects. Wiley, Chichester, U.K.

Pich, M.T; Loch, C.H. and De Meyer,A. (2002). On Uncertainty, Ambiguity and Complexity in Project Management. Management Science. 48(8):1008-1023. August.

Sabherwal, R; Jeyaraj, A; Chowa, C. (2005). Information Systems Success: Dimensions and Determinants. Invited Presentation, College of Business Administration, University of Illinois, October.

Sauer, C. and Cuthbertson, C. (2003). The State of IT Project Management in the UK. Report of Templeton College, Oxford University. November.

Schneberger, S.L. and McLean, E.R. (2003). The Complexity Cross: Implications for Practice. Communications of the ACM 46(9):216-225. September.

Stanley, H. (2001). Service Leadership on the Edge of Chaos. MBA Thesis, University of Nyenrode, Breukelen, The Netherlands.

Sterman, J.D. (1992). System Dynamics Modeling for Project Management. Sloan Management Review.

Symons, V.J. (1994). Evaluation of Information Systems Investments: Towards Multiple Perspectives. Chapman and Hall, London. ISBN 0-412-41540-2.

Tatikonda, M. V. and Rosenthal, S.R.(2000). Technology Novelty, Project Complexity, and Product Development Execution Success. IEEE Transactions. Engineering. Management 47: 74–87.

Vitale, M. (1986). The Growing Risks of Information Systems Success. Management Information Systems Quarterly, 10(4): 327-334. December.

Williams,D. (2002). An Application of System Dynamics to Requirements Engineering Process Modeling. PhD Thesis, London South Bank University.

Williams, D. (2004). Dynamics Synthesis Methodology: A Theoretical Framework for Research in the Requirements Process Modeling and Analysis. Proceedings of the 1st European Conference on Research Methods for Business and Management Studies. Cambridge.

Zmud, R.W. (1980). Management of Large Software Development Efforts. MIS Quarterly. 4. (1):45-55.

20
A Visualization Framework For Discovering Prepaid Mobile Subscriber Usage Patterns

John Aogon and Patrick J. Ogao

Telecommunications operators in developing countries are faced with a problem of knowing their prepaid subscriber usage patterns. They have a challenge of reducing prepaid churn and maximizing the lifetime value of their subscribers. The prepaid subscriber is anonymous, and the only way a prepaid subscriber gives information to the operator is through call records of events on the use of the telecommunications network. Thus, the call details in their raw form do not provide any useful information. In addition, these details provide an overwhelming amount of data that is not easy to analyze. To assist the telecommunications operators, this study undertook to develop a visualization framework that helps telecommunication operators discover prepaid subscriber usage patterns. An exploratory approach was used to unravel subscriber usage patterns from call data records obtained from a local telecommunication operator in Uganda. Five visualization tools that were selected based on their functionalities. Based on the findings, a visualization framework for discovering subscriber usage patterns is presented. The framework is evaluated using call data with known knowledge obtained from the local telecommunication operator. Results outline the strengths of various visualization techniques as regards to specific prepaid usage patterns.

Introduction

The growth in prepaid mobile subscribers in the telecommunication industry is increasing at a fast rate. Up to 90 percent of mobile subscribers are prepaid (Meso *et al*., 2005). Equally, the turnover of these subscribers, as they cease to use mobile services or switch to a competitor is quite high. This migration is referred to as churn in the mobile telephone industry. Information available indicates that up to half or more prepaid customers are likely to change operators in a 12-month period. To reduce churn, operators are starting to rethink their prepaid strategies. In this respect loyalty and customer care programmes specifically tailored for prepaid customers are critical in the prepaid mobile market. In order to minimize churn it is imperative that potential "churn" is identified in advance (Karen, 2004). Tailoring specific loyalty and customer care programmes for these subscribers is therefore an important business requirement for all mobile telecommunication operators.

The challenge is that, most telecommunication companies do not have sufficient information about their prepaid subscribers (Shearer, 2004). This is despite, most

prepaid subscribers giving the operator information about themselves through recorded events on the use of the telecommunication network - Call Detail Records (CDRs). The raw CDRs do not provide immediate useful information to the telecommunication operators. Turning this raw CDRs into giving significant insights of customers and markets is the main challenge.

Therefore this study aims at using the CDRs to enable the telecommunications operators overcome this anonymous nature of prepaid subscribers, discover valuable information and gain insight of its subscribers.

Methodology

An exploratory research approach that utilized the visualization techniques and interactive functionalities to unravel hidden patterns from detail call data from a local telecommunication company was used.

Visual Data mining

Visual data mining aims at integrating the human in the data analysis process, applying human perceptual abilities to the analysis of large data sets available in today's computer systems ((Bustos *et al*., 2003).

The basic idea of visual data mining is to present the data in some visual form, allowing the user to gain insight into the data, draw conclusions, and directly interact with the data (Chung, 1999).

Visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps.

Visual data exploration often follows a three step process: Overview first, zoom and filter, and then details-on-demand (Shneiderman, 2002). In other words, in the exploratory data analysis of a data set, an analyst first obtains an overview. The data analyst directly interacts with the visualizations and dynamically changes the visualizations according to the exploration objectives. This may reveal potentially interesting patterns or certain subsets of the data that deserve further investigation. The analyst then focuses on one or more of these, inspecting the details of the data (Shneiderman, 1998).

Five different visualization tools were selected for this study based on their encompassing spectral capabilities. The tools included both 2-dimensional and Multi-dimensional display techniques. This included the use of scatterplots suitable for identifying trends and outliers, parallel coordinates that are suitable for viewing subscriber mobility, outliers, relationships, barcharts, piecharts, that are good for ranking patterns like high and low usage subscribers.

Requirements Analysis

Analysis of the business requirements was done by reviewing existing reports and interviewing various users in the local telecommunication company. This was done to identify and clarify the user needs that were subsequently used to understand and identify the data required to meet these needs. The billing systems

proved to be the rich and convenient source of data used in this study because it stores online historical data of up to four months in an oracle database.

Table 1 shows statistical text reports that are currently generated and the questions that need to be answered.

Table 1. Reports generated from the call records.

Text Report	Questions that Need to be Answered
Calls by tariff plan	Which is the most used tariff plan? Which customer group is highly profitable, which one is not?
Calls by subscriber	Which subscribers are high users? Which subscribers are low end users? To which customers should we advertise speciall offer?
Calls by location	What is the likely home location of the subscriber? Which locations have low and high traffic?
Calls by destination	What is the most called destination? Which are the least called destination?
Calls by time of day	Which subscribers are business users? Which subscribers are personal users?
First and last call by subscriber	Which subscribers have just joined the network? Which subscribers have left the network? Which subscribers are about to leave the network?
Daily call summary	What is the call usage trend? Which days have low and high usage? Is there anything unusual?

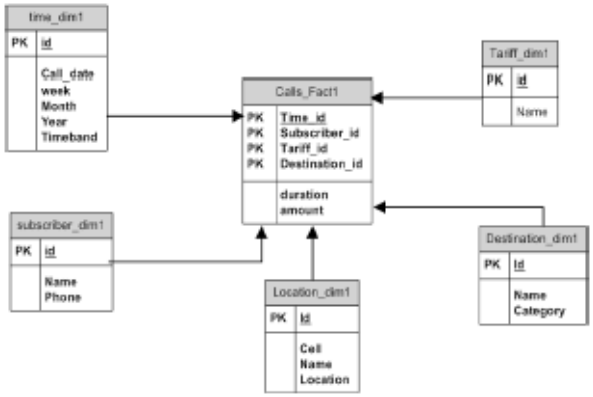
Developing the Data Mart

The requirements analysis stage identified the following key call detail attributes;

- Originating number - the calling number.
- Tariff plan - tariff plan associated with the subscriber at the time of making a call.
- Terminating number - call destination (called number).
- Call time - date and time of starting the call.
- Location - network location where the call was made.
- Duration - the duration of the call in seconds.
- Charge - the charge for the call.

The data requirements identified above were transformed into dimensions that constitute a model for the data mart (Figure 1).

Fig 1. Data Mart Dimension Model

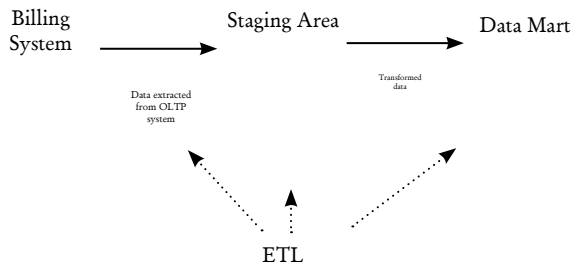


Creation of the Database

Oracle RDBMS was used to implement the above model. To extract, transform and load the data into the data mart, SQL scripts were written. The data was extracted from the billing system (OLTP system) and loaded to the staging area where cleaning and transformation took place (Figure 2).

The following details for each call were extracted; calling number, called destination, location, call date, tariff plan, duration of call and the charge of call. After cleaning and transforming the data, it was loaded into the data mart in summarized form. The day was chosen at the lowest granularity of data. The summarized data was used because the size of the call details data stored in the billing system was large.

Fig 2: Data Mart Architecture



Developing a Visualization Framework

Fayyad *et al*. (1997) describe data mining as a collection of powerful analysis techniques for making sense out of very large datasets. There is no one data mining

approach, but rather a set of techniques that often can be used in combination with each other to extract the most insight from the data. This study used visual data mining approach to gain insight into the usage patterns hidden in the subscriber call detail records.

Visualization Tools

The following five different visualization tools were used in this study.

a) Eureka - (http://infovis.cs.vt.edu/demos/)
b) Advizor (http://www.visualinsights.com/)
c) GGobi - (http://www.ggobi.org/)
d) XmdvTool (MultivariateData Visualization Tool) – (http://davis.wpi.edu/xmdv/)
e) Omniscope – (http://www.visokio.com/)

The interactive techniques that are imbued in these tools included, *filtering, brushing, sorting, identification, grouping* and *linked (multiple) views*. Each tool was used to explore the CDRs and observing the different visual views presented by each tool, identify the technique suitable for visualizing particular kind of patterns.

Suitable technique to reveal specific patterns in the CDR data was established based on the guidelines detailed below (Pillat *et al*., 2005). Based on these guidelines, each tool and the techniques used were assessed for suitability for use in telecommunications subscriber usage pattern discovery.

a) Ability to reveal patterns in call records data. The purpose of the visualization tool is primarily to gain knowledge through the recognition of patterns in the data. In this respect, the ability of a visualization technique to provide answers to questions such as those in table 1.
b) Ability to visualize massive telecommunications call data
c) Support for multi-dimensional data. The call data typically contains several dimensions such as subscriber, tariff plan, time, location, destination, the visualization technique must be able to display many dimensions in a single display.
d) User interaction - To support the user to gain insight of the displayed data, the technique should provide the ability to dynamically manipulate the display.
e) Easy to use - The visualization tool should be ease to use in terms of loading data in different formats; presenting clear displays that are easily perceived by users.

The figures that follow below are examples of exploring the CDRs for patterns using the five visualization tools. These are meant to highlight the out put of techniques used by the tools.

Fig 3: Table lens Shows the Focused Details and the Rest of the Display Remains in Context.

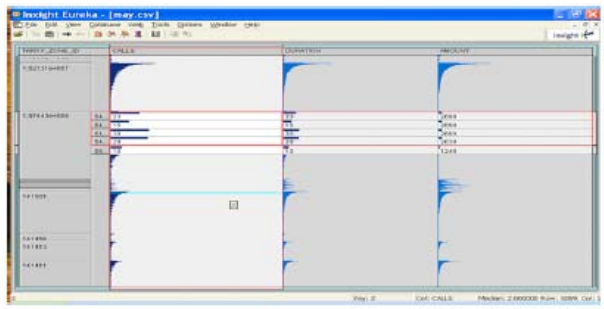


Figure 3 shows the table lens tool display of data, using focusing interaction technique. Details of a particular point of interest in the display can be revealed.

Fig 4: XmdvTool Parallel Coordinates Display of Call Data with brushed data points

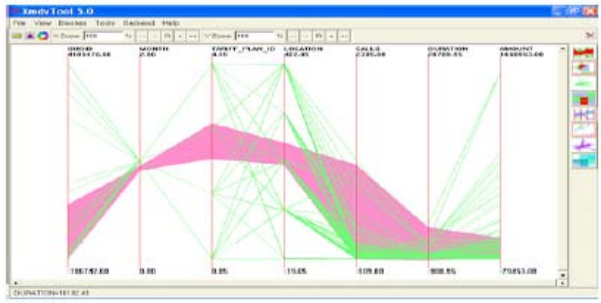


Figure 4 shows the XmdvTool brushing technique used for linking multiple displays of the same data. One of the displays is used to select the data elements of interest by brushing the points with color and immediately the corresponding elements are highlighted on the other existing displays.

Fig 5: GGobi Scatterplot of Subscribers by Call Date

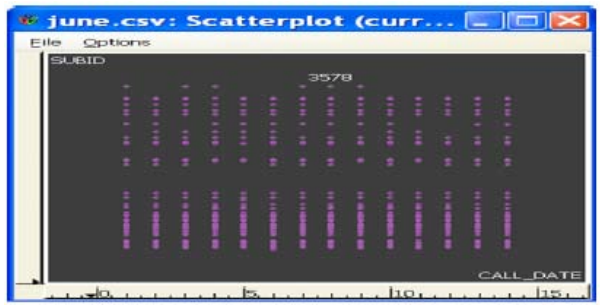


Figure 5 shows the Ggobi scatterplot display for identifying churn and joiners. In the display the dot represents the subscriber that made calls on a given day.The display is useful for telling the subscribers who have left the network and those who joined for the period in consideration.

Fig 6: Omniscope Filtering Technique Used to Select Data

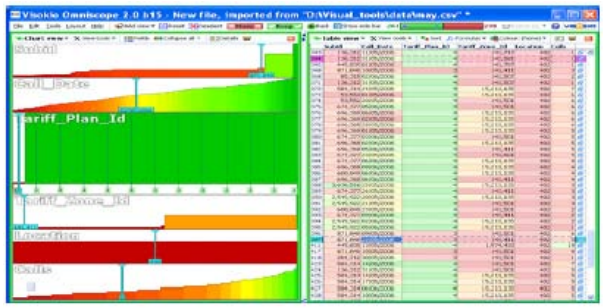


Figure 6 shows Omnicscope tool filtering interaction technique used to graphically select data of interest. The display shows the filtered data by selecting on the chart view only tariff plan 4. The table view is automatically updated to show only tariff plan 4 details.

Fig 7: Advizor Multiscape(3D) Display of Calls for each Subscriber by Tariff Plan and Location

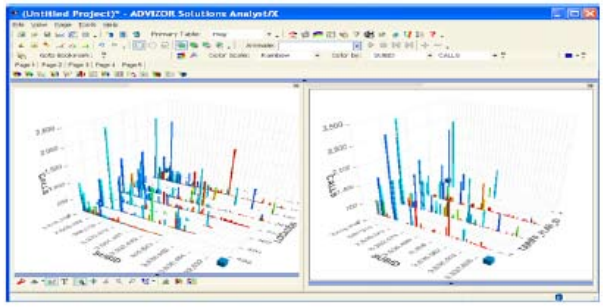


Figure 7 shows Advizor tool ability to display multiple displays on one page. The 3D display are useful in analyzing the relationship between three variables. For example displaying the subscriber calls by location and tariff plan on the same page, it possible to tell the most likely home location of the subscriber and the most common tariff plan used by subscribers.

Validation of the Framework

The validation of the framework is a testing phase, in which the developed visualization framework was applied to real world telecommunications call data obtained from the local telecommunication company.

Table 2. Summary of visualization tools Display and Interaction Techniques

	Visualization Tools				
Display Technique	Eureka	XmdvTool	GGobi	Omniscope	Advizor
Bachart			Y	Y	Y
Line Graph				Y	Y
Scatterplot		Y	Y	Y	
Scatterplot Matrix		Y	Y		
Time Series			Y		Y
Histogram	Y			Y	Y
Table lens	Y				
Parellel Coordinates		Y	Y		Y
Multiscape (3D)					Y
Interaction Techniques					
Linking/Brushing	Y	Y	Y	Y	Y
Focusing	Y				
Sorting	Y			Y	Y
Grouping	Y			Y	Y
Linking/Coloring		Y	Y	Y	Y
Identifying	Y	Y	Y	Y	Y
Filtering	Y			Y	Y
Other Features					
Concurrent Multiple Displays		Y	Y	Y	Y
Database Connectivity	Y			Y	Y
Data Size	Large	Small	Small	Large	Large
Usage	Commercial		Free	Free	Commercial

Table 3 represents the techniques and the patterns. The rows represent visualization techniques and each column represents a pattern of interest in the call records data. Again,"Y" in the column signifies that yes, the visualization technique can be used to reveal the pattern satisfactorily. A blank signifies "Not Applicable". The results show that techniques such as parallel coordinates and multiscape can reveal highest number of patterns in the call records. Scatterplot, Time series and multiscape can reveal churn, one of the most interesting patterns in telecommunications.

Table 3. Visualization Technques and Patterns

	Patterns						
Visualization Technique	Outliers	Churn & Joiners	Subscriber Mobility	High & Low Using subscribers	Most Used Tariff Plan, Location	Correlations	Non Obvious patterns
Bachart				Y	Y		
Pie Chart				Y	Y		
Scatterplot	Y	Y				Y	
Time Series	Y	Y					
Line Chart	Y			Y	Y		
Table lens	Y			Y	Y		
Parellel Coordinates	Y		Y	Y	Y	Y	
Multiscape (3D)	Y	Y	Y	Y	Y	Y	
Interaction Techniques			Y				Y

Results

The proposed framework is based on the premise that a visualization tool consists of two techniques; display and interaction techniques. The display techniques present the graphical display of data and range from common visual displays to special purpose displays.

Interaction techniques are for manipulating the data in the displays. These techniques are used to pose graphical queries on the data so that further insight

can be gained about the data. In this study, display techniques have been divided into two categories:- specialized and the traditional or common techniques.

Based on the number of attributes that can be displayed, the specialized techniques are further divided into two; multidimensional and two-dimensional.

Multidimensional display techniques: Are those that can display more than two variables (attributes) at the same time, for example multiscape (3D), parallel coordinates and Table lens. Multidimensional display techniques can reveal two types of patterns using the telecommunications call data, namely;

a) ***Ranking patterns:*** These are patterns that either rank or categorize the attributes. For example: most called destination, tariff plan with highest and lowest number of subscribers, locations with high and low number of calls, high and low using subscribers.

b) ***Unique patterns:*** These are patterns that are not easily seen with text data or reports. These include; churn - subscribers who have left the network, Joiners new subscribers that have just joined the network, changes in subscriber profiles (subscribers who keep changing tariff plans), subscriber mobility from one location to another.

Two Dimensional Techniques: Are those that can display only two variables (attributes) at the same time, for example time series plots, scatterplots, and line graphs. The time series techniques reveal trends or changes with time and therefore in telecommunications this technique is useful for identifying the following; churn - subscribers who have left or about to leave the network. Joiners - subscribers who have just joined the network. Outliers exceptions that need not to be ignored but investigated. Line graphs can reveal trends, for example subscribers who have stayed longer on the network are generally high users as compared to new subscribers. Scatteplots can reveal relationships between attributes that need to be investigated further. In addition scatterplots can be used to tell churn, joiners and outliers. The regular or traditional techniques (barchart, piechart) are also display techniques that can be used for revealing ranking patterns in telecommunications.

Interaction Techniques: Are used to manipulate the displays presented by the display techniques. Although in visualization the display techniques can immediately tell facts about the data being displayed, interaction techniques are used to uncover patterns that cannot easily be seen on a single display of data that is gaining insights into hidden patterns.

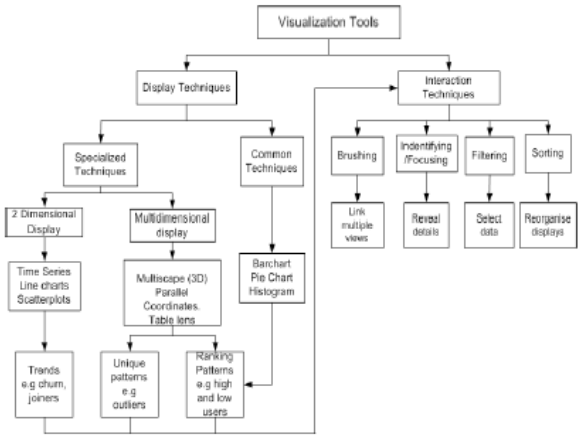
These techniques include brushing, focusing, filtering, labeling, and sorting. In order to benefit from these techniques, more than one display of the data is used. Interaction techniques help pose questions on the data being displayed and generate answers to these questions. Using telecommunications call data records, below are typical questions that can be answered with the help of interaction techniques;

- Where do high using subscribers originate their calls from?
- Which subscribers should be targeted with advertising?

- Where do the high users call most?
- Which tariff plan has the greatest number of low users?

After exploring the call data records for patterns using visualization tools as described in the previous section, a framework for visualizing subscriber usage patterns in telecommunications is presented in figure 7.

Figure 7. Framework for Visualizing Subscriber Usage Patterns in Telecommunications



Discussion

Visualization approach is a preferred alternative to automated data mining approaches. The results show that, with exception of Eureka, the five visualization tools have almost the same techniques; the only significant difference is the size of the data set, and the data formats supported by each tool.

Consequently, a framework for visualizing usage patterns in telecommunications is described. The framework shows techniques and the patterns each technique is suitable for in telecommunications subscriber call records. The framework classifies the visualization techniques into two; the specialized techniques and common techniques. The patterns in the telecommunications call records have also been divided into three; trending patterns, ranking patterns and unique patterns. The results also show that Omniscope, XmdvTool and GGobi are limited in the size of data set they can support. They are not robust as Advizor and Eureka.

Conclusion

In this study we have suggested a visualization framework that serves as a model for discovering prepaid usage patterns using existing visualization techniques. The paper has described a range of existing visualization techniques; simple patterns can be seen on just a single display, but hidden patterns can be discovered by using multiple displays together with interaction techniques. This work has further demonstrated that the visualization techniques cut across all the application domains with varying degrees of strengths and weaknesses.

An interesting area of study is to find out the relationships between the subscriber calling patterns and their airtime loading patterns. This study did not use the subscriber demographic data due to its unavailability from the local telecommunications operator whose data was used in the study. Calling patterns with supporting demographic data can make telecommunications operator gain more insight of their subscribers.

References

Bustos, B., Keim, D. A., Panse, C., Schneidewind, J., Schreck, T., Sips, M. and Wawryniuk, M. (2003). Pattern Visualization. http://dke.cti.gr/panda/tasks/deliverables/DLV-2-3.pdf. (Accessed: March 29, 2006).

Chung,W. P. (1999). Visual Data Mining. http://www.pnl.gov/infoviz/visual_data_mining.pdf. (Accessed: March 30, 2006).

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (1996). From data mining to knowledge discovery in databases. AI Magazine, Fall 1996, pp. 37-54.

Meso, P., Musa, P. and Mbarika, V. (2005). Towards a model of consumer use of mobile information and communication technology in LDCs: the case of sub-Saharan Africa. Information Systems Journal. Vol. 15(2): pp119-146.

Karen, G. S. (2004). Customer-Centered Telecommunications Services Marketing. London: Artech House.

Pillat, R. M., Valiati, E. A. and Freitas, C. M. D. S. (2005). Experimental Study on Evaluation of Multidimensional Information Visualization Techniques, *Proceedings of the 2005 Latin American conference on Human-computer interaction*, Cuernavaca, Mexico.

Shearer, C. (2004). Anticipating Consumer Behavior with Analytics. http://www.crm2day.com/library/

(Accessed: February 8, 2006).

Shneiderman, B. (1998). The Eyes Have It: User Interfaces for Information Visualization. *http://hci.stanford.edu/cs547/abstracts/97-98/980220-shneiderman.html*. (Accessed: March 28, 2006)

Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining, *Journal of Information Visualization*, **1**(1): 5-12.

PART 4

Data Communication and Computer Networks

27
The Diminishing Private Network Security Perimeter Defense

Prof. Joseph M. Kizza

The 3rd Annual International Conference on Sustainable ICT Capacity in Developing Countries - SREC07, August 5-8, 2007, Makerere University, Kampala, Uganda

Introduction

It is difficult to believe that just a few years ago, no one would worry about the security of a computer or a computer network for that matter. Computer security literally meant protecting your computer from thieves. Then the landscape changed and computer after computer become interconnected forming computer networks as we know them today. As this was happening, our personal attributes were finding their way into the computer and eventually into the computer network. Computer networks multiplied, merged and grew bigger resulting into this one huge global super network that seem to include every connected computer around the globe. This global super network is actually a mosaic of thousands of smaller public and private networks of varying sizes. Because of the interconnection and interdependence of these networks, what happens in one network has the potential to affect every network and every computer in the super global network.

To talk about the security of this super global network, it suffices to talk about the security of anyone of the contributing smaller network since this super network is as secure as the weakest member network. Because of this fact, we will focus our discussion in this paper on a company network we will refer to as a private network.

Today, the security of a company is measured by the security of its network. The life of a company from its employee records to its daily transaction records is stored on this private network. Company networks have become the backbone of and the weakest links in the company's security. Because of this realization, company networks, have over the years, been protected by a fortified security perimeter consisting of firewalls, DMZs, intrusion detection and intrusion prevention sensors and many other types of sensors the purpose of which is to stop and prevent any unauthorized penetration into the company network. The perimeter defense has been working well in many networks. Unfortunately, due to the changing technology landscape and other forces, this perimeter defense is

crumbling with serious consequences. The Diminishing Private Network Security Perimeter Defense discuses the causes of this erosion and suggests workable solutions.

Forces Weakening the Private Network Security Perimeter Defense

There is no one particular force that one can point at as the sole cause of this weakening process but a combination of many forces most of them caused by the rapid advances in computer technology, the plummeting prices of information processing and indexing devices, and the development of sprawling global networks. All these are creating an environment where the generation, collection, processing, indexing, and storage of and access to information is made easy. In addition, the ease of use has brought users of all character shades on the global super network, thus making it a real wild west where laws are followed at will and only the stronger and most ruthless survive. Let us look at some of these forces:

Proliferation of Malware

Although the practice of security for the private network has been constant for a good number of years, all based on a few well-known standard protocols and best practices, the new and changing technologies have been constantly cheeping away on the private network security perimeter defenses through a flotera of software tools commonly referred to as malware. Malware is an umbrella term for a collection of unfriendly software tools that include viruses, worms, Trojan horse, key loggers, dialers, stealth passwords, password crackers, sniffers, web traffic generators, advertisement popups, exploits, rootkits, botnets and zombie agents [7]. Malware has become one of the most effective force in degrading the private network security perimeter defenses.

According to a study from Sophos, an antivirus and anti-spam company, there was a sharp increase in the number of malware from 9,450 in 2006 to 23,864 new threats in the first three months of 2007 [11]. According to Sophos, the majority of these new threats are now being embedded in malicious Web sites even though historically malware have been plaguing e-mail and attachments. This change is a worrisome development because many company websites on company private networks are falling victims.

Among the most popular website based malware is currently the Trojan Fujif, which according to Sophos report, accounted for 50.8% of all the malware hosted on Web sites in the first quarter of this year. 12.8% of the website checked were hosting malicious script, Windows malware was responsible for infecting 10.7%, Adware was found on 4.8% of these pages, and porn dialers on 1.1% [11].

Emanation and Emission

Identity theft, the stealing and malicious use of personal information for personal motives, is now the fastest growing crime of the digital age. This is a result of rapid growth in information gathering technologies such as surveillance, snooping,

and sniffing. It is a new "gold rush" of the information age by hackers and other unscrupulous individuals. A lot of this information is gathered by hackers as a result of emanation and emissions from electronic devices. Emanation is any modulated signal such as sound or electromagnetic radiation, leaking from a device that may be used to reconstruct information being processed or transmitted by that device. Emission on the other hand is the giving off of something like gas.

Many data communication equipments, and even data encryption devices, sometimes emit modulated optical signals that carry enough information for an eavesdropper to reproduce the entire data stream. It requires little apparatus and it can be done at a considerable distance, and is completely undetectable [10]. Even flickering light from a common screen reflected off a wall according to one study, can reveal whatever appears on the screen of a PC monitor. Also blinking lights and other optical signals from the little flashing LED lights dotting everything from modems to network cards, routers and keyboards, can be captured with a telescope or long-distance lens and processed to reveal all the data passing through the device[10].

All electronic and electric devices when in use, give off information in a form of radiated electromagnetic signals. When this happens, the intruder near by just needs to intercept the signals with an appropriate recorder. Radiated signals may carry actual information that attackers may want to capture and recreate some or all of the original information.

There several emanation and emission sources including:

- Device Emanations
  - All video displays such as CRTs and LCDs emit a weak TV signal.
  - All cabling, including serial cables using by ATMs and Ethernet cable used by PCs.
  - Keyboard RF or Bluetooth keyboards give off emissions modulated by pressing a key.
- Power line leakage.
  - Power circuits pick up RF signals from nearby devices and circuits and conduct them to neighboring buildings.
- Sound.
- Smartcards such credit-card with embedded microprocessor or memory emit signals when in use.

Emanations and emissions leakages can be captured either passively or actively. In passive capturing, techniques such as wardriving and electromagnetic eavesdropping are used. In both wardriving and electromagnetic eavesdropping, the recorder is set up in a slow moving or parked vehicle. The record can also wirelessly captures wireless broadcast signals. Several techniques are used in active signal capturing, including:

- Using specially designed software commonly referred to as TEMPEST viruses which includes malware that scans infected computer for desired information, which it then broadcasts it via RF signals.
- Using phones near transmitters can cause data to be modulated by the phone and transmitted.
- Listening to high frequency signals on connected cables like power lines and electromagnetic radiation leaked from computer devices.
- Disrupting computations by inserting power glitches.
- Causing glitches or jamming by inserting transients into power or clock signal to induce useful errors.
- RF Fingerprinting by identifying RF device based on the frequency behavior.
- Radio Direction Finding (RDF) by locating the direction to a radio transmission.
- Traffic analysis using cryptanalysis techniques to capture and analyze data ever a period of time
- Signals collection of different signals and extracting information from them.
- Sniffing monitored traffic.

Theft

A considerable amount of information is lost via theft of digital devices including laptops and memory sticks. Company secrets including network access information is always lost along with these devices.

Advances in Wireless Ad-hoc Network Technology

A wireless ad-hoc network is a computer network where each node communicates with each other through wireless links. The network is ad-hoc because each node acts as a router to forward packets of other nodes, and as such the routing determination from source to destination is made dynamically based on the network connectivity. Mobile Ad-hoc Networks (MANeTs) are extremely helpful in supporting and forming an instant, self-organizing and infrastructure-less network.

Wireless ad-hoc routing protocols are designed to discover and maintain an active path from source to destination. Many protocol designs, such as Dynamic Source Routing (DSR), Ad hoc On Demand Distance Vector (AODV), and Greedy Perimeter Stateless Routing (GPSR), are defenseless protocols because they assume that all wireless nodes will follow the specified protocols in a benign environment. However, some wireless nodes may deviate from protocols in various ways, and traditional network security solutions are not applicable to wireless networks due to the lack of physical boundaries. Also defenseless protocols are not able to handle misbehaving nodes during packet delivery. Reliability and security concerns, therefore, are abound in MANeTs because they can create opportunities via misbehaving nodes.

The Peripheral Loopholes

Modern computers use a large collection of peripheral devices to provide communication and storage. These devices, usually optional hardware pieces to enhance the computer's functionalities, interact with the operating systems kernel which controls them via their drivers. Based on the history of viruses, peripheral devices have been used to propagate viruses. For example floppies were used to spread viruses. Also iPods have been known to spread security threats. USB devices have also been shown to have high risks of disclosing of proprietary information. Also certain functionalities of USB devices such as storage have been shown to hide functionalities such as keyboard and mouse that can be automatically used to launch an attack when connected to another computer.

Open Architecture Policy

Cyberspace infrastructure was developed not following a well conceived and understood plan with clear blueprints, but it was developed in steps in reaction to the changing needs of a developing intra and inter communication between computing elements. The hardware infrastructure and corresponding underlying protocols suffer from weak points and sometimes gaping loopholes partly as a result of the infrastructure open architecture protocol policy. This policy, coupled with the spirit of individualism and adventurism, gave birth to the computer industry and underscored the rapid, and sometimes motivated, growth of the Internet. However, the same policy acting as a magnet has attracted all sorts of people to develop exploits for the network's vulnerable and weak points, in search of a challenge, adventurism, fun, and all forms of personal fulfillments.

Lack of Basic Trust Relationships

Compounding the problem of open architecture is the nature and the working of the communication protocols. The Internet as a packet network works by breaking data, to be transmitted into small individually addressed packets that are downloaded on the network's mesh of switching elements. Each individual packet finds its way through the network with no predetermined route and is used in the reassembling of the message by the receiving node. Packet networks need a strong *trust relationship* that must exist among the transmitting nodes. For example, if one node is malicious or misbehaving, there is no way to deal with such a node and no way to inform other nodes in the network about it.

Such a relationship is actually supported by the communication protocols like three way handshake. The three-way handshake establishes a trust relationship between the sending and receiving nodes. However, network security exploits like IP-Spoofing, SYN-Flooding and others that go after infrastructure and protocol loopholes do so by targeting and undermining this very trust relationship created by the three-way handshake.

Distributed Denial of Service Attacks (DDoS)

Distributed denial of service (DDoS) attacks are generally classified as nuisance attacks in the sense that they simply interrupt the services of the system. System interruption can be as serious as destroying a computer's hard disk or as simple as using up all the system available memory.

Although we seem to understand how the DDoS problems do arise, we have yet to come up with meaningful and effective solutions and best practices to deal with it. What makes the search for a solution even more elusive is the fact that we cannot even notice that a server is under attack since the IP spoofing connection requests, for example, may not lead to a system overload. While the attack is going on, the system may still be able to function satisfactorily establishing outgoing connections. Which brings one to wonder how many such attacks are going on without ever being detected, and what fraction of those attacks are ever detected.

Network Operating Systems and Software Vulnerabilities

Network infrastructure exploits are not limited to protocols. There are weaknesses and loopholes in network software that include network operating systems, web browsers, and network applications. Such loopholes are quite often targets of hacker aggressive attacks like planting Trojan horse viruses, deliberately inserting backdoors, stealing sensitive information, and wiping out files from systems. Such exploits have become common.

Limited Knowledge of Users and System Administrators

The limited knowledge computer users and system administrators have about computer network infrastructure and the working of its protocols does not help advance infrastructure security. In fact it increases the dangers. In a mechanical world where users understand the systems, things work differently. For example in a mechanical system like a car, if such a car has fundamental mechanical weaknesses, the driver usually understands and finds those weak points and repairs them. This, however, is not the case with computer networks. As we have seen, the network infrastructure has weaknesses and this situation is complicated when both system administrators and users have limited knowledge of how the system works, its weaknesses and when such weaknesses are in the network. This lack of knowledge leads to other problems including [2]:

- Network administrators not using effective encryption schemes, and not using or enforcing sound security policies.
- Existence of a persistent class of less knowledgeable administrators and users who quite often misuse passwords and they rarely care to change passwords.
- Network administrators' lack of understanding of social engineering practices and techniques. Social engineering is a process in which users carelessly give away information to criminals without being aware of the security implications.
- Network administrators failing to use system security filters. According to security experts, network servers without filters "are the rule rather than the exception."

- Network administrators often forgetting to use recommended patches.

The Growing Dependence on Computers

Our dependence on computers and computer technology has grown in tandem with the growth of computer networks. Computer are increasingly becoming part of our everyday life. From Wall Street to private homes, dependency on computers and computer technology shows no signs of abating. As this dependence grows, the family cabinet that used to store the family's treasures has given way to the family computer. So has the family doctor's files, our children's school file cabinet, and our business and workplace employee cabinets. Yet as we get more and more entangled in a computer driven society, very few have a sound working knowledge and an understanding of the basics of how computers work. Indeed, few show any interest in learning. In general, technology works better and is embraced faster if all its complexities are transparent to the user, and therefore, user-friendly.

Lack of Planning

Despite the potential for computer and computer network abuses to wreak havoc on our computer dependent society, as demonstrated by the "Love Bug" and the "Killer Resume" bug, there are few signs that we are getting the message and making plans to educate the populace on computer use and security [5]. Beside calling on the law enforcement agencies to hunt abusers down, apprehend them, bring them to book with the stiffest jail sentences to send a signal to other would-bes, and demanding for tougher laws, there is nothing on the horizon. There is no clear plan or direction, no blueprint to guide global even national efforts in finding a solution; very little has been done on the education front.

Complacency

People tend to develop a certain degree of security with a growing domain of experience usually acquired from long hours of working in the same trade. Network administrators also do the same. After several years of experience working with a computer network, they acquire that false sense of knowing and being able to handle every possible eventuality that may happen to the network. But as is always the case with technology, this is dangerous and as it is a false sense of security.

Inadequate Security Mechanism, Solutions and Best Practices

Although computer network software developers and hardware manufacturers have tried to find solutions to the network infrastructure and related problems, sound and effective solutions are yet to be found. In fact all solutions and best practices that have been provided so far by both hardware and software manufacturers have not been really solutions but patches. These best known security mechanisms and solutions, actually half solutions to the network infrastructure problems, are inadequate at best. More effective solutions to the network protocol weaknesses are not in sight. This, together with the lack of apprehending the perpetrators highlight

an urgent need for effective solutions and best practices that are still elusive. Yet the rate of such crimes is on the rise. With such rise, the law enforcement agencies are not able to cope with the rising epidemic as they lack both technical know-how and capacity.

Poor Reporting of Computer Crimes and a Slow Pace of the Law

Meanwhile headline-making vandals keep on striking, making more and more daring acts with impunity. Along with those headline makers, there are thousands of others not reported. The number of reported cyber crimes tracked by the Computer Emergency Response Team (CERT), the FBI, and local enforcement authorities is low. In fact according to reports, two-thirds of computer firms do not report hacker attacks [3]. Similar numbers are probably found in the private sector. In a study by the Computer Security Institute (CSI), of the 4,971 questionnaires sent to Information Security practitioners, seeking information on system intrusions, only 8.6 percent responded. Even those few responding, only 42 percent admitted that intrusions ever occurred in their systems [4]. This low reporting rate can be attributed to a number of reasons including:

- Many of those who would have liked to report such crimes do not do so because of both economic and a psychological impact such news would have on both the shareholders' confidence and the overall name of the company. Lack of customer confidence is a competitor's advantage and it may spell financial ruin to the company. Some companies are reluctant to report any form of computer attacks on their systems in fear that others, including shareholders, will perceive company management as weak with poor security policies.
- There is little to no interest in reporting.
- The law enforcement agencies, do not have highly specialized personnel to effectively track down the intruders. Even those few are overworked and underpaid according to the ABCNews report [3].
- Companies and businesses hit by cyber vandalism have little faith in the law enforcement agencies.

The law enforcement situation becomes even more murky when it is on a global scale. The global mosaic of laws, political systems, and law enforcement capacity make badly needed global efforts even more unattainable. Yet, as the "Love Bug" e-mail attack demonstrated, computer attacks have become global. This is making the search for perpetrators even more difficult.

Also current wiretap laws were designed for lengthy surveillance in one place in order to build a case. And if there is a cause to track down a perpetrator, court orders must be sought in every judicial district, which takes time and may lead to evidence getting altered or destroyed altogether. However, cyber attacks that are quick and can instantaneously have a global reach cannot be monitored from one place, and evidence cannot wait for court orders. To be effective laws must allow investigators, among other things, to completely trace an online communication to its source without seeking permission from each jurisdiction.

Security Loopholes in New Technologies

New technologies are creating gateways to restricted areas of any network in a number of ways including:

- Utilizing unknown weaknesses in the new technology
- creation of new opportunities in the existing technologies to access the network's restricted areas.

Bluetooth technology is a good example of these new technologies.

Bluetooth Technology

Bluetooth is a pervasive technology that supports wireless technologies in limited environments such as buildings and rooms within buildings. This technology can support and unifies a variety of everyday devices such as PCs, cellular phones, TVs, Wi-Fis, point-of-sell terminals, and many household devices including refrigerators and washing machines. The devices with the greatest growth and of interest to us are the smart wireless phones such as the blackberries. These devices offer all the functionalities of a wireless phone but in addition, they integrate more advanced functionalities of handheld devices such as operating systems services. Smart wireless phones are, therefore, able to perform most of the services of a PC and at the same time retain the phone services. They can send SMS, MMS, email, play videos and MP3 and are used to surf the internet and can do file transfers.

Even though Bluetooth technology has well known security protocols, it potentially has security loopholes in its core specifications which can lead to compromise in cryptographic algorithms used in Bluetooth communication such as sniffing. Other Bluetooth security concerns include [8].

- **BlueSnarf** – which uses object exchange (OBEX) to push services commonly used to exchange files such as business card files that can give an attacker access to valuable information found in a user card file.
- **Bluejacking** – a process to hijack exchanged device identifications during associations which may offer a way for an attacker to trick a user into allowing an attacker to access codes thus allowing a hacker into the system.
- **BlueBug** – which allows unauthorized access to the phone's set of "AT commands" thus allowing access to vital phones services including voice and text messaging.
- **BlueBump** – which utilizes Bluetooth link keys to allow unauthorized access to most device services leading to information theft.
- **BlueSmack** – a denial of service (DoS) attack which has the potential to knock out device services.
- **HeloMoto** – is a combination of BlueSnarf and BlueBag, first discovered on Motorola phones; hence the name.
- **BlueDump** – which causes device to dump the stored link keys thus creating a weakness for a key-exchange sniffing.

- **CarWhippers** – which allows the default configuration of hands-free headset devices to be hijacked resulting in a potential for PINs being accessed illegally.
- **BlueChop** – is also a DoS attack which can disrupt any established Bluetooth piconet.

Many of these attacks can get an intruder into systems using mobile technology. These attacks can also be performed at a distance using languages antennas and modified Bluetooth dongles [8].

Efforts, Solutions and Best Practices

Preventing the erosion and thus finding an effective secure perimeter defense for a private network is the Holy Grail of the solution to the network security problem. Can this be done? Probably not but efforts must not cease because as we know, we are chasing a solution that is a moving target sometimes moving faster than our efforts.

It is not our intention to leave you with the impression that there are no solutions yet and that we are fighting a losing battle. Solutions and best practices that can create a secure private network that has a high level of privacy, security, reliability, and integrity of information can be found. For such solutions and best practices to be effective, they must involve three strategies: user moral education concerning the use of the infrastructure and computer technology in general, nut-and-bolt security solutions and best practices to protect the networks, and a strong updated legal framework and an effective law enforcement regime [6].

- Security Through Moral and Ethical Education of the User in which we explore:
  - Morality and Ethics
  - An Ethical Education for the Worker: Codes of Conduct
- Building Secure Hardware and Software Systems in which we explore:
  - Network Security Basics
  - Security Threats and Vulnerabilities
  - Security Assessment, Analysis, and Assurance
  - Access Control, Authorization and Authentication
  - Perimeter Defense: The Firewall
  - System Intrusion Detection and Prevention
  - Security in Wireless Networks and Devices
  - Best Practices to Deal with Emanation and Emissions
- Security Through Deterrence: Computer Crime Investigations in which we explore:
  - Digital Evidence and Computer Crime
  - Computer Forensics Investigation Process

- Digital Evidence Collection and Controls
- Evidence Acquisition
- Computer Forensics Analysis
- Writing Investigative Reports

Conclusion

We have shown how technological developments, especially in information technology, are slowly but surely degrading the security of the private network security perimeter defenses. While new technologies create new applications and systems, they also create new but unwanted opportunities for hackers and unauthorized system access. We have given examples of many new information technology-related applications that have been welcomed because they are value adding technologies and yet they are degrading the effectiveness of the security cornerstone of the private network, its security perimeter defenses. Although this is happening, it is not all in despair, we have proposed ways through which these degrading forces may be contained.

References

Wikipedia – "Information superhighway". http://en.wikipedia.org/wiki/Information_highway.

Kizza, Joseph. M. Computer Network Security and Cyber Ethics. Jefferson, NC, McFarland, 2002.

ABC News. "Online and Out of Line: Why Is Cybercrime on the Rise, and Who Is Responsible?" http://www.ABCNews.com/sections/us/DailyNews/cybercrime_000117.html.

"Security in Cyberspace." U.S. Senate Permanent Subcommittee on Investigations, June 5, 1996. http://www.fas.org/irp/congress/1996_hr/s9606053.html.

Netscape. "'Love Bug' Computer Virus Wreaks Fresh Havoc." http://www.mynetscape.com/news/.

Kizza, Joseph. M. Securing the Information Infrastructure. IDEA Group publishers. (forthcoming 2007).

Arce, Ivan. "A Surprise Party (on Your Computer)?" IEEE Security and Privacy, vol. 5, no. 2, March/April, 2007, pp 15-16.

Merloni, Claudio, Luca Corettoni and Stephano zanero. "Studying Bluetooth Malware Propagation?" IEEE Security and Privacy, vol. 5, no. 2, March/April, 2007, pp 17-25.

Arce, Ivan. "Bad Peripherals?" IEEE Security and Privacy, vol. 3, no. 1, January/February, 2005, pp 70-73.

Loughry, Joe and david A. Umphress. "Information Leakage from Optical Emanations". ACM Transactions on Information and System Security, Vol. 5, No. 3, August 2002.

Gaudin, Sharon. " Malware Spikes in 1Q As Hackers Increasingly Infect Websites". Information Week, April 24, 2007.

28

Improving QoS with MIMO-OFDM in Future Broadband Wireless Networks

Bulega Tonny Eddie, Gang Wei, Fang-Jiong Chen

Designing very high speed wireless links that offer good quality-of-service (QoS) and range- capability in non-line-of-sight (NLOS) environments constitutes a significant research and engineering challenge. In this article, we provide an analytical overview and performance analysis, on the key issues of an emerging technology known as Multiple-Input Multiple-Output (MIMO) - Orthogonal Frequency Division Multiplexing (OFDM) wireless that offer significant promise in achieving high data rates over wireless links. MIMO technology holds the potential to drastically improve the spectral efficiency and link reliability in future wireless networks while the OFDM transmission scheme turns the frequency-selective channel into a set of parallel flat fading channels, which is an attractive way of coping with Inter-Symbol-Interference (ISI). The combination of MIMO and OFDM has been designed to improve the data rate and the QoS of the wireless system by exploiting the multiplexing gain and/ or the diversity gain which is a major problem in communication. In this article, key areas in OFDM-MIMO like Space-time block-coding (STBC), channel modeling and channel estimation are presented. Inter-channel-interference (ICI) which causes channel degradation is analyzed and we conclude by highlighting areas of further research.

Introduction

Demands for future wireless communication networks are to provide high data rates, better QoS while supporting an open architecture for a multiplicity of international standard wireless network technologies ranging from second-/ third-/fourth-generation (2G/3G/4G) cellular radio systems such as: Global System for Mobile (GSM), General Packet Radio Service (GPRS), Universal Mobile Telecommunications System (UMTS) to Wireless Local Area Networks (WLANs), Broadband Radio Access Networks (BRANs), Digital-Video Broadcast (DVB) and Digital-Audio Broadcast (DAB) networks [1, 2, 10]. While the available radio spectrum is a scarce resource, rising consumer demand for high bandwidth applications with diverse QoS guarantees over wireless networks has created an unprecedented technological-challenge to develop efficient coding and modulation schemes along with sophisticated signal and information processing algorithms to improve the quality and spectral efficiency of wireless communication links [3]. At the physical layer, QoS is synonymous with an acceptable signal-to-noise ratio (SNR) level or bit error rate (BER) at the receiver, while at the MAC or higher layers, QoS is usually expressed in terms of minimum rate or maximum delay guarantees. The fulfillment of QoS requirements depends

on procedures that span several layers. At the MAC layer, QoS guarantees can be provided by appropriate scheduling [4] and channel allocation methods [5]. At the physical layer, adaptation of transmission power, modulation level or symbol rate helps in maintaining acceptable link quality [6], [7]. Moreover, smart antennas constitute perhaps the most promising means of increasing system capacity through OFDM [8].

In recent years, the use of multiple-input multiple-output (MIMO) wireless technologies that offer significant promise in achieving high data rates over wireless links have captured a lot of interest. With a MIMO system, the capacity can be improved by a factor equal to the minimum number of transmit and receive antennas if perfect channel state information (CSI) is available at the receiver, compared with a single-input single-output (SISO) system with flat Rayleigh fading channels [9,10]. Another promising candidate for next-generation fixed and mobile wireless systems is the combination of MIMO technology with OFDM. OFDM transmission scheme turns the frequency-selective channel into a set of parallel flat fading channels and is, hence, an attractive way of coping with inter symbol interference (ISI) while providing increased spectral efficiency and improved performance. Since data is multiplexed on many narrow band subcarriers, OFDM is very robust with typical multi-path fading (i.e., frequency-selective) channels. Furthermore, the sub-carriers can easily be generated at the transmitter and recorded at the receiver using highly efficient digital signal processing schemes based on Fast Fourier Transform (FFT). The combination of MIMO and OFDM has been designed to improve the data rate and hence QoS by exploiting the multiplexing gain and the diversity gain.

Given these interesting benefits, several wireless networking (e.g., IEEE 802.11 and 802.16) and wireless broadcasting systems (e.g., digital video transmission-terrestrial (DVT-T), DAB) have already been developed using OFDM technology and are now available in mature commercial products.

The goal of this article is to provide an analytical review of the basics of MIMO-OFDM wireless systems with a focus on signal processing techniques for MIMO-OFDM. Channel modeling, channel estimation and the channel interference is analyzed. A system model is presented followed by a summary of performance results for a Rayleigh fading channel. Finally, we provide a list of relevant open areas for further research.

MIMO with OFDM Modulation

In wireless communications, multi-path propagation induces fading. This degrades the performance and capacity of the wireless link leading to poor QoS. A system that has demonstrated the ability to reliably provide high throughput in rich multipath environments is MIMO. By sending signals that carry the same information through these different paths, multiple independently faded replicas of the data symbol can be obtained at the receive end thereby providing the receiver with diversity [11]. A MIMO arrangement is shown in figure 1.

There are three kinds of MIMO techniques. The first aims to improve power efficiency by maximizing spatial diversity using such techniques as delay diversity, space-time block codes (STBC), and space-time trellis codes (STTC). The second uses a layered approach to increase capacity (e.g., V-BLAST architecture), while the third exploits knowledge of the channel at the transmitter. MIMO systems have proven to be very effective at combating time varying multipath fading in broadband wireless channels. In these systems, replicas of the transmitted signal, with uncorrelated variations in time, frequency, or spatial domain, or

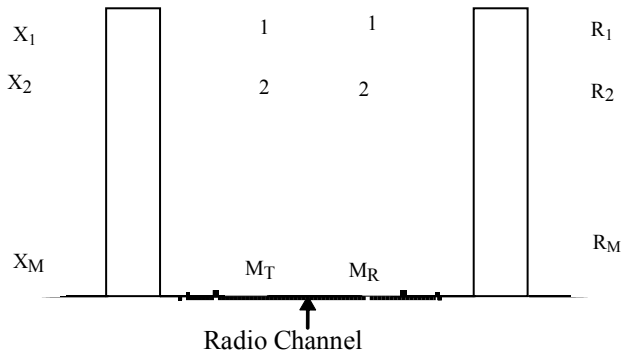
a combination of all three arrive at the receiver. The replicas are combined in such a way as to minimize the transmission degradation that could be caused by fading of each of the individual channels. Effectiveness or "diversity gain" of an implementation is dependent on the scenario for the propagation channels (e.g., rural or urban), transmission data rate, Doppler spread, and channel delay spread. When multiple antennas are used instead of single antenna systems, the following benefits are achieved:

Spatial multiplexing yields a linear (in the minimum of the number of transmit and receive antennas) capacity increase if perfect CSI is available at the receiver, compared to systems with a single antenna at one or both sides of the link, at no additional power or bandwidth expenditure [12,13]. The corresponding gain is available if the propagation channel exhibits rich scattering and can be realized by the simultaneous transmission of independent data streams in the same frequency band. The receiver exploits differences in the spatial signatures induced by the MIMO channel onto the multiplexed data streams to separate the different signals, thereby realizing a capacity gain.

Diversity leads to improved link reliability by rendering the channel "less fading" and by increasing the robustness against co-channel interference. Diversity gain is obtained by transmitting the data signal over multiple (ideally) independently fading dimensions in time, frequency, and space and by performing proper combination at the receiver. Spatial (i.e., antenna) diversity is particularly attractive when compared to time or frequency diversity, as it does not incur expenditure in transmission time or bandwidth respectively. Space-time coding realizes spatial diversity gain in systems with multiple transmit antennas without requiring channel knowledge at the transmitter.

Array gain can be realized both at the transmitter and the receiver. It requires channel knowledge for coherent combining and results in an increase in average receive signal-to-noise ratio (SNR) and hence improved coverage. Multiple antennas at one or both sides of the wireless link can be used to cancel or reduce ICI, and hence improve cellular system capacity.

Fig 1: A block diagram of MIMO transceiver



A MIMO system with a transmit array of M_T antennas and a receive array of M_R antennas is shown in the figure 1. The transmitted matrix is a M_T x 1 column matrix X where X_i is the ith component, transmitted from antenna i. The channel is considered to be a Gaussian

channel such that the elements of X are independent identically distributed (i.i.d.) Gaussian variables. If the radio channel is unknown at the transmitter, the signals transmitted from each antenna have equal powers of Es/M_T. The covariance matrix for this transmitted signal is given by

(1)

Where Es is the power across the transmitter irrespective of the number of antennas M_T and I_{MT} is an M_T X M_T identity matrix. The channel matrix **H** is an M_R X M_T complex matrix. Therefore, the received vector can be expressed as the input–output relation over a symbol period for a single-carrier modulation and is given by

(2)

Where **y** is the M_R x 1 received signal vector, X is an M_T X 1 transmitted signal vector, **H** is the M_R x M_T MIMO channel matrix and n is additive temporally white complex Gaussian noise.

Capacity of a MIMO channel

The Shannon capacity of a communication channel is the maximum asymptotically (in the block-length) error-free transmission rate supported by the channel. If the transmitted codewords span an infinite number of independently fading blocks, the Shannon capacity also known as ergodic capacity is achieved by choosing to be circularly symmetric complex Gaussian which results in Rss = $I_{MT.}$ It has been established in [12] that at high SNR,

$$C = \min(M_R, M_T)\log_2 \rho + O(1) \quad (3)$$

With ρ = Es/No, which clearly shows the linear increase in capacity with the minimum number of transmit and receive antennas.

OFDM Modulation

OFDM can be thought of as a hybrid of multicarrier modulation (MCM) and frequency shift keying (FSK) modulation. Orthogonality among carriers is achieved by separating them by an integer multiple of the inverse of the symbol duration of the parallel bit streams, thus minimizing ISI. Carriers are spread across the complete channel, fully occupying it and hence using the bandwidth efficiently.

OFDM is a block modulation scheme where a block of N information symbols is transmitted in parallel on N sub-carriers.

The time duration of an OFDM symbol is N times larger than that of a single-carrier system. An OFDM modulator can be implemented as an inverse discrete Fourier transform (IDFT) on a block of information symbols followed by an analog-to-digital converter (ADC). To mitigate the effects of ISI caused by channel time spread, each block of IDFT coefficients is typically preceded by a cyclic prefix (CP) or a guard interval consisting of samples, such that the length of the CP is at least equal to the channel length. Under this condition, a linear convolution of the transmitted sequence and the channel is converted to a circular convolution. As a result, the effects of the ISI are easily and completely eliminated.

MIMO OFDM System Model

A multi-carrier system can be efficiently implemented in discrete time using an IFFT to act as a modulator and an FFT to act as a demodulator. The transmitted data are the "frequency" domain coefficients and the samples at the output of the IFFT stage are "time" domain samples of the transmitted waveform.

Let $X = \{X_0, X_1, ...X_{N-1}\}$ denote the length-N data symbol block. The IDFT of the date block yields the time domain sequence $x = \{x_0, x_1,....x_{N-1}\}$, i.e.,

$$x_n = IFFT_N\{X_k\}(n). \quad (4)$$

To mitigate the effects of channel delay spread, a guard interval comprised of either a CP or suffix is appended to the sequence X. In case of a CP, the transmitted sequence with a guard interval is

$$n = -G,.....,-1, 0, 1,.....N-1 \quad (5)$$

where G is the guard interval in samples and $(n)_N$ is the residue of n modulo N.

To avoid ISI, the CP length G must equal or exceed the length of the discrete time channel impulse response.

Fig 2: Block diagram of OFDM modulation with MIMO

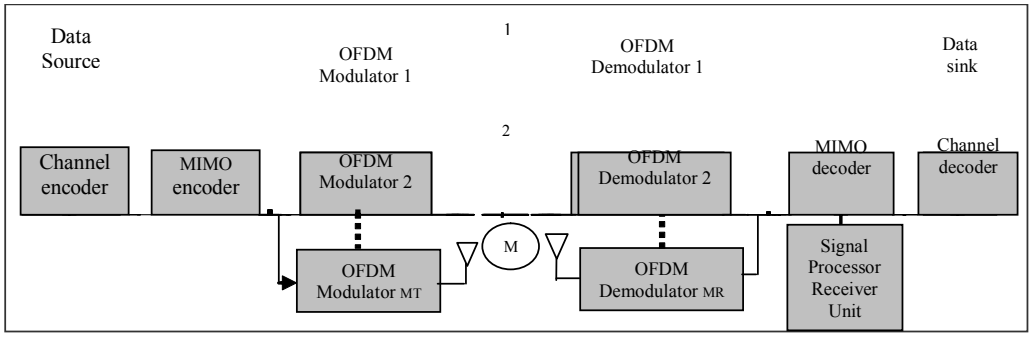


Fig 3: Frame structure of the OFDM symbol block

Data	Pilot	Data	Data	CP

The OFDM preamble consists of Q training symbols (where Q is the number of antennas) of length, N_I+G where $G \leq N_I \leq N$, N_I = N/I and I an integer that divides N. Often the length of the guard interval in the training period is doubled; for example, in IEEE802.16a [14], to aid in synchronization, frequency offset estimation and equalization for channel shortens in cases where the length of the channel exceeds the length of the guard interval.

$$y = \sqrt{\frac{E_s}{M_T}}\, Hx + n$$

$$\{R_k^{(\ell)}\}_{k=0}^{N_I-1}$$

OFDM Channel Analysis

Let $H_{i,j}$ be the vector of sub-channel coefficients between the *i*th transmit and the *j*th receive antenna and let be the received sample sequence at the *l*th receiver antenna.

After removing the guard interval, the received samples are repeated *I* times and demodulated using the *N*-point FFT as

$$R_k^{(l)} = FFT_N\{r^{(\ell)}\}(k) \tag{6}$$

$$= \sum_{q=1}^{Q} H_k^{(q,\ell)} S_k^{(q)} + W_k^{\ell} \tag{7}$$

Where $k = 0,....N\text{-}1$. The demodulated OFDM sample matrix R_k of dimension (QxL) for the *k*th sub-carrier can be expressed in terms of the transmitted sample matrix S_k of dimension (QxQ), the channel coefficient matrix H_k of dimension (QxL) and the additive white Gaussian noise W_k matrix of dimension (QxL) [16] as

$$R_{k,QxL} = S_{k,QxQ} \cdot H_{k,QxL} + W_{k,QxL} \tag{8}$$

Where R, H and W can be viewed as either a collection of *N* matrices of dimension $Q \times L$ or as a collection of $Q \times L$ vectors of length *N*.

ICI caused by Time varying channels

Assuming that the multi-path fading channel under consideration consists of Q discrete paths, the received baseband OFDM signal after FFT can be written as

$$R(k) = \left(\sum_{q=0}^{Q} H_q(0) e^{-j2\pi n_q k/N} \right) X(k) + I(k) + W(k) \tag{9}$$

Where *N* and n_q denote the number of sub-carriers and the time delay of the *q*th path respectively. Also, $H_q(k)$ and $I(k)$ represent the frequency response of a time-variant channel, $h_p(n)$, and the ICI caused by the time-variant channel respectively. Here $H_p(k)$ and $I(k)$ are given by

$$H_p(k) = \frac{1}{N} \sum_{n=0}^{N-1} h_q(n) e^{-j2\pi nk/N} \tag{10}$$

$$I(k) = \sum_{\substack{m=0 \\ m \neq k}}^{N-1} \left(\sum_{q=0}^{Q-1} H_q(k-m) e^{-j2\pi n_q m/N} \right) X(m) \tag{11}$$

If the channel is time-invariant during a symbol period, the term in the right hand side of (11) contains only the multiplicative distortion, which can easily be compensated by a one tap frequency-domain equalizer. However, in fast fading channels, the time variation of a fading channel over an OFDM symbol period

causes a loss of sub-channel orthorgonality resulting in an error floor that increases with the mobile system's speed [18].

The time variant channel within an OFDM block can be approximated by a *D*-th order polynomial function as

$$\hat{h}_q(n) = \sum_{d=0}^{D} a_{q,d} n^d + b_q \qquad n = 0, 1, ..., N\text{-}1 \qquad (12)$$

Here the parameters $a_{q,d}$ and b_q can be found by solving the least square equation. Then, the frequency response of the approximated time-variant channel up to 2nd order $\hat{h}_q(n)$ can be written as

$$\hat{H}_q^2(k) = \begin{cases} \hat{H}_q^1(o) + a_{q,2}(N-1)(2N-1)/6 & for\ k=0 \\ \hat{H}_q^1(k) + a_{q,2}\dfrac{(N-2)e^{-j2k/N} - N}{(1-e^{-j2k/n})^2} & for\ 1 \le k \le N-1 \end{cases} \qquad (13)$$

Here $\hat{H}_q^2(k)$ represents the frequency response of the approximated time-variant channel of the 1-*st* order and is given by

$$\hat{H}_q^1(k) = \begin{cases} b_{q,1}(N-1)/6 & for\ k=0 \\ \dfrac{a_p 1(-1+j\cot(k/N)}{2} & for\ 1 \le k \le N-1 \end{cases} \qquad (14)$$

The above equation shows the approximate effect of ICI caused by a fast fading channel. If the channel is time-invariant during an OFDM symbol period, i.e. $a_{q,2} = a_{q,1} = 0$, there exists no ICI since for $k \neq 0$. However, in the case where the channel is time

invariant within a symbol period, i.e. $a_{q,1} \neq 0$ or $a_{q,2} \neq 0$, the leakage signals are distorted over all other sub-carriers resulting in ICI.

Interference Reduction

Co-channel interference arises due to frequency reuse in wireless channels. When multiple antennas are used, the differentiation between the spatial signatures of the desired signal and co-channel signals can be exploited to reduce interference. Interference reduction requires knowledge of the desired signal's channel. Exact knowledge of the interferer's channel may not be necessary. Interference reduction (or avoidance) can also be implemented at the transmitter, where the goal is to minimize the interference energy sent toward the co-channel users while delivering the signal to the desired user. Interference reduction allows aggressive frequency reuse and thereby increases multi-cell capacity.

We note that in general it is not possible to leverage all the advantages of MIMO technology simultaneously due to conflicting demands on the spatial degrees of freedom (or number of antennas). The degree to which these conflicts are resolved depends upon the signaling scheme and transceiver design.

MIMO-OFDM Channel Estimation

Channel state information is required in MIMO-OFDM for space–time coding at the transmitter and signal detection at receiver. Its accuracy directly affects the overall performance of MIMO-OFDM systems. Channel estimation for OFDM can exploit time and frequency correlation of the channel parameters. A basic channel estimator has been introduced in [19]. As discussed before, for a MIMO system with Q transmit antennas, the signal from each receive antenna at the kth sub channel of the nth OFDM block can be expressed as;

$$R_{n,k} = \sum_{q=1}^{Q} H_{n,k}^{(q)} X_{n,k}^{(q)} + W_{n,k} \tag{15}$$

where $H_{n,k}^{(q)}$ is the channel frequency response at the kth sub channel of the OFDM block corresponding to the qth transmit antenna, and $W_{n,k}$ is the additive white Gaussian noise.

The challenge with MIMO channel estimation is that each received signal corresponds to several channel parameters. Since the channel response at different frequencies is correlated, channel parameters at different sub-carriers can be expressed as

$$H_{n,k}^{q} = \sum_{q=1}^{N_o-1} h_{n,k}^{(q)} + W_N^{kq} \tag{16}$$

for k = 0, ... N-1, q = 1,....Q and . The parameter N_o depends on the ratio of the delay span of wireless channels to the OFDM symbol duration, and

$W_N = \exp(-j(2\pi/N))$. Hence, to obtain $H_{n,k}^{q}$ we only need to estimate $h_{n,m}^{q}$. If the transmitted signals $X_{n,k}^{q}$ from the qth transmit antenna are known for $q = 1,...Q$ then $\hat{h}_{n,k}^{q}$ a temporal estimation of $h_{n,k}^{q}$ can be estimated by minimizing the cost function

$$\sum_{k=1}^{N-1} \left| R_{n,k} - \sum_{q=0}^{Q} \sum_{m=0}^{No-1} \hat{h}_{n,m}^{(q)} W_K^{kq} x_{n,k}^{q} \right|^2 \tag{17}$$

The estimated channel parameter $\hat{h}_{n,m}^{q}$ can be decomposed into the true channel parameter $h_{n,m}^{q}$ and the estimation error $e_{n,m}^{q}$ that is

$$\hat{h}_{n,m}^{q} = h_{n,m}^{(q)} + e_{n,m}^{(q)} \tag{18}$$

$e_{n,m}^{(q)}$ can be assumed to be Gaussian with zero-mean and variance σ^2, and independent for different qs, ns or ms. If the parameter estimation quality is

measured by means of the normalized mean square error (NMSE), which is defined as

$$NMSE = \frac{E\left|\hat{H}_{n,k}^{q} - H_{n,k}^{q}\right|^2}{\left|H_{n,k}^{q}\right|^2} \tag{19}$$

Then it can be calculated directly that the NMSE for the estimation in (19) $\mathrm{NMSE}_r = N_o\sigma^2$

Space-time Coding for MIMO-OFDM

Quantitatively, the maximum achievable diversity order is a product of the number of transmit antennas, the number of receiver antennas, and the number of resolvable propagation paths (i.e., the channel impulse response length) [16], [17]. To achieve this full diversity requires that the information symbols be carefully spread over the tones as well as over the transmitting antennas. A space-time code- or more generally, a space–time-frequency code-is a strategy for mapping information symbols to antennas and tones as a means for extracting both spatial and frequency diversity.

Space-time coding jointly encodes the data streams over different antennas, and therefore aims to maximize diversity gain. Two main space-time coding schemes, STBC and STTC, are mentioned here. STTC obtains coding and diversity gains due to coding in space-time dimensions. However its' decoding complexity increases greatly as the size of modulation constellations, state number, and code length increase. STBC based on orthogonal design obtains full diversity gain with low decoding complexity, and therefore has been widely used. The well-known Alamouti code is just a special case of STBC with double transmit antennas. Space-frequency block code (SFBC) is based directly on space-time codes (with time reinterpreted as frequency).

STBC is optimally designed under the assumption that the fading channel is quasi-static.

Therefore, the time or frequency selectivity degrades the performance of STBC and SFBC. Between SFBC and STBC, one is selected based on the selectivity of the channel in the time or frequency domain. Whatever the delay spread of the channel, STBC is chosen only if the channel is slowly varying in the time domain when the terminal moves slowly. Similarly, at whatever speed the terminal moves, SFBC is chosen only if the channel is slowly varying in the frequency domain when the delay spread of the channel is small.

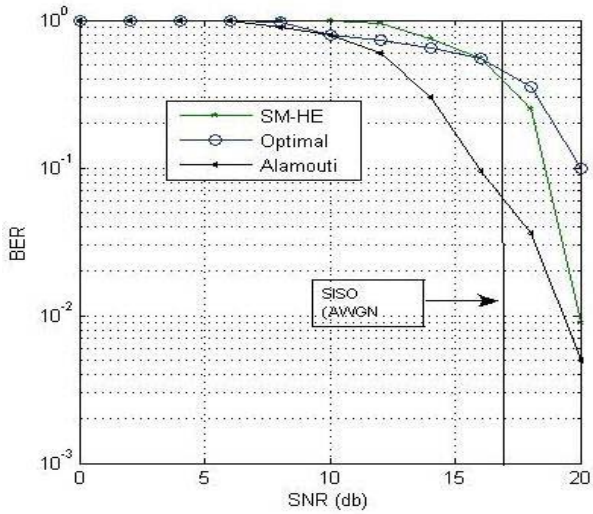
MIMO-OFDM Performance Analysis

For the sake of clarity of exposition we consider a simple MIMO system with $M_T = M_R = 2$. We assume that the transmitter has knowledge of the channel and the SNR p. Fig. 4 plots the BER as a function of SNR for a fixed transmission rate of 6 b/s/Hz. The magnitude of the slope of the BER curve has been shown

to be $M_T M_R$ [14] for a fixed rate and at high enough SNR. This indicates that for fixed rate transmission, optimal coding yields full $M_T M_R$.th-order spatial diversity inherent in the channel. In comparison, the BER curve for a single input single output (SISO) AWGN channel with a signaling rate of 6 b/s/Hz is a vertical line at $p =$ 18 dB, i.e., an error is always made if we attempt to transmit at 6 b/s/Hz over the SISO AWGN channel when $p < 18$ dB. The result confirms the notion that an AWGN channel has infinite diversity [15] and furthermore shows that for SNR below 18 dB, the MIMO fading channel has better performance in terms of BER than the SISO AWGN channel. The same slice for the optimal coding is depicted for fixed transmission rate. For optimal coding for a fixed transmission rate, we can trade an increase in SNR for a reduction in BER (diversity gain equal to $M_T M_R$), and conversely for a fixed BER, we can trade an increase in SNR for a linear increase in data rate.

We note that the two schemes (spatial multiplexing with horizontal encoding (SM-HE) and STBOC) used for MIMO OFDM transmission lie in the achievable region. Indeed STBOC has been shown to achieve full diversity of $M_T M_R$ [20]. Furthermore, at low SNR, SM-HE outperforms the Alamouti scheme. However, due to the higher diversity gain of the Alamouti scheme, at high SNR the situation reverses. We can see that the question of which scheme to use depends significantly on the target BER and the operational SNR. These encouraging results show that MIMO-OFDM creates a new way to achieve high bandwidth efficiency without sacrificing additional power or bandwidth.

Fig 4: BER performance of a MIMO OFDM with STBOC Alamouti spatial multiplexing



Areas for Future Research

We conclude the article with a brief discussion of open problems in the area of MIMO-OFDM that need to be addressed so that the gains promised by the technology can be fully leveraged in practical systems. Multi-user MIMO systems

are largely unexplored. Making progress in the area of multi-user MIMO systems is of key importance to the development of practical systems that exploit MIMO gains on the system level. The recently launched EU FP6 STREP project MASCOT (Multiple-access Space-Time Coding Test bed) is aimed at developing, analyzing, and implementing (in hardware) concepts and techniques for multi-user MIMO communications. Specific areas of relevance in the context of multi-user MIMO systems include multiple-access schemes, transceiver design (including precoding), and space-frequency code design. In particular, the variable amount of collision-based framework for multiple access, introduced, needs to be further developed to account for the presence of out-of-cell interference and to allow for variable amounts of collision in space, time, and frequency. Another critical area is MIMO OFDM channel estimation using blind methods. By using blind methods, a great deal of bandwidth which is used for pilots can be utilized to increase capacity.

Conclusion

We provided a brief overview of MIMO-OFDM wireless technology covering some key aspects of the system design such as; channel modeling, ICI analysis, channel estimation and space time block coding aimed at increasing the transmission rate and providing reliable QoS to users. The field is attracting considerable research attention in all of these areas. Performance results are presented showing the potential of MIMO combined with signal processing techniques in achieving high data rates through OFDM. Significant efforts are underway to develop and standardize channel models for different systems and applications. Understanding the information-theoretic performance limits of MIMO-OFDM systems, particularly in a channel design and estimation context, is an active area of research because they form the basis for reducing ICI and ISI hence improving the QoS. Finally, we feel that a better understanding of the system design implications of fundamental performance tradeoffs (such as rate versus BER versus SNR) is required to enhance performance. The high bandwidth efficiency obtained shows that MIMO-OFDM is a potential candidate for future broadband wireless access.

References

Alamouti, S. M. "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, October 1998

Jankiraman, M. *Space-time codes and MIMO systems*, Artech House London, 2004.

Bingham, J. A. C. Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Transaction on Communications*, vol. 28, pp. 5-14, May 1990

Andrews M., Kumaran K., Ramanan K., A. Stolyar and Whiting P., "Providing Quality of Service over a shared wireless link", *IEEE Commun. Mag.*, vol.39, no.2, pp.150-154, Feb. 2001.

Katzela I. and Naghshineh M., "Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey", *IEEE Pers. Commun.*, vol.3, no.3, pp.10-31, June 1996.

Grandhi S.A. , Vijayan R., Goodman D.J. and Zander J. , "Centralized power control in cellular radio systems", *IEEE Trans. Veh. Tech.*, vol.42, no.4, Nov.1993

Morinaga N. , Nakagawa M. and Kohno R., "New concepts and technologies for achieving highly reliable and high-capacity multimedia wireless communication systems", *IEEE Commun. Mag.*, vol.37, no.1, pp.34-40, Jan. 1997

Sheikh K., Gesbert D., Gore D. and Paulraj A., "Smart antennas for broadband wireless access networks", *IEEE Commun. Mag.*, vol.37, no.11, pp.100-105, Nov. 1999

Foschini, G. J. "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, pp. 41-59

Tarokh V., Jafarkhani H. , and Calderbank A. R., "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, July 199, autumn 1996.

Paulraj A. J. and Kailath T., "Increasing Capacity in Wireless Broadcast Systems Using Distributed Transmission Directional Reception," U.S. Patent no. 5, 345, 599, 1994

Telatar I. E. , "Capacity of Multi-Antenna Gaussian Channels," *Euro. Trans. Telecom.*, vol. 10, no. 6, Nov./Dec. 1999, pp. 585–95

Li Y. G., Winters J. H. and Sollenberger N. R., MIMO-OFDM for wireless communications: signal detection with enhanced channel estimation," *IEEE Transactions on Communications*, vol. 50, pp. 1471{1477, Sep. 2002

Blum R., Li Y., Winters J., and Yan Q., "Improved space–time coding for MIMO-OFDM wireless communications," *IEEE Trans. Commun.*, vol. 49, no. 11, pp. 1873–1878, Nov. 2001

ETSI, "Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television," July 1999, EN 300 744 V1.2.1

Paulraj A. J. and Papadias C. B., "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 46, no. 6, pp. 49–83, November 1997

Foschini G. J. and Gas M. J., "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal* Comm., vol: 6, pp. 311-335, 1998

Marzetta T. L. and Hochwald B. M., "Capacity of a mobile multiple antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999

Zheng L. and Tse D. N. C., "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003

Jakes W., *Microwave Mobile Communications*. New York: Wiley, 1974

29
Analysis of Free Haven anonymous storage and publication system

Drake Patrick Mirembe, Francis Otto

In this paper we evaluate the design of a distributed anonymous storage and publication system that is meant to resist even attacks of most powerful adversaries like the government. We present a discussion of whether the assumptions held in the design of Free Haven System (FHS) are achievable in the real world and the potential implementation handles. We point out the different possible attacks and voice our opinions on the suggested solutions. We end with recommendations on how to improve the FHS design and offer direction for future research efforts in distributed anonymous storage services.

Background

The need to share information anonymously has been in existence since time immemorial. According to Anderson [2], in medieval times, knowledge was guarded for the power it gave those who possess it. He noted that, for example, the bible was controlled by the church for the knowledge it gave those who possessed it. Besides being encoded in Latin, bibles were often locked up in what we would call highly restricted areas. However, the spread of technical know-how destroyed the guild that had accreted abuses over centuries of religious monopoly. Today, many oppressed individuals desire to publish their criticisms anonymously in order to avoid persecution from those in authority. It is situations like those mentioned above that have inspired the design and deployment of anonymous storage and publishing services or networks such as the Eternity service, Mojo Nation, Gnutella, Napster and among others. The events that befell Napster and Gnutella in 1999 and 2000 respectively [1] and the work of Anderson [2] provided extra motivation for the development of a new generation of anonymous publishing service such as FreeNet and Free Haven System (FHS). As one can imagine the threat model of such a system would include most powerful adversaries such as "the governments and terrorists".

A. FHS Overview

The FHS concept is based on the use of a network of servers called ServNet; each server in the community holds segments of some documents referred to as shares which are created from documents by use of Rabins Information dispersal algorithm [4] and are traded between servers based on a buddy system [1]. To introduce accountability in the system, FHS uses a reputation system, which is

based on the performance of a server in transactions within the ServNet instead of a more complex digital cash system [5] as proposed by Anderson. To achieve server anonymity, servers are referred to by their pseudonym in the network. FHS relies on the secure MixNet for communication between ServNet nodes and the buddy system to check corrupt nodes on the network. To retrieve a document, the reader generates a key pair (PKclient, SKClient) and broadcasts the hash of the document together with a one time remailer reply block. A node that has a share of the requested document will reply the request using the PKClient and the remailer reply block information, and after receiving sufficient shares, the client reconstructs the document. In all, the FHS design emphasizes distributed, reliable, anonymous storage and publication over efficient information retrieval [8].

While in systems like FreeNet a publisher refers to a server, in the context of FHS, a publisher is an entity that places document in the system, while an author is the entity that originally created a document.

B. Paper structure

The paper begins with a brief background on the subject of anonymous publication and FHS in Section One. We present an overview on the design of FHS and a discussion on the subsystems that make up FHS which include the publication system, the reputation system and the communication system in Section Two. We proceed with an enumeration of possible attacks on system in Section Three, which may be social, political, technical and legal attacks. We present a discussion about the successes and failures of FHS in Section Four and we end with recommendations and future research directions in Section Five

FHS Design and Operation

The design of free haven was partly inspired by the earlier work of Anderson [2] when he proposed the Eternity anonymous storage service in 1996. Other projects, like FreeNet [3], Mojo Nation, Gnutella, Publius and Napster have highlighted the need to design a more robust anonymous storage and publication system. The design of FHS consists of mainly two parts, the publication system, which is responsible for storing and serving documents and the communications systems which provides a channel for anonymous communication between servers [1]. Little is added on the communication system as FHS basically uses the remailer network concepts which are already deployed in Mixmaster remailer networks including onion routing, freedom network and cypherpunk [1]. To improve the performance of FHS, the design allows publishers to set share expiry date, which guarantee that even if the document is unpopular, it will stay in the system as long as desired by the publisher. This is one of the improvements FHS added to the concept of anonymous publishing adopted earlier by FreeNet, Mojo Nation among others. In the following sections we give a detailed description of goals, design, and operation of FHS.

A. Goals of Free haven Project

Most of the early anonymous systems aimed at providing storage anonymity but not publication anonymity which led to the events that befall Napster and Gnuttela. In order to come up with a true anonymous storage and publishing system, the FHS team set the following as their design goals.

Provide **author anonymity;** according to [1] the FHS is designed to offer author anonymity. A system is said to be author anonymous if an adversary cannot link an author of document to the document itself.

Document anonymity; this means the contents of the document can not even be read by the server that is storing the document. For the survival of FHS, this property is crucial, as possession of an illegal content in a given jurisdiction can be a cause of censor to the server operator.

Publisher anonymity; this property is important in order to prevent an adversary from linking a publisher of a given document to the document itself.

FHS seeks to provide **Server anonymity;** this property means, given document identifiers, an adversary is not closer enough to link a document to the server(s) where it is stored.

FHS is **reader anonymity**; in order to enhance the privacy of the readers, FHS is designed with a module that makes its hard for an adversary to link a reader of a document to the document itself.

Query anonymity; by using the concepts of private information retrieval [6], FHS aims at preventing a server from determining which document it is serving to answer a user request.

Longevity of unpopular documents; The FHS system is designed with a mechanism that allows documents to remain in the system as long as the publisher desires. The duration of a document depends on the publisher of the document but not on the popularity of the documents itself as the case with FreeNet and others.

B. The Publication System

The FHS publication system uses the concept of (n, k) secret sharing schemes. A good example of these schemes is a polynomial of degree n $P = f(x)$, $f(x) = a_n x^n + a_{n-1} + x^{n-1} + \ldots + a x + a$ for where by only k out of n points are sufficient to evaluate. Thus, for an author to publish a document in FHS, he first looks for a ServNet node willing to store his document D, then applies Rabins information dispersal algorithm (RID) [4] on the document to break it into n shares of which any k out of n shares are sufficient to reconstruct the document. The publishing server then generates a public-secret key pair (PK_{doc}, SK_{doc})(axaxaxaxfnnnn++++==−−)(xfdoc, SKdoc), constructs and signs a data piece for each share (for accountability purposes during trading sessions) as well as insert other control information into the data piece like time stamp, expiry date for the share, share number and Pkdoc for verification purposes. For robustness, optimal k must be used, since a large k comparative to n makes the document D unrecoverable if a

few numbers of shares are corrupted while a small k means more storage space for a single share [3].

To provide cover for the publisher of a document and to allow servers to freely enter and leave the network without suffering damage to their reputation, servers periodically trade their shares within the ServNet. This provides a moving target to an attacker and therefore makes it difficult for any adversary to discover who the publisher of a given document is.

For secure and efficient retrieval, document shares in FHS are indexed by the hash of their public keys. And therefore, to retrieve a document, a searcher who is also a member of the ServNet generates the key pair (PKclient, SKclient) and one time remailer reply block for this transaction. Then, the searcher server broadcasts a document request H (PK_{doc}) along with its client public and reply block to its buddies. Any server that has a share of a document with H (PK_{doc}) as its index replies the request by encrypting the index of the share using the client's public key and forwarding it to the address provided in the reply block. After receiving at least **k** shares out n of **D**, the client reconstruct the document [1, 3 and 12].

C. The Communication System

For anonymous communication, FHS relies on the existing anonymous communications schemes such as Mixmaster remailers, Onion routing, and freedom network among others. However, the current design of FHS uses Mixmaster remailers and cypherpunk as anonymous communication channels. These schemes use the chaining technique to send encrypted messages to other nodes via a chain of nodes which can not read the contents of the packets. Since little is added to these schemes in the design of FHS, the reader is therefore referred to [1, 2] for a detailed description.

D. The Reputation System

To prevent attacks engineered by some malicious servers within the ServNet and to improve the trust between nodes, FHS relies on the reputation system. Many opportunities exists for servers to be naughty, servers can delete shares before their expiry date, wrongfully accuse other nodes for deleting shares, refuse to send acknowledgements after transactions among others [1, 5 and 7]. The design of FHS incorporates a "buddy subsystem" that creates dynamic association of servers holding shares of a given document. In this scheme every server maintains the reputation and credibility values of other servers in the buddy list basing on the number of successful transactions a given server say (A) has made, complaints received from other server about A, reputation of new servers A has introduced, validity of complaints A broadcasts among others. This reputation gives a degree of confidence servers put in a given server in regard to its adherence to the free haven protocol and the value of control information that server generates. This careful monitoring of buddies allows servers to keep track of buddies to trust in the ServNet [1, 7].

Also the system provides an easy way to add new servers and removing inactive ones. As servers with good reputation scores are trusted to act as introducers of newcomers in the ServNet via anonymous communication channel. Hence, allowing the dynamic growth and shrinking of the ServNet [7].

Attacks on Free Haven

The threat model of Free Haven system comprises some of the most advanced adversaries; Governments, corporations, and individuals all have reasons to oppose the deployment of FHS. The reasons may include copy right protection, fight against child pornography, fight against terrorism among others. These adversaries may employ technical, legal, political and social schemes as means to undermine a successful deployment of the system. In the proceeding sections we elaborate on these attacks.

A. Technical attacks

Technical attacks come from individuals, corporations and national intelligence agencies, targeting either the system as a whole or particular documents or servers in a given location so as to reduce the quality of service or gain control of part of the network. Given the post 9/11 world we live in with diminishing privacy and anonymity, national security agencies will deploy all the resources at their disposal to again access to documents in the system since the deign of FHS may provide a secure communication channel for terrorists and criminals. By use of viruses, worms and spy ware, security agencies and individuals can join the network and collect vital user identifiable information hence; violating the privacy of the users. Beside, the worms and viruses may be used to shutdown sections of ServNet.

Adversaries may also flood the system with queries so as to use up available resources hence, the denial of service attack [1, 7]. Other attacks like buddy co-opting in which the adversaries may join the ServNet and try to gain control of a good number of nodes, simple betrayal, server refusal to issue receipts at the end of a trading session, share hoarding in which a server trades away enough garbage so as to gain control of a significant amount of shares of a targeted content, false referrals in which a server broadcasts false reputation scores of other servers or simply forwards scores to only selected collaborating servers, log publishing and traffic analysis are all possible [1, 9].

B. Political and Legal attacks

The most difficult attacks to defend against are those that are politically engineered; individuals in the positions of authority in a given jurisdiction can use their influence to discourage use of the system. The authorities can attempt to find a physical server containing controversial documents and order its operators to shut it down or even prosecute the owners.

In some cases ordinary citizens may also employ the power of the government through lawsuits, multinational corporations who feel threatened by the deployment of FHS as it can encourage corporate espionage and infringement of

copy rights could persuade countries in which they operate to pass laws that bar the deployment of anonymous storage and publishing networks such as FHS. The Motion Picture Association of American (MPAA) and its world wide counterpart Motion Picture Association (MPA) and the Recording Industry Association of America (RIAA) have demonstrated resilience in fighting P2P networks such as Napster and Gnuttela for their facilitation in the distribution of copy righted content. We strongly believe it is organization like MPA and RIAA that pose a serious threat to a successful deployment of FHS.

C. Social attacks

The degree of social attacks on anonymous systems depends on the culture of the society in which the system is deployed. Some cultures like African cultures associate privacy and anonymity with evil, since in their culture; evil acts are committed in secret places and thus, whoever demands privacy and anonymity is a suspect of evil doing. Therefore, operating FHS in such societies will always be difficult; citizens will try to influence their governments so as to undermine the trust in the security of the system, as well as question the moral justification of ServNet node operators. These attacks can take the form of demonstrations and campaigns against the deployment of FHS, black mailing the operator of ServNet nodes of inappropriate conduct in society like facilitating child pornography among others. The current social pressure exerted on other P2P networks like KaZaA, Limewire, Napster among others strengthen our claim that FHS will also be subjected to the same.

Evaluation

Given the fact that at the time of writing this paper, FHS was and (is) still largely a conceptual design though with a proof of concept implementation by Dingledine [1] and his team, to our knowledge little in literature is discussed about the success and failures of FHS as a system. Although, FHS shares most of its properties with other distributed anonymousing systems like FreeNet, its good to mention that it's design introduces a number of concepts in the study of anonymous publication and storage systems such as query anonymity, document anonymity, publisher defined life span among others and our evaluation in this section is entirely based on the design of the FHS, the validity of assumptions made and the literature about similar systems particularly FreeNet.

A. Weakness

One of the weaknesses of FHS is its communication infrastructure. FHS relies on the existing anonymous communication channels for linking nodes within the ServNet, but these communication schemes such as Mixmaster and Onion routing (e.g. TOR) are unreliable and inefficient and therefore suffer from number of attacks, such as;

Traffic analysis attacks [1, 7, 8, and 9]; Mixmasters can be subjected to traffic analysis attacks by either a local or a global observer who can monitor traffic in

the ServNet and basing on the statistics (such as bandwidth) he can deduce, the identity of the communicating parties (source or destination) hence the loss of publisher and reader anonymity which is one of the goals of FHS. FHS team has developed a new generation of Onion routers called TOR [7] which is hoped to be more resistant to traffic analysis, however TOR has a weak threat model and a low user base. Since its inception, TOR has only remained in the hands of enthusiastic privacy and anonymity researchers due to its failure to win general user confidence and thus it seems less promising than first thought. Technically given its weak threat model TOR, does not promise much in the prevention of traffic analysis.

Usability weakness; the current design of FHS assumes that users have global knowledge of the network (keep a database of all public keys of nodes in the ServNet and reputation of buddies), which is absolutely impractical in an ideal peer-to-peer network made up of hundreds of thousands of peers. Users often have heterogeneous computing resources which may not be adequate to allow such extra processing and the size and dynamism of the P2P system makes this concept even more remote to achieve.

The design specifies the use of a buddy system which creates an association of nodes based on shares the server is holding so as to achieve accountability [1, 7]. Since servers query one another so as to determine the buddies of shares, it mean servers in FHS can read share identifiers and can determine which server hold what document as long as they hold buddy shares of a given document. Once more, receipts exchanged during the trading sessions between servers reveal more share identifiable information such as share number, expiry date, size among others. Thus the design only offers partial document anonymity from external attacks not from collaborating ServNet nodes and as such, document anonymity is hard to achieve in the current circumstances.

The claim that passive-server document anonymity property of FHS can protect a ServNet node against legal action is unrealistic. No technological means can supercede the laws of the land even in highly democratized societies plausible denial of data stored on ones server can not easily be an excuse against wrong doing in society and thus prosecution and arrests and other form of social and legal harassments are still possible.

The current design has no realist protection against "share hoarding attack" [1, 7]. The assumption that a collaborating group of individual server operators will find it expensive to obtain a fraction of a document due to the size of the network can not be realistically accepted. Dedicated and well facilitated spy organizations like the central intelligence agency (CIA) and terrorist organization like Al Qaeda can find enough resources to fund and form such expensive group.

Server anonymity is hard to achieve against bad hosts in the system and bad guys can easy join the system as they is no robust protocol defined to evaluate the reputation of new members, the reliance on introducers is not enough as introducers themselves mighty be collaborating with bad guys. The design of FHS

promises a lot, but in the current environment, implementing such a design is not easy and some of the assumptions do not pass some credibility tests.

We note that, the current design of FHS does not support document revocation as possessing extra knowledge to allow document revocation makes one a target of physical and legal attacks [1, 3 and 7]. However, in some case document revocation may be a desired property.

B. Successes

In comparison with its predecessors such as Mojo Nation, Napster, Gnutella, and FreeNet, FHS has a number of recommendable successes in its design even at its infant stage. Some of the components of the system are built on strong logical concepts of its predecessors.

In comparison with FreeNet and Gnutella, FHS maintains a document in the system even if the document is unpopular; this is one of the contributions made by FHS team to the study of anonymous storage and publication systems [1, 7].

Unlike Mojo Nation and Napster which have some of their system services centralized (like content tracker and publication tracker in Mojo Nation and central indexer in Napster), FHS is a purely decentralized system which can not suffer from a single point of failure. Peers have both serve and client functionality and they rely on the reputation and accountability systems to monitor the integrity of peers and to find which peers to trust.

Based on its trading scheme and high number of ServNet node in different jurisdictions (assumption), then FHS offers more publisher anonymity then any of its predecessors and this has been the reason why KaZaA has remained put amidst legal pressures from MPA and RIAA. If the trades are frequent and the number of servers is in hundreds of thousands say, it's not easy for an adversary to assume that the server trading a given share is the publisher of the document. The trading also provides a moving target to would be adversaries, thus improving the confidentiality of share and document anonymity. The trading scheme also protect the system from malicious share flooding as a server can only send into the network as much information as it can store.

By using selective query flooding [1, 7 and 12], FHS performs better than most of its predecessors in network resource optimization. When retrieving a document, the reader randomly selects a server his knows and sends out his request. This improves the over all system perform as other channels are made available to other communicating nodes, which is not the case with say FreeNet, Eternity service and Mojo Nation which relies on query broadcasts.

Based on FHS perfect forward anonymity (a system is said to offer perfect forward anonymity if no identifiable information of participants remains after a transaction is complete). It works on the premise that no key used for the transfer of data may be used to derive any keys for future transmission. FHS achieves author anonymity because authors communicate with publishers via an anonymous channel and share contains no extra information about authors.

However this is only true if authors behave and they don't give extra identifiable information in the documents they write.

C. Suggestions

We suggest that shares of a given document should be stored based on a quota (threshold) scheme per server and per jurisdiction. A given server should not store beyond a given percentage of shares of a given document (say 15%) and at least every share must have one of its buddies in every jurisdiction.

Buddies should have a short life span and nodes should not be allowed to associate with the same nodes over and over again. This will decrease the chances of buddy co-opting by a malicious server from happening. Of course short buddy life span means more network reconfigurations and hence more computational resources. Therefore a compromise should be made between computational efficiency and system robustness.

Since servers in the system maintain share identifiable information of which shares they have traded away and received, we can exploit this important information to improving efficiency of document retrieval by adapting the Intelligence Search Algorithm (ISA) [10, 12] instead of the current random walk algorithm.

Conclusions and Future

In all fairness one would say that, the chances of success of such a complex anonymous publication and storage system are slim. The greatest challenge of all is the requirement of implementing strong anonymity and strong authentication. These two properties are like an egg and chicken puzzle. Making one strong leads to the other being weak besides, the current design of FHS is partially anonymous and preventing the adversary from obtaining user identifiable information like their bandwidth, geographical location of the place where they live and share locations is practically impossible. The type of threats the system faces are very complex and adversaries can deploy extensive techniques to undermine the success of the system.

However, on the bright side, the popularity of Napster and Gnuttala systems provides hope that if technical and social handles are minimized to acceptable levels, The Free Haven service can be a success since the potential of attracting a critical mass of users is there. We have to mention that FHS design introduced more concepts in the area of anonymous storage systems, like the enhancement of publisher anonymity, server anonymity, document anonymity, query anonymity and reputation system based on buddies [1, 7].

The possibility of attacks through compromised or malicious servers hasn't been considered well in the current design and therefore requires further understanding and many of the issues that apply to other anonymous storage service like FreeNet do apply to FHS. Therefore, more research is needed in the direction of improving the accountability and reputation systems without comprising the robustness

and anonymity of servers, document and users. Our consensus is that under the current assumptions the use of discretionary distribution of shares in FHS is an improvement to the design of distributed anonymous storage systems.

It is clear from the literature that FHS promises a lot, but implementing such a system in a real world is still a long way off, the core issue being the provision of anonymity. Guaranteeing anonymity is very hard since it largely depends on building trust along the communication channel which in FHS design the communication channel is public subjected to all sorts of attacks. The failure of its predecessors due to legal pressures and lack of general acceptability in the society cast doubts on the successful implementation of Free Haven. It is our view that Roger Dingledine and his team have to faces an uphill task of achieving their dream.

References

Roger D, J.F Michael and M David "The Free Haven Project; Distributed Anonymous storage Service", International workshop on desiging privacy enhancing technologies: design issues in anonymity and unobservability, Berkeley, CA, USA, pp 67-95, Springer-Verlag, NY

Ross J. Anderson "The Eternity Service". In the proceedings of Pragocrpt 1996, Cambridge University Computer Laboratory

Ian Clarke "the free network project." http://freenet.sourceforge.net accessed on September, 30th 2006

Michael O. Rabin, "Efficient dispersal of information for security, load balancing, faulty Tolerance", Journal of the ACM (JACM) Vol. 36 , Issue 2 (April 1989) pp 335 - 348

David Chaum "Security Without Identification: Transaction Systems to Make Big Brother Obsolete". Commun. ACM 28(10): 1030-1044 (1985)

Y.Gertner, Y. Ishai, E. Kushilevitz and T. Malkin, "Protecting Data Privacy in Private, Proceedings of the thirtieth annual ACM symposium on Theory, 1998

Information Retrieval Schemes" http://theory.lcs.mit.edu/~cis/pir.html accessed on July 20th, 2006

FreeNet systems, www.freehaven.net accessed on August, 2006

The onion routr, http://www.onion-router.net/Publications.html accessed on July 22, 2005

Jean-François Raymond: Traffic Analysis: Protocols, Attacks, Design Issues, and Open problems. Proc. Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability (2000): Vol. 2009, pp 10-29

Demetrios Zeinalipour-Yazti "Information Retrieval in Peer-to-Peer Systems" Masters Thesis , June 2003, citeseer.ist.psu.edu/article/zeinalipour-yazti03information.html, accessed, November 5th, 2006

F. Otto and S. Ouyang: Improving Search in Unstractured P2P Systems: Intelligent Walks (I-Walks). In proceeding of 7th IDEAL conference; Sept 2006, Burgos, Spain; LNCS 4224, pp 1312-1319, springer.

30
Subscriber Mobility Modeling in Wireless Networks

Tom Wanyama

In this paper, a simplified model for user mobility behavior for wireless networks is developed. The model takes into account speed and direction of motion, the two major characteristics of user mobility. A user mobility simulation based on the model is developed, and its results are compared with those in published work. The transient performance metrics of a mobile network are analyzed in terms of trajectory prediction, mean location update rate, and new/handoff call residence time distribution. The proposed mobility model yields results that match very well with those obtained from previously reported complex mobility models.

Introduction

The challenge of supporting rapidly increasing number of mobile subscribers, while constrained by limited radio spectrum, is being met through increasingly smaller radio cell sizes. However, this results in increased signaling for location management procedures, which reduces the bandwidth available for user traffic, as well as additional transmission and processing requirements on the mobile network. The fundamental procedures that make up the basis for location management are location updates and pages. At one extreme, the location of a subscriber is maintained on a per cell basis. Whenever the mobile terminal moves to a new cell, which may happen very frequently in case of an automobile-mounted mobile terminal, a location update is triggered. This is clearly inefficient in terms of bandwidth and base station processing power usage. However, paging messages need only be sent to one cell, since the exact location of the subscriber is known. At the other extreme of location updating, all the cells in the network belong to one location area. In this case, location updates are not required at all but, for every incoming call, the network must page every cell. This approach is also unsatisfactory due to high paging traffic processing. To improve the situation described above, different location management techniques have been proposed to reduce the signaling required to locate any active subscriber on the network (Cho, 1994), (Hong et al., 1986). All such techniques make use of subscriber mobility models to determine the cells where the subscriber is most likely to be located. Mobility models also play an important role in examining different issues involved in wireless mobile networks such as handover, offered traffic, dimensioning of signaling network, user location update, paging, and multi-service provider network management.

In the general case, a mobility model should take into consideration changes of both speed and direction of a mobile terminal. Since the moving direction and speed of a mobile terminal are both random variables, the path of a mobile terminal is a random trajectory. Tracing this trajectory requires a systematic formulation of geometrical, speed and time relations that govern the complex problem of random movement (Liu et al., 1998). A review on the available literature on this subject reveals that Hong and

Rappaport (Hong et al., 1986) modeled the mobility with random direction motions in two-dimensional environments. Cho (Cho, 1994), studied the effect of terminal mobility in rectangular shaped urban micro cellular systems, and Kim et al. (Kim et al., 2000) estimated the number of handoffs in three-dimensional indoor environments. However, these studies were mainly concerned with the problems of handoff as

opposed to location management. Markoulidakis et al. (Markoulidakis et al., 1997) proposed a mobility model based on the transportation model. This model generalizes the mobility behavior of people, hence making it difficult to use the model to study mobility management procedures in a system. Therefore, the model is only suitable for planning purposes. Zonoozi and Dassanayake (Zonoozi et al., 1997) proposed a

mobility model, that takes into account most of the possible mobility-related parameters. Pollini et al. (Pollini et al., 1995) evaluated the signaling traffic volume for mobile and personal communications. They (Pollini et al., 1995) considered the simple terminal mobility model and evaluated the performance of mobile networks in terms of the mean number of location updates or handoff calls at steady state.

However, terminal mobility behaviors can vary according to time periods. For example, during an officegoing hour, most of the mobile users move from home to offices, and they return home from their offices during the office closing-hour. Mobility management traffic also varies according to the users' initial positions. Therefore, characterization of mobility management traffic in the transient period is sometimes

more important than that of steady-state. Most models proposed in the above-cited literature are complicated, and require a lot of computing resources and time, making them unsuitable for real-time purposes. There is also a tendency of working backwards, that is, mobility attributes such as velocity or the cell residence time are assumed to follow a particular distribution without any real-life data used to justify the assumptions.

The focus of this paper is the proposal of a simple but yet realistic mobility model for wireless networks. The model takes into account the two major characteristics of user mobility, namely, randomly changing speed and randomly changing direction. The simplicity of the proposed model makes it suitable for use in real-time simulation situations. The model produces reliable results because at every moment in time, it considers four major mobility attributes: direction,

position, speed and acceleration. Unlike previous work, no particular distribution is assumed for any of the attributes of mobility or the associated network performance metrics. Therefore, the model is based on the basic principles of mobility. A user mobility simulation based on the model is developed, and the transient performance metrics of a mobile network are analyzed in terms of trajectory prediction, mean location update rate, and new/handoff call residence time distribution.

The outline of this paper is as follows. In section 2, the structure of location areas of mobile networksand the description of terminal mobility are introduced. The section also deals with the analysis of trajectory prediction, mean location update rate, and new/handoff call residence time distribution. Section 3 covers the performance evaluation of the proposed mobility model, and the results obtained are compared with those in published literature. Finally, section 4 presents the main conclusions.

Proposed Mobility Model

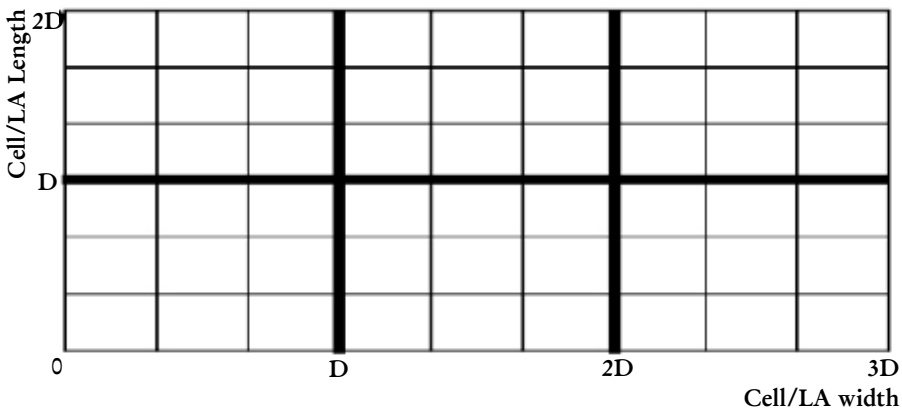
In this paper, the following assumptions are made for computational and geometrical simplicity.

A1) Square cell shape, each cell is of size d

A2) Each location Area (LA) has a square shape of size D.

Figure 1 shows an example of the arrangement of cells and LAs in a given service area. First note that the behaviors of mobile users can vary according to time periods in a day. For example, during the officegoing hour, most mobile users move from their homes to offices, while during the office-closing hour the users move from offices to their homes.

Fig 1: The assumed wireless service area



However, some mobile users (for example, salesmen) move from place to place during the daytime. Based on the foregoing, user mobility is modeled considering these user behaviors, as reflected in the following assumptions:

A3) Movement direction changes occur after a fixed time interval.

A4) Mobile users change their direction with probabilities p_x, q_x, p_y and q_y for forward, backward, up and down directions, respectively, where, $p_x + q_x + p_y + q_y = 1$.

A5) The acceleration of a mobile user is a random variable that is correlated in time.

Assumption A4) follows from the assumed square shape for cells and location areas.

Trajectory Prediction. In predicting the trajectory of motion for a user, the first step is to specify the probabilities px, qx, py and qy. Let (v, w) denote the current location of a user whose destination location (selected randomly in the service area) is denoted by (i, j). Expressions for calculating the probabilities p_x, q_x, p_y and q_y of selecting directions are given in Table 1. These expressions are derived assuming knowledge of user destination, and that users take the shortest route to destination. Secondly, the probability distribution takes into account the fact that some users may choose to take a longer route to the destination. The four moving directions, namely; forward, backward, up and down are boundaries to four possible quadrangles into which a user may move. To take the shortest distance to destination, the user must move in a direction whose x (i.e., forward or backward) and y (i.e., up or down) components are in the same directions as the two moving directions that make up the boundary into which the destination coordinates are located. For example, if v > i and w > j , moving in forward or up directions means moving further away from the destination. Hence, the user must move, with a high probability, in directions with backward and down components in order to reach the destination. In other words, the forward and upward moving directions have small moving probability which we assume to be equal, that is, $p_x = p_y$. By the total probability theorem and from Assumption A4:

$$p_x = p_y = 0.5\ 1\ [1 - (q_x + q_y)] \quad (1)$$

Equation (1) is the entry in row 1, column 2 of Table 1. The other entries in rows 2 to 4 of column 2 corresponding to other possible user's current location relative to the destination are derived in the same manner. It now remains to derive the expressions for q_x and q_y. First note that dx and dy are the respective x and y components of the distance between a user's current location and the destination.

Now, if dx > dy the user must give a higher priority to moving in the direction of the longer of the two distances. Therefore, x y $q_x > q_y$ whose expressions (row 1, column 3 of Table 1) are given by:

$$q_x = 0.4 + 0.3\ (dx - dy)/dx] \quad (2a)$$

$$q_y = 0.4 + 0.2\ (dx - dy)\ /dx] \quad (2b)$$

An explanation of eqn. (2) is given as follows: Based on the condition that $q_x > q_y$, the multiplicative factor in the second term on the right hand side of eqn. (2b) is not equal to that of eqn. (2a) and this is intended to address the fact that

priority given to one moving direction is not always equal to the resentment towards another direction. If at some moment during the movement dy becomes longer than dx , then q_y and q_x will trade places, as shown by the entries in row 2, column 3 of Table 1. The other entries in rows 3 to 8, column 3 of Table 1 are developed using similar reasoning. The foregoing demonstrates the variation of the probabilities representing user mobility behavior with respect to user's current location relative to the destination. In addition, the probabilities can also vary according to time periods. For example, during the office-closing hour, a user generally returns home. Therefore, the probability of selecting a moving direction towards his/her home is higher than those of the other directions.

Table 1: Moving direction selection probabilities of a user at location (v,w) with destination location (i,j)

Current Location	Moving Direction Selection Probabilities		Condition
(v,w) for *v>i* and *w>j*	$p_x = p_y = 0.5[1-(q_x+q_y)]$	$q_x = 0.4 + 0.3[(dx-dy)/dx]$ $q_y = 0.4 - 0.2[(dx-dy)/dx]$	*dx>dy*
		$q_y = 0.4 + 0.3[(dy-dx)/dy]$ $q_x = 0.4 - 0.2[(dy-dx)/dy]$	*dy>dx*
(v,w) for *v>i* and *w<j*	$p_x = q_y = 0.5[1-(q_x+p_y)]$	$q_x = 0.4 + 0.3[(dx-dy)/dx]$ $p_y = 0.4 - 0.2[(dx-dy)/dx]$	*dx>dy*
		$p_y = 0.4 + 0.3[(dy-dx)/dy]$ $q_x = 0.4 - 0.2[(dy-dx)/dy]$	*dy>dx*
(v,w) for *v<i* and *w>j*	$p_y = q_x = 0.5[1-(q_y+p_x)]$	$p_x = 0.4 + 0.3[(dx-dy)/dx]$ $q_y = 0.4 - 0.2[(dx-dy)/dx]$	*dx>dy*
		$q_y = 0.4 + 0.3[(dy-dx)/dy]$ $p_x = 0.4 - 0.2[(dy-dx)/dy]$	*dy>dx*
(v,w) for *v<i* and *w<j*	$q_y = q_x = 0.5[1-(p_x+p_y)]$	$px = 0.4 + 0.3[(dx-dy)/dx]$ $py = 0.4 - 0.2[(dx-dy)/dx]$	*dx>dy*
		$py = 0.4 + 0.3[(dy-dx)/dy]$ $px = 0.4 - 0.2[(dy-dx)/dy]$	*dy>dx*

Note: In the Table, dx = | v-i | , dy = | w-j |

Having specified the movement probabilities, the next step in predicting the trajectory of motion is to derive the dynamic equations for continuous-time movement, presented as follows. In two-dimensional Cartesian coordinates, subscriber movement can be described by a vector equation of the form **R**(t) = [x(t), v_x (t), y(t), v_y (t)] where *x*(t) and y(t) represent the position at time t, and v_x(t) and v_y(t) represent the relative speed along the x and y direction during the time interval dt between times t and t+dt.

Furthermore, let $\mathbf{A}(t) = [a_x(t), a_y(t)]$ denote the two-dimensional random acceleration vector. From Assumption A5), it follows that if a mobile is accelerating at a rate $\mathbf{A}(t)$ at time t, it continues to accelerate at this rate for a small time interval dt. Hence, the relative movement in the x and y directions is described using the equations of motion at constant acceleration for every small time interval dt. The velocity in the *x* direction, $v_x(.)$, is given by:

$$v_x(t + dt) = v_x(t) + a_x(t)dt \tag{3}$$

The distance Sx covered in the x direction during the time interval dt is determined by the equation:

$$S_x(t) = \begin{cases} \dfrac{[v_x(t+dt)]^2 - [v_x(t)]^2}{2a_x(t)}, & a_x(t) \neq 0 \\ v_x(t)dt, & a_x(t) = 0 \end{cases} \tag{4}$$

Let P(*x*) denote the probability of the direction selected along the x direction. That is, P(*x*) equals p_x or q_x for the forward or backward direction, respectively. The expected distance covered in the x direction, $E[S_x(t)]$, is then given by:

$$E[S_x(t)] = P(x)S_x(t)$$

By replacing x with y in eqns. (3) to (5), the corresponding equations of motion in the y direction are obtained. The expected distance covered during time interval dt can then be written as:

$$E[S(t)] = \begin{cases} E[S_x(t)] \\ E[S_y(t)] \end{cases} \tag{6}$$

Equation (6) is the main result for predicting the trajectory of a mobile user during the time interval *dt*.

Location Update. Location update (LU) is the process by which a mobile terminal (MT) makes its position known to the network. This is performed either by zone, distance or timing method. The zone method is considered in this paper. Zone based location updating requires splitting of the whole service area into zones called location area (LA) as shown in Figure 1. Location updating is performed every time the MT crosses the LA boundary. Mobility and LA size are related to LU through the location update rate,
$\lambda_{LU}(t)$, given by:

$$\lambda_{LU}(t) = \frac{E[S(t)]}{D}$$

where E[S(t)] is given by eqn. (5). Clearly $\lambda_{LU}(t)$ is time-dependent.

Handoff. When a mobile user engaged in conversation travels from one cell to another cell, the call must be transferred to the new base station to prevent abrupt termination of the call. This ability for call transfer is referred to as handoff in mobile cellular systems. Handoff is related to mobility and cell size through a parameter called handoff rate, λ_{HO} (t) , and is written as

$$\lambda_{HO}(t) = \frac{E[S(t)]}{d}$$

Therefore, the expected handoff rate E[λ_{LU} (t)] is then given by

$$E[\lambda_{HO}(t)] = P_{HO}\lambda_{HO}(t)$$

where PHO is the handoff probability associated with either the new call residence time or handoff call residence time. New call residence time is defined as the length of time that a call originating from a cell stays in that cell before crossing into another cell. In other words, handoff is initiated when the call duration exceeds the new call residence time. Similarly, handoff residence time is the time a handoff call stays in a cell before crossing into another cell. That is, handoff of a previously handoff call is triggered when the call duration is greater than the handoff call residence time. The next step in the analysis is to calculate the expressions for the handoff probability associated with new call and handoff call residence times.

Handoff Probability due to New Call Residence Time. The probability density function for the new call residence time, α(t), is given by [3]:

$$\alpha(t) = \begin{cases} \dfrac{8R}{3\pi V_m t^2}\left\{1-\left[1-\left(\dfrac{V_m t}{2R}\right)^2\right]\right\}^{\frac{3}{2}}, & 0 \le t \le \dfrac{2d}{V_m} \\ \dfrac{8r}{3\pi V_m t^2}, & t > \dfrac{2d}{V_m} \end{cases} \quad (10)$$

where Vm and R are the speed and cell radius respectively d is the cell diameter which we approximate to the cell length d in the proposed mobility model. Equation (10) is derived under the assumptions of constant speed and one direction of movement in an hexagonal-shaped cellular structure. We adopt eqn.

(10) in this paper but replace the constant speed with mobile expected speed and also select the cell size assuming the square cell area is equal to the hexagon area. The handoff probability due to new call residence time is determined by numerical integration of the modified version of eqn. (10).

Handoff Probability due to Handoff Call Residence Time. By definition, a handoff call is itself handed off if the call duration, Tcall exceeds the handoff call residence time, Tres. The handoff probability due to handoff call residence time, b is given by:

$$\beta = Pr\{T_{call} > T_{res}\}$$

It turns out that both Tcall and Tres are random variables so that is determined from knowledge of their respective distribution functions. Voice call duration is assumed, in most analysis, to be exponentially distributed whereas the distribution for cell residence time is strongly dependent on the mobility model (Fang et al., 2002). Since our proposed mobility model makes no assumption regarding cell residence time, eqn. (11) cannot be used. Instead, in this paper, we assume that the probability density function (pdf) for

handoff call residence time is a scaled version of that for new call residence time where the scaling factor M is greater than unity. A scaling factor greater than unity is selected to minimize the number of interruptions experienced by previously handed off calls.

Performance Evaluation

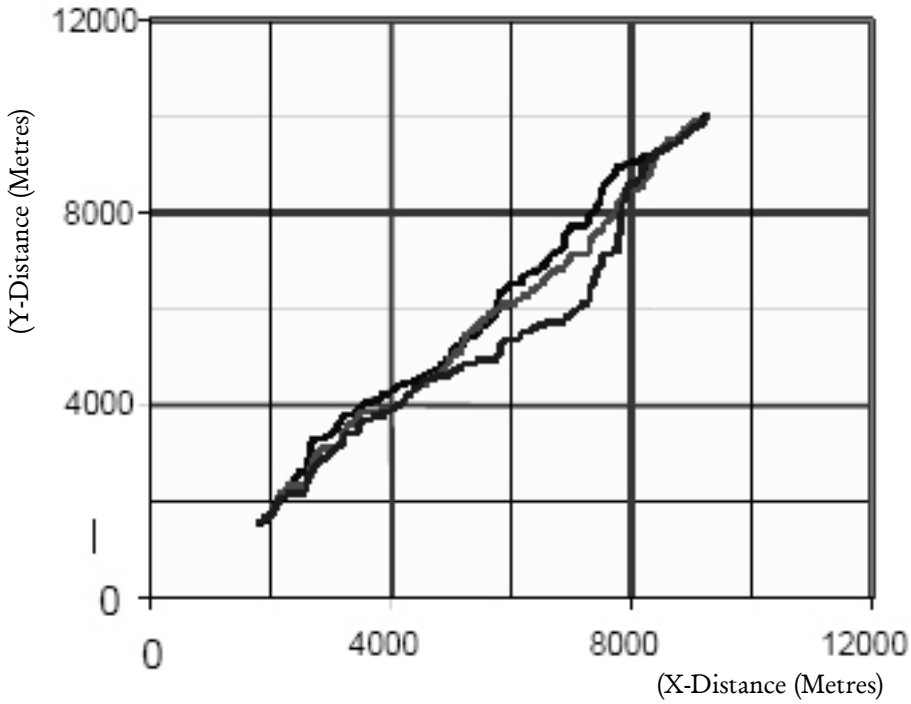
In this section, performance of the proposed mobility model is evaluated using Monte Carlo simulation approach making use of the equations derived in section 2. A number of simulations are carried out to determine how well the simulation tool predicts the cells the user in most likely to cross. Secondly, other mobility parameters are deduced from the simulations. Simulation results are generated assuming the following network and mobility parameter values:

- Size of service area: 12 km x 12 km
- Cell of square shape, d = 4 km
- Four cells per location area
- Acceleration is selected randomly in the range: [-5, 5] m/s2.
- Speed is selected randomly in the range: [30, 60] km/hr.
- Scaling factor for pdf of handoff call residence time, M: 2

Simulation results are discussed under appropriate sub-headings.

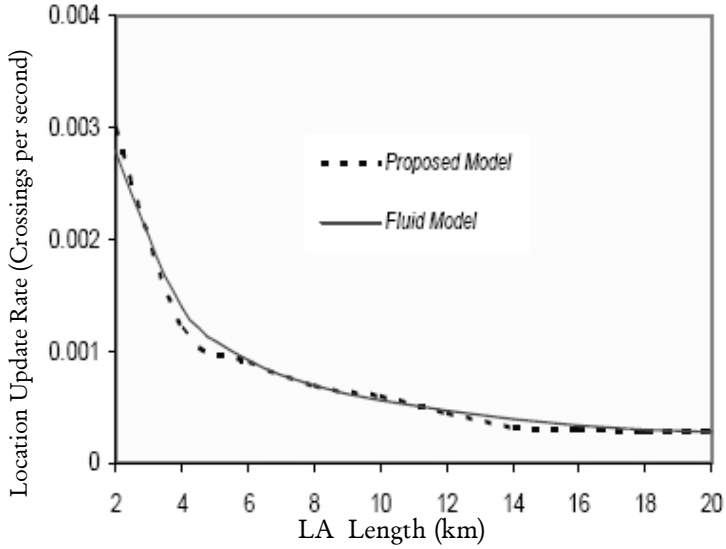
Trajectory Prediction. Given the initial location and the destination of a subscriber, the model was simulated several times to determine the subscriber trajectory. Fig. 2 shows the predicted trajectory in three of the simulation runs. Fig. 2 is useful for cellular network design because it predicts the cells that a mobile terminal in motion will pass through, so that resources are reserved to serve the MT in case it arrives at the cell boundary while engaged in a call.

Fig 2: Predicted motion trajectory for three mobility simulations



Location Update Rate. Figure 3 shows the calculated location update rate plotted as a function of the location area (LA) size. The results indicate that location update rate can be modeled as a decaying exponential function. Also shown in Figure 3 is the calculated location update rate based on the fluid flow mobility model (Thomas et al. 1988). There is very good agreement between the predicted location update rates by the two models.

Fig 3: Variation of location update rate with LA size



Handoff Rate. As a prelude to presenting the calculated handoff rate, we first show in Figure 4 the probability distribution for new call residence time obtained using our simplified mobility model compared with that of a previously proposed model by Zonoozi and Dassanayake (Zonoozi et al., 1997). The results presented in Fig. 4 are generated for a cell size d of 4 km. For new call residence times less than 2.5 minutes, the proposed model predicts a higher handoff probability than that of Zonoozi and Dassanayake's model where the new call residence time is assumed to have the gamma distribution function (Zonoozi et al., 1997). It is interesting to find that, beyond new call residence time of 2.5 minutes, our results agree with those in Reference (Zonoozi et al., 1997), that new call residence time follows the gamma distribution.

Fig 4: Probability distribution for new call residence time

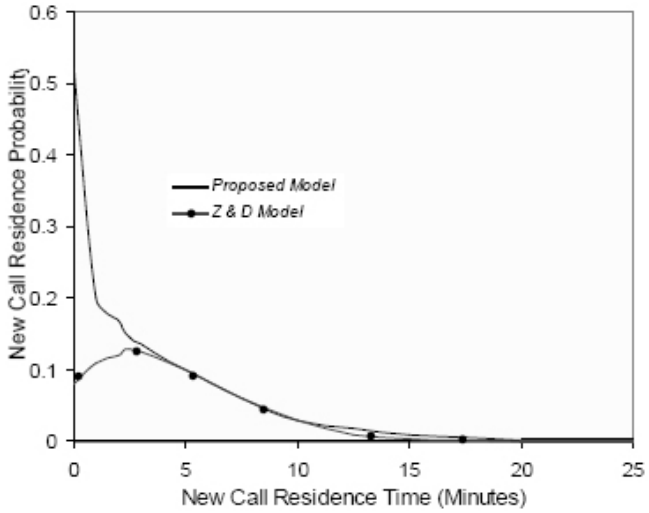
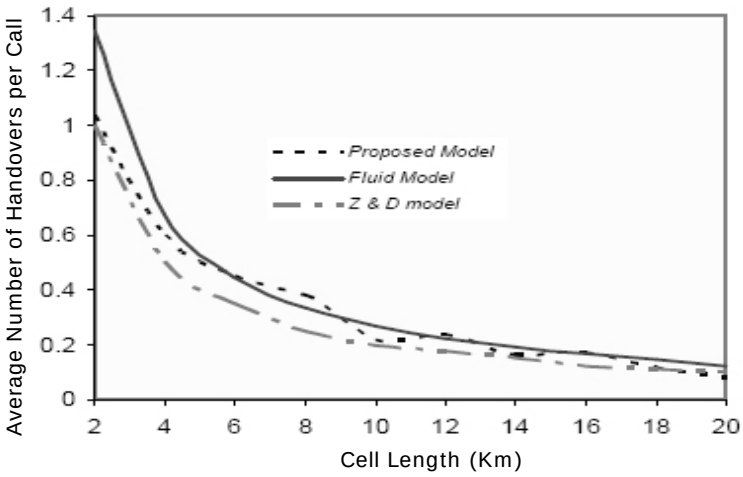


Figure 5 depicts the average number of handoffs per call determined by the proposed mobility model. Results obtained with Zonoozi and Dassanayake's (Zonoozi et al., 1997) and fluid flow (Thomas et al. 1988) models are shown for comparison. In the calculations, the average call duration call T is assumed to be 120 seconds. The results shown in Figure 5 assume the worst case scenario that all moving mobile terminals are engaged in a call. It is seen that the handoff rates predicted by the proposed mobility model are roughly equal to those predicted by the fluid flow model but are slightly higher than those predicted using Zonoozi and Dassanayake's model. Lower handoff rate is achieved with Zonoozi and Dassanayake's model because the model predicts a lower average speed than that obtained with the proposed model.

Fig 5. Variation of average number of handovers with cell size (assumed average call duration = 120 secs.)



Conclusions

A simple mobility model that does not assume any specific distribution for the underlying mobility parameters is proposed. The mobility model is based on user speed and direction of motion selected according to user behavior. Analytical expressions are derived for the expected distance traveled as function of time along with the expressions for network performance metrics related to mobility management. A Monte Carlo simulation of the proposed mobility model was conducted and results are compared with those from previous work in the published literature. It is concluded from the results that location update rate can be represented by a decaying exponential function, and the new call residence time (exceeding 2.5 minutes) follows the gamma function. Agreement of the results from the simplified model demonstrates its usefulness as a viable candidate for location management planning in practical mobile radio networks.

References

Cho H. S. (1994). "Analysis of Signaling Traffic Related to Location Registration/Updating in Personal Communication Networks", MS thesis KAIST, Korea

Fang Y., and I. Chlamtac (2002). "Analytical Generalized Results for Handoff Probability in Wireless Networks", IEEE Transactions on Communication, Vol. 50, no. 3, pp. 396-399

D. Hong D., and S. S. Rappaport (1986). "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures", IEEE Transactions on Vehicular Technology, Vol. VT-35, No. 3, pp. 77- 92

Markoulidakis J. G., G. L. Lyberopoulos, D. F. Tsirkas, and E. D. Sykas (1997). "Mobility Modeling in Third-Generation Mobile Telecommunications Systems", IEEE Personal Communications, Vol. 4, No. 4, pp. 41-56

Liu T., P. Bahl, and I. Chlamtac (1998). "Mobility Modeling. Location Tracking, and Trajectory Prediction in Wireless ATM Networks", IEEE Journal on Selected Areas in Communication, Vol. 16, No. 6, pp. 922-936

Kim T. S., J. K. Kwon, and D. K. Sung (2000). "Mobility Modeling and Traffic Analysis in Three-Dimensional High-Rise Building Environment", IEEE Transactions on Vehicular Technology, Vol. 49, No. 5, pp. 1633-1640

Pollini G. P., K. S. Meier-Hellstern, and D. J. Goodman (1995). "Signaling traffic Volume Generalized by Mobile and Personal Communication, IEEE Communication Magazine, Vol. 33, No. 6, pp. 60-65

Zonoozi M. M., and P. Dassanayake (1997). "User Mobility Modeling and Characterization of Mobility Patterns", IEEE Journal on Selected Areas in Communication, Vol. 15, No. 7, pp. 1239-1252

Thomas, R., H. Gilbert and G. Mazziotto (1988). "Influence of the movement of the mobile station on the performance of the radio cellular network", in Proc. 3rd Nordic Seminar, Copenhagen, Denmark, pp. 94-106

31
An Evaluation Study of Data Transport Protocols for e-VLBI

Julianne Sansa

This paper compares TCP-like data transport algorithms in the light of e-VLBI requirements and proposes HTCP with bandwidth estimation (HTCP-BE) as a suitable candidate by simulating its behaviour in comparison with seven existing TCP variants; HighSpeed TCP for large Congestion Window (HSTCP), Scalable TCP (STCP), Binary Increase TCP (BI-TCP), Cubic TCP (CUBIC-TCP), TCP for Highspeed and long-distance networks (HTCP), TCP Westwood+ (TCPW) and standard TCP (TCP). Using average throughput efficiency and stability as the performance metrics we show that HTCP-BE better suits e-VLBI needs than any of the other seven protocols in environments with random background traffic.

Introduction

The European VLBI Network (EVN) is an array of radio telescopes located throughout Europe and as far away as China and South Africa. These radio telescopes produce data at rates of up to 1 Gbps each. Until recently, these data streams were recorded on tapes, nowadays on hard disk drives, and shipped to the correlator located at JIVE, the Joint Institute for VLBI in Europe, in Dwingeloo, the Netherlands. During the last few years JIVE, in collaboration with the European National Research Networks and the pan-European Research Network GEANT, has worked on a proof-of-concept (PoC) project to connect several telescopes across Europe in real-time to the correlator via the Internet (electronic VLBI or e-VLBI). This project has led to an EC sponsored project called EXPReS, which over the next few years will transform the EVN to a fully functional real-time e-VLBI network. During the PoC project it became clear that in spite of the vast capacity of the connecting networks, the actual transport of large data streams poses quite a challenge. The most critical requirement for e-VLBI transport is delay-sensitivity with the advantage of being loss tolerant.

The Mark5 [7] application that handles e-VLBI data uses the Transport Control Protocol (TCP). By its nature, e-VLBI involves transporting huge amounts of data via the Internet over long distances from geographically dispersed telescopes to one central correlator. TCP is somewhat problematic in combination with long distance high speed links [1, 4, 17]. Having identified the congestion control algorithm of TCP as the bottleneck under certain circumstance as reported in [15], we investigate e-VLBI data transport with several TCP-like transport protocols. Based on the observations we then propose yet a new congestion control algorithm (a minor modification to HTCP), which suits e-VLBI requirements better than the

other seven protocols considered. In the next section we give a brief explanation about the congestion control algorithms of the TCP-like protocols we are evaluating including our own HTCP-BE, then we explain the setup of the simulations, after which we present the performance evaluation and conclude.

TCP-like CONGESTION CONTROL ALGORITHMS

In this section we present the congestion control algorithms and response functions of standard TCP and its recent modifications. We categorize them in two, the non-adaptive and the adaptive algorithms.

Non-Adaptive TCP algorithms

These are described as the algorithms which modify the CWND with increase and decrease factors (and respectively) which are functions of the value of CWND at the point of modification independent of any previous CWND modifications.

This results in CWND values that are changed by specific pre-defined percentage and CWND modifications that are oblivious of each other. In the following subsections we describe the algorithms which are non-adaptive.

Standard TCP [17]

The TCP congestion avoidance algorithm is presented in the equations (1) and (2), while the resulting response function is equation (3)
On Acknowledgement:

$$CWND \leftarrow CWND + \alpha/CWND \tag{1}$$

On Packet Loss Indication:

$$CWND \leftarrow CWND - \beta \times CWND \tag{2}$$

Where $\alpha = 1$ and $\beta = 0.5$

$$CWND_{average} = \frac{1.2}{p^{0.5}} \tag{3}$$

Where p is the steady state packet loss rate.

Highspeed TCP (HSTCP) [4]

HSTCP's congestion avoidance algorithm is similar to that of standard TCP as presented in the equations (1) and (2), but differs in the values of α and β.

HSTCP obtains α and β from a pre-computed table based on the network's bandwidth delay product (BDP) in such a way that if its present value is below a certain pre-set threshold termed low window, HSTCP sets $= 1$, $\beta = 0.5$ thus reverting back to standard TCP congestion control algorithm. However if the BDP value exceeds the low window, α and β are then assigned from the table. For our network in steady state the appropriate $\alpha = 26$ and $\beta = 0.22$.

HSTCP then presents a response function shown in equation (4), which is able to reach a higher CWNDaverage than that reached by standard TCP in equation (3) but takes a longer time to converge.

$$CWND_{average} = \frac{0.12}{p^{0.835}} \quad (4)$$

Where p is the steady state packet loss rate.

Scalable TCP (STCP) [8]

STCP's congestion avoidance algorithm is a modification of HSTCP and is presented in the equation (5) for each packet acknowledged and equation (2)for each packet lost.

On Acknowledgement:

$$CWND \leftarrow CWND + \times CWND \quad (5)$$

The resulting STCP response function is as shown equation (6) which reaches a higher

$CWND_{average}$ than that reached by HSTCP in equation (4) for the same steady state packet loss rate. STCP however takes a longer time to converge than HSTCP.

$$CWND_{average} = \frac{0.038}{p^{1}} \quad (6)$$

Where p is the steady state packet loss rate.

Binary Increase TCP (BI-TCP)[18]

BI-TCP is a semi-adaptive algorithm, which uses a combination of the binary search increase and the additive increase for incrementing the CWND on arrival of acknowledgements when the current CWND exceeds a preset low window threshold.

Should the CWND value be below the low window threshold the standard TCP CWND increment is reverted to. Binary search increase follows the binary search algorithm concept and sets the next CWND value to a value halfway between its current value and currently known maximum. The binary search increase is used when the CWND increment is small while the additive increase is used when the CWND increment is large. Occasionally when the known maximum window is exceeded, BI-TCP goes into a slow start phase in which it probs for a new maximum. The BI-TCP CWND increment is summarised in the equation (7).

$$CWND \leftarrow \begin{cases} CWND + 1/CWND & CWND < lw \\ CWND + (tw - CWND)/CWND & tw - CWND < Sx \\ CWND + Sx/CWND & tw - CWND > Sx \\ CWND + SS\text{-}CWND/CWND & CWND > Wx \end{cases} \quad (7)$$

Where *lw* is the CWND threshold beyond which BI-TCP engages otherwise standard TCP is used, tw is the midpoint between the maximum and minimum window sizes, S*x* is the maximum increment, SS CWND is a variable to keep track of CWND increase during the BI-TCP slow start and W*x* is the maximum window size; initially set to a default large integer.

BI-TCP CWND increase is semi-adaptive as it follows adaptive patterns implicitly. It progresses through binary increase, additive increase and finally slow-start probing, each of which relies on the previous episode receiving acknowledgements.

In addition these functions consecutively increase the CWND in greater proportions than the previous function.

On detection of packet loss BIC reduces it's CWND based on multiplicative decrease pattern following equation (8). This decrease function is non-adaptive.

$$\begin{aligned} &\text{CWND} \leftarrow \text{CWND} * (1 - \beta) && \text{CWND} \geq \text{low}_{window} \\ &\text{CWND} \quad \text{CWND} * 0.5 && \text{CWND} < \text{low}_{window} \end{aligned} \tag{8}$$

Where the decrease factor is fixed at 0.125

Adaptive TCP algorithms

These we describe as the algorithms that modify the consecutive CWND differently considering a variable related to the previous packet loss event e.g. the time elapsed since the last packet loss event or the throughput attained just before the last packet loss event. They use these variables (duration and /or throughput) to gauge the changing level of congestion, hence modifying the CWND not by a fixed percentage but to a value that indicates the prevailing congestion conditions.

The increase and decrease factors are functions of either the time elapsed since the last congestion event or the throughput just before the last congestion event. In the following subsections we discuss some of the recently proposed algorithms which are adaptive.

Westwood TCP (TCPW) [5, 11, 10]

TCPW connection establishment is based on standard TCP (as in equation(1)) until a packet loss is encountered. At that point the adaptive decrease is invoked by setting the ssthresh (and thus CWND) to a value that is larger than the corresponding value set by standard TCP. ssthresh is set based on the available bandwidth estimate B for the connection. The adaptive decrease algorithm is defined as follows.

On Packet Loss Indication:

$$\text{ssthresh} = \frac{B \times RTT_{min}}{\text{seg size}}$$

If ssthresh ≤ 2 then ssthresh $= 2$ (9)

If the packet loss is detected by 3 duplicate packets

$$\text{CWND}_{new} = \text{ssthresh} \tag{10}$$

If congestion is detected with a timeout of an unacknowledged packet

$$\text{CWND}_{new} = 1 \tag{11}$$

TCPW setting of the ssthresh value results in the TCPW connection utilising more bandwidth than standard TCP due to two factors. 1) When the loss is signalled by

3 duplicate ACKs TCPW's CWND does not decrease as much as standard TCP decrease. 2) When the loss is signalled by timeout of an unacknowledged packet both TCPW and standard TCP decrease the CWND to just one packet, however since TCPW sets the ssthresh to a much larger value than what standard TCP does, the TCPWconnection has a longer slowstart phase (in which its CWND grows exponentially) than the standard TCP connection. TCPW increase function is non-adaptive while its decrease function is adaptive. TCPW however presents the challenge of accurate estimation of the available bandwidth. The bandwidth estimation algorithm used for TCPW+ [11] is more accurate than that used for an earlier version of TCPW [5, 10].

TCP for High-speed and long-distance networks (HTCP)[14]

Uses an adaptive congestion control algorithm in which i the increase factor of source i is set for each acknowledgment as a function of the time elapsed since the last packet loss event. On receipt of each acknowledgement the CWND increment is defined by the same equation as standard TCP (1), but differs in the assignment of i, which follows equation (12).

$$\alpha_i \leftarrow \begin{cases} 1 & \Delta_i \le \Delta^L \\ 1 + 10(\Delta_i - \Delta^L) + \frac{(\Delta_i - \Delta^L)^2}{2} & \Delta_i > \Delta^L \end{cases} \tag{12}$$

Where Δ_i is the elapsed time since the last congestion event experienced by the flow i, Δ^L is the time duration used as the threshold for switching from the low to high speed regimes. In the first case, i is increased in the same manner as the standard TCP algorithm. In the second case however i is increased by a function relative to the elapsed time since the last congestion event, whose response function is similar to that of HS-TCP.

HTCP's decrease factor β_i for a particular source *i* is computed relative to that flow's throughput just before and after the last congestion event. The decremented CWND is thus set according to equation (13).

$$\text{CWND} \leftarrow \beta \times \text{CWND} \tag{13}$$

Where

$$\beta_i(k+1) \leftarrow \begin{cases} 0.5 & \frac{B_i^-(k+1) - B_i^-(k)}{B_i^-(k)} > 0.2 \\ \frac{RTT_{min,i}}{RTT_{max,i}} & \text{otherwise} \end{cases} \tag{14}$$

Where $B_i^-(k)$ is the throughput of flow i immediately before the k'th congestion event, $B_i^+(k)$ is the throughput of flow i immediately after the k'th congestion event, RT Tmin,i is the minimum RTT experienced by the i'th source and RT $T_{max,i}$ isthe maximum RTT experienced by the i'th source.

The current implementation of HTCP uses equation (15) to estimate

$$\frac{RTT_{min,i}}{RTT_{max,i}}$$

$$\beta_i\, i(k+1) = \min \beta_i\,(j)\; B_i^{-}(j) \quad B_i^{+}(j) \tag{15}$$

Cubic TCP (CUBIC-TCP)[13]

CUBIC-TCP is an algorithm that aims at improving and simplifying the BI-TCP algorithm, hence its resulting response function is similar to that BITCP. It is however a fully adaptive algorithm unlike BI-TCP since CWND modifications are computed relative to the time elapsed since the last packet loss event in a similar manner to HTCP.

The CWND of CUBIC is determined by equation (16).

$$\text{CWND} = C(t - K)^3 + \max\text{-win} \tag{16}$$

Where C is a scaling factor, t is the elapsed time from the last window reduction, max win is the window size just before the last reduction, and $K = \sqrt[3]{\max \text{win}\ \beta / C}$ where is a constant multiplication decrease factor applied for window reduction at the time of the loss event.

Our Proposed Algorithm (HTCP with Bandwidth Estimation (HTCP-BE))

Basing on the observations from the above stated algorithms [4, 8, 18, 11, 14, 13] as well as some comparative studies [3, 9, 6, 16], we propose an algorithm that combines the strengths of HTCP and TCPW. With this algorithm a connection will be established following HTCP-like rules, however at a packet loss, we propose that the TCPW's adaptive decrease be evoked. From our simulations with NS2, on a packet loss event a TCPWflow sets its CWND to a value averagely 17% greater than that set by an HTCP flow. The increase factor will be controlled by the HTCP adaptive increase, which controls increment relative to the elapsed time since the last packet loss event. From our simulations the HTCP increase function yields much faster growth of the CWND than that obtained with TCPW. The aim is to use the tested band width estimation mechanism employed by TCPW algorithm and combine it with adaptive increase mechanism of HTCP. The expectation is that this will ensure higher link utilisation than what is possible with any of TCPWor HTCP. The weaknesses of HTCP in short RTT flows as well as parallel flows revealed in [6] we propose to solve by employing the tested TCPW adaptive backoff. The non-adaptive nature of TCPW increase factor will be taken care of by the HTCP adaptive increase factor. Our proposed algorithm is summarised in (17) and (19).

On Acknowledgement:

$$\text{CWND} \leftarrow \text{CWND} + \alpha/\text{CWND} \tag{17}$$

Where

$$\alpha_i \leftarrow \begin{cases} 1 & \Delta_i \le \Delta^L \\ 1 + 10(\Delta_i - \Delta^L) + \left(\frac{\Delta_i - \Delta^L}{2}\right)^2 & \Delta_i > \Delta^L \end{cases} \tag{18}$$

Where Δ_i is the elapsed time since the last congestion event experienced by the flow i, Δ^L is the time duration used as the threshold for switching from the low to high speed regimes.

On Packet Loss event:

$$CWND \leftarrow B \times RTT_{avg} \tag{19}$$

Where

$$B = CWND/RTT_{avg} \tag{20}$$

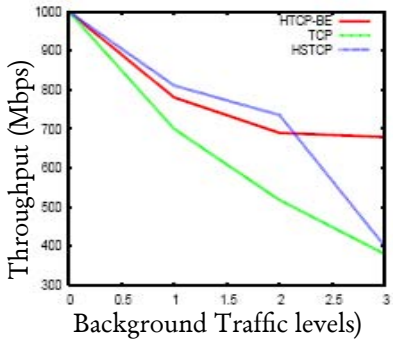
Setup

Using NS-2 simulator [2, 12], we run simulations for 300 seconds, three times for each protocol and computed the averages. We simulate (scripts available on request), typical e-VLBI setting within the EVN with round trip times varying between 10ms and 40ms and bottleneck links of 1 Gbps.

Performance Evaluation

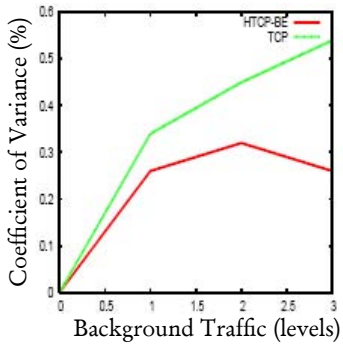
Throughput Efficiency

For each of the protocols we plot here the average throughput achieved across varying backround traffic environments.



Stability

Similarly for each of the protocols we plot here the average coefficient of variation across varying backround traffic environments. The lower this value the more stable the protocol.



Conclusion and future work

Our results show that HTCP-BE is of high stability across varying background traffic environments and maintains a high throughput, thus suitability for the delay-sensitive e-VLBI application. To further improve data transport for e-VLBI we shall embark on the following;

- Vary Background traffic to include non-TCP flows b'se much as the most internet traffic is TCP, a significant portion is non TCP and it may affect the performance differently.
- Implement HTCP-be and run e-VLBI tests on a real network.
- Increase the throughput effeciency of HTCPBE by maintaining a relatively large enough CWND (not reducing at all) in periods of low packet loss and only decreasing it when the packet loss gets out of hand. This is especially useful because e-VLBI is not loss sensitive but delay sensitive.
- Exploring other transport protocols that are not TCP related such as User DatagramProtocol (UDP) and Real-Time Protocol(RTP)

References

M. Allman, Paxson V., and W. Stevens. Tcp congestion control. RFC 2581, Internet Engineering Task Force, 1999.

L. Breslau, D. Estrin, S. Fall, J. Heidemann, P. Helmy, H. S. McCanne, K. Varadhan,Y. Xu, and H. Yu. Advances in network simulation. IEEE Computer, 55(5):59–67, 2000.

H. Bullot, L.S. Cottrell, and R. Hughes-Jones. Evaluation of advanced tcp stacks on fast long-distance production networks. PFLDnet, 2004.

S. Floyd. Highspeed tcp for large congestion windows. RFC 3649, Internet Engineering Task Force, 2003.

M. Gerla, M. Sandaidi, M. Valla, and R. Wang. Tcp westwood with adaptive bandwidth estimation to improve efficiency/friendliness tradeoffs. Computer Communication Journal, 2003.

S. Ha, Y. Kim, L. Le, I. Rhee, and L. Xu. A step toward realistic evaluation of high-speed tcp protocols. Technical report, Department of Computer Science, North Carolina State University, 2006.

Mark5 vlbi data system. Haystack observatory Website. http://web.haystack.mit.edu/mark5/.

T. Kelly. Scalable tcp: Improving performance in highspeed wide area networks. ACM SIGCOMM Computer Communications Review, 33(2), 2003.

Y. Li, D. Leith, and R. Shorten. Experimental evaluation of tcp protocols for high-speed networks, 2005.

S. Mascolo, M.Y. Sanadidi C. Casetti, M. Gerla, and R. Wang. Bandwidth estimation for enhanced transport over wireless links. In Mobile Computing and Networking, pages 287 – 297, 2001. 5

S. Mascolo, L. A. Grieco, R. Ferorelli, P. Carmada, and G. Piscitelli. Performace evaluation of westwood+ tcp congestion control. Performance Evaluation, 55:93 – 111, 2004.

The ns2 simulation. www.isi.edu/nsnam/ns.

I. Rhee and L. Xu. Cubic: A new tcp-friendly high-speed tcp variant. Proc. PFLDnet, 2005.

R.N.Shorten and D.J.Leith. H-tcp: Tcp for high-speed and long-distance networks. Proc. PFLDnet, Argonne, 2004.

J. Sansa, A. Szomoru, and J.M. van der Hulst. On network measurement and monitoring of end-to-end paths used for e-vlbi. Chapter in Advances in Systems Modeling and ICT Applications ISBN: 13: 978-9970-02-604-3, 2006.

M.C. Weigle, P.Sharma, and Jesse R. Freeman. Performance of competing high-speed tcp flows. Proceedings of IFIP Networking, 2006.

Gary R. Wright and W. Richard Stevens. Tcp/Ip Illustrated: The Protocols, volume 1. Addison-Wesley, 1994.

L. Xu, K. Harfoush, and I. Rhee. Binary increase congestion control for fast long distance networks. INFOCOM, 2004.

Part 5

Software Engineering

32
A Comparison of Service Oriented Architecture with other Advances in Software Architectures

Benjamin Kanagwa and Ezra K. Mugisa

Service Oriented Architecture (SOA) allows software systems to possess desirable architecture attributes such as loose coupling, platform independence, inter-operability, reusability, agility and so on. Despite its wide adoption in the form of Web services, many stakeholders both from academia and industry have limited understanding of its underlying principles. This has led to faulty implementations of SOA and in some cases, it has been implemented in places where it is not suitable. In this paper, we investigate and show the relationship between SOA and other advances in software architecture. The paper relates SOA to Architecture Patterns, Components and Connectors. We advance the view that SOA's uniqueness and strength does not lie in its computational elements but in the way it enables and handles their interaction. In this way, we facilitate the understanding of SOA in terms of other advances in software architectures that are already well understood. We believe that a good understanding of SOAs in terms of other advances in software architectures will help to reap its enormous benefits.

Introduction

SOA has gained more popularity of recent, largely due to web services (Hashimi, 2004) and partially due to natural evolution in software engineering (Tsai, 2005). At the same time, there is a systematic progression from object oriented systems, distributed objects, components and then service oriented systems. This evolution in software engineering shows a systematic move from tightly coupled and platform specific systems to more robust loosely coupled platform-independent systems.

Despite this progression, and widespread interest in SOAs, most publications on the subject fail to explain SOA or simply assume it merely defines a synonym for a stack of extensible markup language (XML) web service protocols and standards (Stal, 2006). Stal argues that developers need to understand the service oriented paradigm from the architecture perspective in order to leverage SOA implementations. This lack of understanding has already manifested itself in faulty implementation of SOA as highlighted by Loughran and Smith (2005). We argue that these faulty implementations

that carry remote procedure call sentiments can only be addressed by understanding the core principles of SOA. Since we cannot divorce services from

SOA, we start by addressing the problem from the core. The focus is to provide an understanding of SOA in terms of well established concepts in software architectures.

It has been reported in literature (OASIS, 2006) that SOA provides a way of building architectures and builds on existing advances in software architecture (Szyperski, 2003), (Gokhale et al., 2002). We view SOA as an alternative approach to building architecture of systems. The architecture solution space includes well formulated approaches such as: Component Based Architectures (CBA), Connectors, Architecture Patterns/Styles and Architecture Description Languages (ADLs). Given that these concepts are well developed and understood, it is important to show how thye relate to SOA. We are particularly interested in those advances that are in line with providing solutions to the software architecture problems. As part of our investigation we show the specific features that have been enhanced by SOA to make it distinct from its predecessors.

Our definition for SOA and service is from our previous work (Kanagwa and Mugisa, 2007) where we define a system (application) that is considered to have a SOA. From this definition, we can establish the principles built within SOA systems. A service is defined as a set of functionality that exposes a resource for use by others and is invoked directly under the conditions it specifies. We say that an application is a SOA system if it is partitioned into sets of independent functionality that interact directly, independent of context.

The concepts in the above two definitions are based on the definitions provided by (Baker and Dobson, 2005) and (OASIS, 2006). In (Baker and Dobson, 2005), a service is defined as "a set of functionality provided by one entity for the use of others". We feel that it is important to emphasize direct interaction to avoid confusion with technologies such as Common Gateway Interface (CGI) (accessed through web servers), and Components, that need component frameworks through which they interact.

SOA Context

We take a brief look at the use of the term SOA. It has been described as a framework, as a paradigm (OASIS, 2006) as a collection of services, plus several other definitions. What is clear from these definitions is that they do not represent the same thing. At this point, we would like to distinguish the two uses of the term SOA. In some contexts, SOA is used to refer to an abstract architecture (architectures that do not represent concrete systems). However, SOA has another meaning to it in which it is used to refer to concrete systems (running applications). The later is common from the industry practitioners.

In the first context, SOA does not represent architecture of a specific system just like Common Request Broker Architecture (CORBA). Definitions such the framework, paradigm refer to this dimension of SOA. Definitions such as collection of services are inline with the second dimension of SOA. In the second context, the definition aims at answering the question, when is a system service

oriented? The ideal position should be that systems designed as per the first SOA context imply the second context. In other words, if SOA is a framework or paradigm, systems designed as per this framework or paradigm should lead to systems whose architecture is service oriented. Unfortunately this is not the case.

As a framework, paradigm or way of building architectures, SOA involves what we have described under the section "architecture patterns for SOA". In addition SOA specifies how to specify and describe services, how to facilitate the actual interaction between services. Although how to 'publish' and 'find' services are key principles in the SOA framework, they are not part of architectures of resulting systems.

Just like component repositories do not contribute to the software architecture, the way the services are published and discovered is not really part of running SOA.

Contribution

Our major contribution is tracing the evolution of SOA and expressing SOA in terms of other advances in software architectures that are already well established and understood. In this way, we facilitate understanding of SOA and foster its proper usage. Proper understanding and usage of SOA will leverage its benefits.

The rest of the paper is organised as follows. We discuss related work, invesigate the evolution of SOA, show how SOA relates to other advances in software architectures, show features inherited by SOA from its predecessors and provide a conclusion.

Related Work

A comparison of Service-Oriented and Distributed Object Architectures is given in (Baker and Dobson, 2005). They state that :"while superficially similar, the two approaches exhibit a number of subtle differences that, taken together, lead to significant differences in terms of their large-scale software engineering properties such as the granularity of service, ease of composition and differentiation properties that have a significant impact on the design and evolution of enterprise-scale systems".

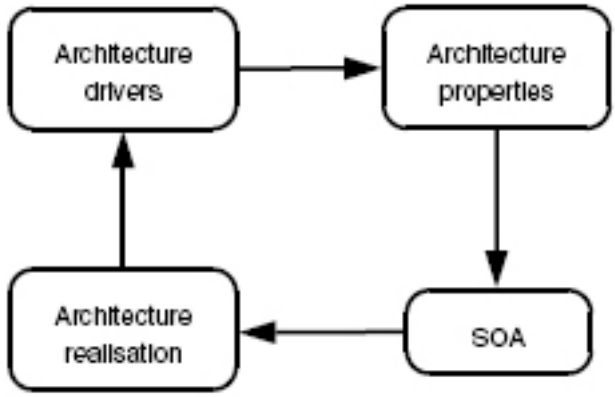
Stal (2006) suggests the use of architecture patterns and blueprints to explain the architecture principles of SOA. The work is based on the driving forces to define a model for the service oriented context. The work addresses the architecture patterns necessary to address the driving forces. However, the work does not do much to facilitate the understanding of SOA. The work focusses on implementation and design patterns. Our effort is aimed at explaining SOAs by stressing the relation of SOA with existing architecture concepts (patterns inclusive).

Evolution of SOA

We have noted earlier that SOA does not represent a concrete system but provides a way of building architectures. So it is part of the software architecture solution

space. We describe the emergence of SOA using figure 1. The aim is to put SOA in context based on its drivers, architecture properties, architecture realisation and instances.

Fig1: Showing relation between SOA concerns (drivers), architecture features and architecture



In figure 1, we argue that architecture evolution is precipitated by limitations of the architecture approaches in addressing existing and emerging situations. We call the limitations architecture drivers. In other words, these limitations create the need for a new architecture approach. The architecture drivers point to specific architecture properties. The architecture properties translate into an architecture framework or style. Finally the realization of the software architecture is obtained. The architecture realization must satisfy the drivers and this completes the relationship.

Architecture Drivers for SOA

SOA is a distributed computing architecture (OASIS, 2006) and so are the predecessors. Remote Procedure Call (RPC), CORBA, Distributed Common Object Model (DCOM), and Remote Method Invocation (RMI) are (were) the most widely used and well known distributed computing architectures. They were developed at a time the Internet was not widely spread or discovered to be a potential distributed computing platform. They therefore failed to rise to the challenges of the Internet as a global computing platform. To fully reap the benefits of the internet, a new way of building software architectures was needed. SOA carries elements of existing architecture approaches but emphasizes the interaction of its architecture elements.

Baker and Dobson (2005) argue that SOA differs from Distributed Object Architectures in a number of aspects that lead to combined significant differences in terms of large scale engineering properties. In (Alencar et al., 2002), we learn that current software applications can be separated into a basic concern and a number

of special purpose concerns. The basic concern is represented by the fundamental computational algorithms that provide the essential functionality relevant to an application domain and the special purpose concerns relate to other software issues, such as user interface presentation, control, timing, and distribution. SOA is driven by the quest to address the special purpose concerns that relate to large scale systems. These concerns are mainly interaction issues. SOAs strength does not lie in its computational elements but in the way it enables and handles their interaction. Interaction manages the behaviour of elements and therefore has substantial influence on the overall system. By focusing on interaction, SOA is able to mitigate the limitations of its predecessors.

Soa And Software Architectures

Most work in SOA emphasises enabling technologies such as in web services. We assess the contribution of SOA to the software architecture solution space as an alternative solution or enhancement of other advances in software engineering.

Architecture Patterns and SOA

SOA is modeled as a relationship of three kinds of participants (roles): the service consumer (client), the service registry, and the *service provider*. This view of three participants is shared by (Papazoglou, 2003), (Hashimi, 2004), (Pilioura and Tsalgatidou, 2001), (Talaei-khoeil et al., 2005), (Huhns and Singh, 2005) and (Gottschalk et al., 2002) among others.

At the center of SOA are the above three roles. It starts with the service provider publishing its existence in some registry, followed by the service consumer finding the service provider by querying the service registry. The service consumer then binds directly to the service provider. This translates to the publish–find–bind architecture pattern. The realization of this pattern is left to different instances of SOA. In (Hull et al., 2003, Wu and Chang, 2005), they relate the SOA pattern with other advances in software architecture especially broker, peer-2-peer and client-server architecture patterns. Web services is the most developed instance of SOA, and we draw valuable lessons from its implementation. Table shows a summary of the interaction styles from web services that are closely related to SOA. The styles in the table are not definitive to SOA because other possibilities are possible.

It is desirable that services be published to a registry. The publish interaction employs the client-server interaction style that is characterized by request–response. However, SOA imposes no restriction.

Table 1: SOA interaction Styles:-CS-client server; CSS-client stateless server; p2p-peer-to-peer

	CS	CSS	p2p
publish	x		
find	x	x	
bind			x

Any form of publishing can be used. For example, if the target consumers are known, the service descriptions can be directly sent to the consumers. This is very likely in highly specialized services such as in the military where the consumers are restricted.

Find relates to discovering binding information for a service. Typically, it is the connection between the consumer and registry. The pattern of finding a service can be carried out using client-server interaction style. The request is inform of a 'service specification' and the reply is the 'service description'. In this case, the consumer is the client and the registry is the server. Other possibilities exist. Direct interaction between the consumer and provider (example of small in-house developers) or in specialized services discussed above. It is also possible for a consumer to make a provider-by-provider search until it finds the right service. Although many possibilities for finding services are possible (especially if fewer services are involved), automating service interactions requires client-server interaction.

Bind, involves the actual interaction between the service provider and consumer or simply services. A service has multiple operations and each interaction is independent of others. In respect to a single interaction, one service acts like a client and the other like a server leading to client-server interaction style. However, the reply is optional and the services are not really clients and servers as defined in client-server architectures .The services simply assume the roles when required for a specific interaction.

Whereas publish and find are important aspects of SOA, they are not part of the resulting architecture. The resulting architecture is the collection of services and how they interact with no regard on the 'publish' and 'find' styles. The publish-find styles enable quick and flexible binding, but do not manage the service binding or provide rational for choice of the services. They are therefore not part of the architecture. The architecture of a service oriented is defined once the service have been found and bound together. Architecture reconfiguration is possible through dynamic discovery and binding.

Thus, the focus of SOA systems should be on enabling quick and flexible binding between services. SOA provides the underlying framework for interaction and dynamic behaviour of services. We believe that this service interaction is reusable across all service oriented systems and is what characterizes SOA.

Architecture Connectors and SOA

Interactions in software architectures are modeled by connectors (Shaw and Garlan, 1996). It has been argued that even components that do not use connectors explicitly have composition operators that can be identified as connectors at different levels of abstraction (Lau et al., 2005). We note that this is true for all systems that are partitioned into subsystems. Such subsystems must communicate to be able to work as a unit. Connectors therefore exist in all systems but in different forms and levels of abstraction. The abstraction levels range from low level programming to more explicit connectors at the architecture.

Even though connectors are not explicitly mentioned in relation to SOA's, it is easier to appreciate the strength of interactions in software architecture by drawing comparisons with connectors. SOA's main uniqueness is derived from its interaction mechanisms that control the overall system rather than structure elements. In relation to connectors, the significance of interactions is given by (Medvidovic et al., 2000), (B´alek and Pl´aˆsil, 2001) and (Guo et al., 2005) among others.

Whereas the main system functional blocks are components, the properties of the system also strongly depend on the character of the component interactions (B´alek and Pl´aˆsil, 2001). They identify benefits of connectors as: increased

re-usability, direct support for distribution, location transparency, mobility of components in a system, support for dynamic changes in the system's connectivity and so on. Medvidovic et al. (2000), point out that in large, and especially

distributed systems, connectors become key determinants of system properties, such as performance, resource utilization, global rates of flow and security.

Connectors help to localize information about interaction of components in a system such that information is no longer spread over all communicating components (Guo et al., 2005). Therefore, connectors have significant impact on the quality attributes and offer a basis for abstracting interactions and reasoning about systems at the architecture level. We note that typical service based systems operate by interacting among constituent services.

SOA focuses on the special purpose concerns that relate to inter-operability and loose coupling. It borrows some of the concerns from its predecessors. However, SOA refines the concerns to provide significant impact on the systems that use it.

Connectors can exist in explicit or implicit form. Although SOA does not define explicit connectors, it is clear that its implicit connectors shape the characteristics of SOA. SOA does not define connectors explicitly but defines the infrastructure for constructing the connectors. The infrastructure generally includes the interface specification, descriptions, message formatting and messaging protocols.

Predecessors of SOA

We would like to think that SOA builds on existing knowledge, particularly architecture approaches. The question to answer here is which old system does SOA compare with or extend? Or what does a service extend? To put the question

in another way, which architecture features does it build on? This question has been answered by (Gokhale et al., 2002) and (Szyperski, 2003). Gokhale et al. (2002) makes a comparison between Web services and CORBA. Szyperski (2003); suggests that some elements of component technology apply to services.

What (Szyperski, 2003) and (Gokhale et al., 2002) do not answer however, is what exactly do we carry forward to SOA? (Wang and Fung, 2004) provide an interesting comparison of Object Oriented Architecture (OOA), Component Based Architecture (CBA) and Service Based Architecture (SBA). It is clear from the characteristics -that SOA emphasizes interaction rather than structure. For example, a one-to-one comparison structural comparison of Web services and CORBA (Gokhale et al., 2002) does not yield much. It only reveals philosophical differences that translate to efficient interaction and only contribute to the overall architecture of the resulting systems.

Features from Predecessors

From its main predecessors:-Distributed Component Model (DCM), a subset of CBSE, SOA carries the following features: composability,separationofconcerns ,loosecouplingandabstraction. SOAcombinesbenefitsandchallenges from COTS and CBD except that services are consumed from where ever they are. We note that SOA relaxes the restrictions on the features to make it unique from its predecessors.

Composability

From CBSE, we carry forward the belief that complex systems can be built from smaller elements. Unlike CBSE, SOA is not constrained by a specific component model. Component models are in fact the major limitation to interoperability

with existing component models such as DCOM, CORBA, EJB not able to easily interoperate. The importance of component models in DCM and CBSE is highlighted by (Lau and Wang, 2005) and (Baker and Dobson, 2005). Comparing SOA and component models, SOA is a relaxation of the restrictions on component based systems. There is no requirement for a specific internal structure, no component model semantics, no containers and application servers. This relaxation fosters software agility that requires software to change frequently and gracefully.

Separation of concerns

This principle states that a given problem involves different kinds of concerns, which should be identified and separated to cope with complexity, and to achieve the required engineering quality factors such as robustness, adaptability and reusabilit y(Aksitetal.,2001). InSOAthemajorconcernsareintegrationofautonomoussystems. Integrationbecomes complex if the systems in question do not match in terms of protocols used and data formats. Thus, the critical starting point addressed by SOA is the way systems are exposed and consumed. SOA requires that systems are accessed through platform independent interfaces.

However, a key concept in Object Oriented Programming (OOP) that is not necessary in SOA is inheritance mainly because SOA is not a programming model and services are self contained, independent entities whose internal details are invisible. Services do not need to inherit any functionality since they can include the functionality through standardised service interactions.

Loose coupling

Loose coupling is the most pronounced architecture property associated with SOA. Loose coupling is not really a new concept that is unique to SOA. It exists in CBSE, except SOA is designed to be more loosely coupled. SOA achieves more loosely coupling through its emphasis on platform neutral interfaces, platform neutral descriptions, message based interactions and self-contained services. In tightly coupled systems, architecture elements are designed for each other, while in loosely coupled systems, architecture systems are designed to interoperate.

Abstraction

SOA extends the concept of abstraction and encapsulation. Unlike in OO which hides internal details (data and methods), SOA aims at hiding all causes of integration problems. Stal (2002) lists the causes of heterogeneity as "network technologies, devices, and OSs; middleware solutions and communication paradigms; programming languages; services and interface technologies; domains and architectures; and data and document formats."

SOA encapsulates a specific set of discrepancies in its domain of application. By hiding these differences, services within the service oriented architecture can be accessed seamlessly. The differences are typically hidden by using a common set of standards. The differences may be in terms of data structures, communication protocols or platforms. For example, SOA's that target the Internet as the mode of delivery, will require standards that hide heterogeneity in platforms, while a simple SOA that is part of an operating system such as accessing devices, will require standards that present the devices in the same way. A typical example in UNIX where everything is a file. In terms of architecture, it implies that standards play a critical role. Through use of standards, protocols and common vocabulary, the differences in participants are resolved thereby allowing all the participants to appear uniform.

Generally, SOA does not eliminate the heterogeneity, but simply hide the heterogeneity causes. The systems and technologies remain heterogeneous (Stal, 2002) but their interfaces and collaborations are standardized. This is a core concept of SOA and the choice of interface specification and standards is left to different instances of SOA. The choice

and design of the standards must be carried out with care lest they will be a bottleneck. They must be simple to use and comprehend without compromising the efficiency. Simple standards tend to be very verbose while short standards are normally hand to understand. A key to this is a compromise between the two

extremes. According to Almeida et al. (2003),the protocols and standards that enable technology abstraction should be suitable and intuitive for application developers that develop and maintain applications in different technology domains.

Uniform access as a requirement for as a strength for SOA is noted by (Curbera et al., 2001), (Baker and Dobson, 2005). Curbera et al. (2001) "a key goal of WS framework is to produce a common representation of applications which use diverse communication protocols and interaction models while at the same time enabling Web services to take advantage of more efficient protocols when they are available at both ends of the interaction".

Conclusion

We have showed the relation between SOA and other advances in software architectures. In so doing, we have clarified most architecture issues surrounding SOA. We have also showed some of the exact features carried by SOA's from its predecessors Through comparison of SOA with other advances in software architecture, we have advanced the view that SOA's uniqueness and strength does not lie in its computational elements but lies in the way it enables and handles their interaction. Whereas SOA relaxes several features from its predecessors, we have showed how such relaxations impact on the architecture of the resulting system. We have backed this claim by looking at the contribution of interactions (connectors) to software architectures.

Therefore, we recommend that future research in SOA should focus on optimizing that view.

References

Aksit, M., Tekinerdogan, B., and Bergmans, L. (2001). The six concerns for separation of concerns. In ECOOP-Workshop on Advanced Separation of concerns, Budapest.

Alencar, P. S. C., Cowan, D. D., and de Lucena, C. J. P. (2002). A logical theory of interfaces and objects. IEEE Trans. Software Eng., 28(6):548–575.

Almeida, J. P. A., Pires, L. F., and van Sinderen, M. J. (2003). Web services and seamless interoperability. In ECOOP 2003 European Workshop on Object Orientation and Web Services , Darmstadt, Germany.

Baker, S. and Dobson, S. (2005). Comparing service-oriented and distributed object architectures. In OTM Conferences (1), pages 631–645.

B´alek, D. and Pl´aˆsil, F. (2001). Software connectors and their role in component deployment. In Third International Working Conference on Distributed Applications and Interoperable Systems (DAIS) .

Curbera, F., W, N., and Weerawarana, S. (2001). Web services. why and how. Workshop on object-Oriented Web Services OOPSLA, Tampa, Florida,USA, 12:591–613.

Gokhale, A., Kumar, B., and Sahuguet, A. (2002). Reinventing the wheel? corba vs. web services. In Proceedings of The Eleventh International World Wide Web conference (WWW2002) , Honolulu, Hawaii.

Gottschalk, Graham, K., Kreger, S., and H. Snell, J. (2002). Introduction to web services architecture. IBM systems Journal., 2(41):170–177.

Guo, J., Liao, Y., Gray, J., and Bryant, B. (2005). Using connectors to integrate software components. 12th IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'05) , 00:11–18.

Hashimi, S. (2004). Service-oriented architecture explained. http://www.ondotnet.com/pub/a/dotnet /2003/08/18/soa explained.html. Huhns, M. N. and Singh, M. P. (2005). Service-oriented computing: Key concepts and principles. IEEE Internet Computing, 9(1):75–81.

Hull, R., Christophides, V., Benedikt, M., and Su, J. (2003). Eservices: A look behind the curtain. In Proceedings of the International Symposium on Principles of Database Systems (PODS) , San Diego CA, USA.

Kanagwa, B. and Mugisa, E. K. (2007). Architecture analysis of service oriented architecture. In SERP'07, Orlando, FL, USA(to appear).

Lau, K.-K., Elizondo, P. V., and Wang, Z. (2005). Exogenous connectors for software components. CBSE 2005, (90–106).

Lau, K.-K. and Wang, Z. (2005). A taxonomy of software component models. In EUROMICRO-SEAA, pages 88–95.

Loughran, S. and Smith, E. (2005). Rethinking the java soap stack. In IEEE International Conference on Web Services (ICWS), Orlando, Frolida, USA.

Medvidovic, N., Gamble, R. F., and Rosenblum, D. S. (2000). Towards software multioperability: Bridging heterogeneous software interoperability platforms. In Proceedings of the Fourth International Software Architecture Workshop (ISAW-4), pages 77–83, Limerick, Ireland.

OASIS (2006). Reference model for service oriented architecture draft 1.0.

Papazoglou, M. P. (2003). Service -oriented computing: Concepts, characteristics and directions. In Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE'a03) . IEEE.

Pilioura, T. and Tsalgatidou, A. (2001). E-services: Current technology and open issues. In TES, pages 1–15, Rome, Italy.

Shaw, M. and Garlan, D. (1996). Software Architecture: Perspectives on an Emerging Discipline. Prentice Hall.

Stal, M. (2002). Web services: Beyond component-based computing. Communications of the ACM, 45(10):71–76.

Stal, M. (2006). Using architectural patterns and blueprints for service-oriented architecture. IEEE Software, 23(2):54– 61.

Szyperski, C. A. (2003). Component technology -what, where, and how?. In ICSE, pages 684–693.

Talaei-khoeil, A., Sheriffan, A. H., Akbari, M. K., and Farzaneh, J. (2005). a new approach for service orieneted architecture. In 3rd International Conference on Information and Communications Technology, 2005. Enabling Technologies for the New Knowledge Society.

Tsai, W. T. (2005). Service-oriented system engineering: a new paradigm. In SOSE 2005, pages 3– 6.

Wang, G. and Fung, C. K. (2004). Architecture paradigms and their influences and impacts on component-based software systems. In HICSS.

Wu, C. and Chang, E. (2005). Comparison of web service architectures based on architecture quality properties. In INDIN 05, pages 746–755.

33
Decision Support for the Selection of COTS

Tom Wanyama* Agnes F. N. Lumala**

Commercial-Off-The-Shelf (COTS)-Based Software Development (CBSD) has the potential to reduce time and resources required to develop software. However, In order to realize the benefits, which COTS bring to software development, it is imperative that the "right" COTS products are selected for projects, because selecting inappropriate COTS products may result in increasing time and cost of software development; which CBSD aims at reducing. COTS selection is a major challenge to COTS-based software developers, due to the multiplicity of similar COTS products on the market with varying capabilities and quality differences. Moreover, COTS selection is a complex decision-making problem that is characterized by the following: uncertainty, multiple stakeholders, and multiple selection objectives. Therefore, the process of selecting COTS products requires addressing multiple challenges, which in turn necessitates using multiple Decision Support Applications (DSA) as well as multiple information repositories. This paper describes an agent-based Decision Support System (DSS), which integrates various COTS selection DSA and repositories to address a variety of COTS selection challenges. Besides managing the COTS selection DSA and repositories, the agents are used to support communication and information sharing among the COTS selection stakeholders.

Introduction

As software development has increasingly become focused on component-based software engineering, the emergence of plug-and-play software components, in the form of Commercial Off-The-Shelf (COTS) components, has been a valuable result. COTS hold the promise of providing versatile, low cost, and efficient solutions for a variety of recurring functional requirements applicable to a wide array of commercial software products. One difficulty in the way of widespread implementation of COTS solutions is the lack of, efficient, effective, adaptive systems, which facilitate the accurate selection of a single, or a group of COTS to fulfill a specific, and often-unique system requirements. As software development continues on in the trend to develop component-based software, the popularity of COTS and the number of COTS also continues to grow. With the growth of this trend, it is becoming clear that there is a need for Decision Support Systems (DSSs) to assist developers of COTS-Based Software (CBS) to select appropriate COTS for their software.

In response to the growing need for COTS Selection DSS (CSDSS), a multiplicity of COTS selection methods have been reported in literature, the

prominent among them being the following: Off-The Shelf Option (Kontio et al., 2000), Comparative Evaluation Process (Cavanaugh et al., 2002), COTS-based Requirements Engineering (Alves et al., 2003) and (Alves, 2003), COTS-Aware Requirements Engineering (Chung et al., 2002), Procurement- Oriented Requirements Engineering (Ncube et al., 2003), COTS Acquisition Process (Ochs et al., 2000), QUESTA (Hansen, 2003), Storyboard (Comella-Dorda et al., 2002), and, Socio-Technical Approach to COTS Evaluation (Kunda et al., 1999), PECA (Comella-Dorda et al., 2002), and Combined Selection of COTS Components (Burgues et al., 2002). An evaluation of these COTS selection methods based on their ability to address the COTS selection challenges presented in Wanyama and Far revealed that they generally have the following major shortfalls:

- Absence of provisions to allow for access to expert knowledge on COTS products and the COTS selection process
- The methods lack support for interaction among stakeholders to share information and experiences
- Knowledge collection and reuse is not clearly handled. While most of the methods recommend documentation of the selection process, they do not have representation models for the existing body of knowledge about COTS evaluation and selection. Moreover, none of them has a model for learning from previous COTS selection processes
- Absence of provisions to support interaction between the users and the decision support tools, so that users can try out different scenarios and weigh their impact on the COTS selection process
- Lack of techniques to appropriately include all aspects of COTS evaluation and selection in the process models. For example, OTSO considers only the functional and cost aspects of the process, leaving out the quality and vendor aspects
- Finally, none of the reviewed methods considers uncertainty caused by evolution of COTS products, and the ability of the selection team

Researchers such as Holmes (Holmes, 2000), and Ncube and Dean (Ncube et al., 2002) have suggested that the problem with the reviewed COTS selection methods is the accuracy of their COTS performance estimates. Therefore, they have proposed new and/or improved COTS evaluation techniques, which provide better accuracy. For example, Ncube and Dean (Ncube et al., 2002) have proposed a COTS evaluation method called gap analysis, with the aim of mitigating the shortfalls of the conventional Multi-Criteria Decision Making (MCDM) techniques. However, a review of gap analysis revealed that the method is just another MCDM technique with its own set of strength, weaknesses, limitations, and assumptions. While we commend work of those developing new and 'better' MCDM techniques for COTS selection, we believe that provided the above shortfalls of the reviewed COTS selection methods are addressed, the

existing MCDM techniques can adequately handle the COTS selection problem. Moreover, improving accuracy of estimates serves no useful purpose, if the input data to the improved models is either unavailable or inaccurate.

To overcome the above-mentioned obstacles to selecting appropriate COTS for software projects, we propose a framework for COTS selection. The framework has a process model that guides the stakeholders throughout the COTS selection process, and a Decision Support System (DSS), which assists the stakeholders to make the necessary decisions at every stage of the COTS selection process. The DSS incorporates group negotiation and conflict resolution in order to exist as a valid and accurate selection support tool. As such, a multi-agent system (MAS) based design was determined to be appropriate, as agents can be used to provide an autonomic, flexible, and modular interface between the end user (stakeholder) and the DSS.

The rest of this paper is arranged as follows. In section 2, the requirements for the framework for COTS selection are presented. Section 3 explains the framework for COTS selection that we developed, presenting how each of the requirements presented in Section 2 were addressed. In addition, the section presents a DSS that integrates the various applications that were developed to provide decision support for the various steps of the COTS selection Process. Finally, section 4 presents the main conclusions

Requirements for the Framework for COTS Selection

The general shortfalls of the COTS selection methods in the reviewed literature reveal that none of the COTS selection methods was designed to be a framework for providing decision support for the process of selecting COTS products. Each of them addresses just a few of the COTS selection challenges; and the focus of most of them is to apply analysis and decision techniques in the COTS selection process, other than providing a framework for supporting COTS selection. Therefore, we decided to develop a framework for the process of selecting COTS products from scratch. To do this, we started by identifying the functional requirements of the framework, based on the challenges of the COTS selection process. Thereafter, we identified the approaches through which the identified functional requirements of the framework for COTS selection would be addressed. Table 1 presents the major requirements of the framework; the approaches, which may be employed to achieve the requirements; as well as the challenges addressed by the requirements.

From Table 1 it is noticed that some COTS selection challenges are addressed by more than one requirement; meaning that sometimes it may not be possible to sufficiently address a challenge through a single requirement. This is something that is generally overlooked by most COTS selection methods in the reviewed literature. The requirement for providing support for concurrent selection of COTS products for the different subsystems of COTSbased software systems (Requirement 12) was not address in the research work reported in this paper. The reasoning, we believe, was that to address this requirement, it is necessary

that the challenges of selecting a single COTS product are first addressed as much as possible.

Table 1: Requirements of a Framework for COTS Selection, Approach for Providing the Requirements, and the Challenges Addressed

No.	Requirement	Approach of providing the requirement	Challenges
1	Support iterative processing	COTS Selection process model	1. Multiple stakeholders with changing preferences) 2. Changing COTS features 3. Hierarchical decision making
2	Support for defining the COTS evaluation criteria	Support for preferences elicitation	Many Similar COTS
3	Means to estimate COTS Capabilities	Multi-Criteria Decision Making	Many Similar COTS
4	Support for identifying agreement options	Technique for determining 'best' fit solutions	1. Many Similar COTS 2. Multiple Stakeholders
5	Support for Tradeoff	Tradeoff analysis technique	1. Multiple Stakeholders 2. Uncertainty management 3. Multiple Objectives
6	Capabilities for information sharing	1. Web-based platform and discussions repository 2. Use predefined model of the COTS characteristics	Multiple Stakeholders
7	Enable learning from previous COTS selection processes	Searchable repository for knowledge, and data mining techniques	Knowledge Management
8	Provision of a COTS repository	Database technology	1. Many COTS products 2. Knowledge Management 3. Multiple Stakeholders

9	Support both individual stakeholder and group processes	1. COTS Selection process model 2. Implementation Technology for the CSDSS	1. Multiple Stakeholders 2. Knowledge management
10	Support integration into existing information system	Implementation Technology for the CSDSS	
11	Provision of component for process modeling and simulation		
12	Support for concurrent selection of COTS products for the difference subsystems of COTS-based software systems	This requirement was not addressed, thus the approaches for providing the requirement were not identified	Multiple objectives

The requirements of the framework for COTS selection can be classified into three categories, as shown in Table 2. Such a classification is of great importance during the establishment of the CSDSS, because it helps the system developer to know when a particular requirement has to be implemented. In view of the requirements classification in table 2, Figure 1 illustrates the relationship among the COTS selection framework, the COTS selection process model, the techniques that facilitate the COTS selection process, and the technologies used to integrate various techniques into a single CSDSS. Moreover, the figure illustrates the following:

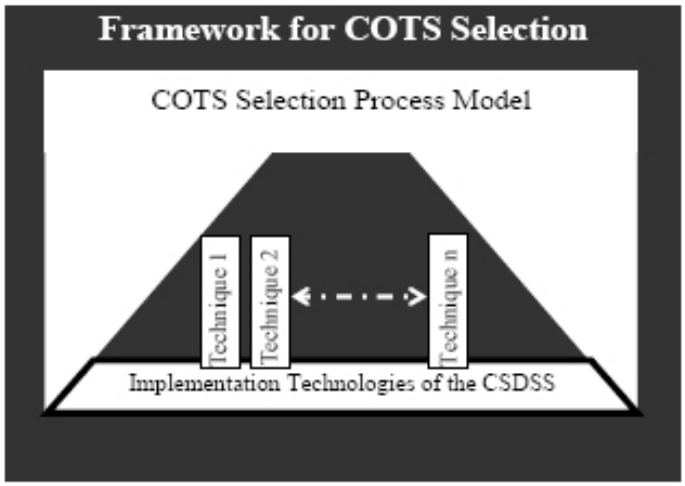
- The techniques that facilitate the COTS selection process model are used within the framework of the process model
- The integration technologies enable the inter-working of various techniques of the CSDSS as required by the process model
- The COTS selection process model, the techniques that facilitate the COTS selection process model and the technologies used to integrate the techniques into a single CSDSS, make up the framework for COTS selection

Table 2: Classification of the Requirements of the COTS Selection Framework

No.	Requirement	Category
1	Support iterative processing	1) Requirements achieved by the COTS selection process model
9	Support both individual stakeholder and group processes	
12	Support for concurrent selection of COTS products for the difference subsystems of COTS-based software systems	
2	Support for defining the COTS evaluation criteria	2) Requirements achieved through the techniques which have to be integrated into the CSDSS
3	Means to estimate COTS Capabilities	
4	Support for identifying agreement options	
6	Capabilities for information sharing	
7	Enable learning from previous COTS selection processes	
12	Support for concurrent selection of COTS products for the difference subsystems of COTS-based software systems	
8	Provision of a COT repository	3) Requirements achieved through the technologies used to implement the CSDSS
9	Support both individual stakeholder and group processes	
10	Support integration into existing information system	
11	Provision of component for process modeling and simulation	
12	Support for concurrent selection of COTS products for the difference subsystems of COTS-based software systems	

It should be noted that each of the techniques in Figure 1 can be developed into a stand alone DSA, and that it is such applications that were integrated into single CSDSS presented in this paper.

Fig 1: Components of the Framework for COTS Selection



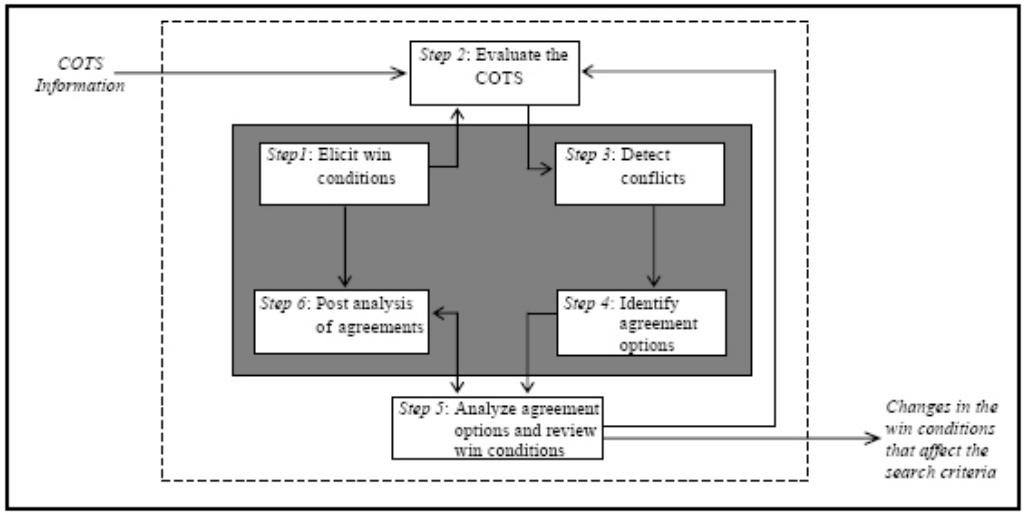
Framework For Cots Selection

We implemented the framework for COTS selection by addressing each of the requirements within the context of its category. The framework has two main components, namely: a process model that forms the basis for achieving all the requirements of the framework and guides the stakeholders through the COTS selection process, and a DSS which has techniques for providing decision support at every stage of the process model.

COTS Selection Process Model. COTS selection is a complex process with intricate relationships and interactions among the various activities of the process. This complexity is managed by breaking the process of selecting COTS products into phases that have related activities. The concept of reducing complexity by breaking the process into many smaller phases is not our creation, neither is it limited to software engineering; it is an old military strategy called divide and conquer. In the context of COTS selection, it is used to arrange the activities of the process in a logical and practical sequence, and to enable evaluation of the progress of COTS selection processes. Moreover, it allows the stakeholders to deal with small components of COTS selection, making it easy for them to master the sequence of activities upfront.

The COTS selection process model that we developed (see Figure 2) supports iterative processing (requirement 1, see Table 1) and the individual stakeholder and group activities (requirement 9, see Table1). We call it the NEgotiation MOdel for COTS SElections (NEMO-COSE), because it enables each stakeholder to individually evaluate the solution options, and it facilitates stakeholder negotiation so to identify the 'best-fit' solution.

Fig 2: The NEMO-COSE Model



To support requirements 1 and 9 of the framework for COTS selection, the model under goes the following steps:

1. ***Step 1:*** Each stakeholder defines his/her evaluation criteria, and sets the weight of each criterion (relative importance of each evaluation criteria). We refer to this information as the stakeholders' win conditions.
2. ***Step 2:*** Using a MCDM technique, every stakeholder evaluates and ranks the COTS alternatives based on his/her win conditions.
3. ***Step 3:*** The rankings of the COTS alternatives with respect to different stakeholders are generated and compared in order to identify conflicts. Though unlikely, if there are no conflicts, go to step 6. Otherwise, go to step 4.
4. ***Step 4:*** Using a conflict analysis model, identify agreement options, as well as a COTS alternative ranking that reflects the combined preferences of all stakeholders.
5. ***Step 5:*** Once the agreement options are made known to every stakeholder, they (stakeholders) each independently carry out the following activities in order to identify, and make tradeoffs required to reach agreement:
   - Using a tradeoff analysis model, analyze the agreement options.
   - Review the win conditions. If an agreement has not been reached and the evaluation criteria as well as their weights are modified, then go to step 2. Otherwise go to step six.

     If from step six, refine the changes made in the win conditions and go to step 2.
6. ***Step 6:*** Compare the new or final win conditions with the initial conditions to determine if the changes are acceptable. If the final win conditions are acceptable the process stops for that stakeholder, otherwise go back to step 5.

Ideally, the NEMO-COSE continues until the point when stakeholders reach agreement on one COTS product to select. Therefore, besides following the NEMO-COSE process, stakeholders need to share information on the tradeoff choices they have and on whether it is possible for them to make those tradeoffs or not. In practice, we do not expect the NEMO-COSE steps to be rigidly followed as presented above, because in some cases it may be necessary to move back and forth between steps; or to skip some steps as mentioned in step 3.

Techniques for Providing Decision Support for the various Steps of the COTS Selection Process. The DSS that we developed had to support all the requirements of category two (see Table 2). To achieve this, we identified or developed techniques for addressing each one of these requirements individually. Thereafter, each of the techniques was developed into an independent DSA, and an appropriate technology was used to integrate the DSAs and the COTS selection repositories into a single DSS. In the following subsections we describe how each of the requirements of Category two were addressed.

Support for Defining the COTS Evaluation Criteria (Requirement 2, step 1 of process model). In Multi-Criteria Decision Making problems such as COTS selection, it is necessary to have a precise model of the user preferences. However, studies have revealed that often people are unable to state their preferences fully up front (Falting et al., 2004). Moreover, research has revealed that people start searching for, and evaluating the solution options with, a small set of high-value preferences; but adjust the values of the original preferences as they discover the solution characteristics that can be incorporated into their preference models (Falting et al., 2004), (Pu et al,. 2003). Therefore we engendered a support system for the elicitation of user preferences based on utilizing the high-value preferences of users to incrementally develop complete and accurate preference models of the COTS selection stakeholders. We call the support system the Enhanced Preference Elicitation Model (EPEM). It is presented in detail in the paper titled, "Using Prediction to Provide Decision Support for the Elicitation of User Preferences" (Wanyama et al., 2005a). The system works as follows:

1. ***Step 1:*** The User states his/her high value preferences.
2. ***Step 2:*** Using the provided user preferences, the computer searches, evaluates and presents solution options. Moreover, the support system utilizes information from step 1 as input to the neural network to predict other preferences and preference values which the user may want to consider.
3. ***Step 3:*** As the user rejects, accepts, and/or modifies the preferences, and the solution options suggested to him/her by the support system, the system automatically establishes a new input to the neural network, resulting in a new prediction, and a new set of solution options.

Ideally, step 3 continues until the user is satisfied with the established preference model, or until the user preference model is the same as the predicted model.

Means for Estimating the Performance of COTS Products (Requirement 3, Step 2 of process model). Evaluation of COTS products usually involves determining the performance of the products in multiple features. Therefore, COTS evaluation is a Multi-Criteria Decision Making (MCDM) problem. We believe that a good CSDSS should incorporate a variety of MCDM techniques to allow users to choose the ones with which they are familiar. Such techniques include Ordered Weighted Averaging (OWA) (Yager et al., 1997), Expected Utility Method (EUM) (Davis et al., 1997) and Analytic Hierarchical Process (AHP) (Saaty, 1980). In the research that led to the framework for COTS selection that is reported in this paper, the model represented by Equation 1 was used to estimate the ability of the alternatives COTS products to satisfy project conditions.

$$Score_j = \sum_{i=1}^{k} a_{ji} w_i \,.$$

Where $Score_j$ is the ability of COTS product j to satisfy the project conditions, k is the number of evaluation criteria, ji a is the strength of COTS product j in criterion i , and w_i is the weight of criterion i . This model was used because of its simplicity, and because of the belief by the researchers that establishing the user preference model and value function, as well as determining and accurately quantifying the strength of the alternative COTS products, is far more important to the decision making process than the level of sophistication of the MCDM technique used to aggregate the evaluation data.

Support for Identifying Agreement Options (Requirement 4, Step 4 of process model). Screening of the COTS products is carried out to reduce the solution space for effective evaluation of the alternative products. However, the screening is usually based on key requirements other than the stakeholder preferences. In order to increase the influence of the stakeholder preferences at an early stage of the COTS selection process; and to further narrow down the number of alternatives to just a few promising ones before carrying out detailed analysis and negotiation among the stakeholders, it is necessary to identify the agreement options from the sets of alternative COTS products (agreement options are the COTS products that are most likely to be selected) (Xue et al., 2004). Determination of agreement options serves as a second layer of alternatives screening based on the stakeholder preferences. This process objectively reduces the 'amount' of stakeholder interaction required to reach agreement. That is, it ensures that stakeholders negotiate to select one COTS product out of just two or three most promising alternatives. We address this requirement by using a Game Theory model presented in the paper titled, "Multi-Agent System for Group-Choice Negotiation and Decision Support" (Wanyama et al., 2004). The input to the model is a set of rankings for the COTS alternatives for the individual stakeholders, and the output is a group ranking of the solutions. Support for Tradeoff. Preference over COTS evaluation criteria depends upon the perspective of the individual stakeholders. For example, the System Design

Team might be more interested in the system quality and architectural issues, whereas the customers are more interested in the domain issues related to their organizations and constraints, such as budget and schedule limits. Moreover, it is required that products are evaluated in isolation, with respect to each category of the evaluation criteria (because of interdependences among evaluation criteria in different categories), thus each stakeholder has more than one evaluation objective, each corresponding to a particular category of the evaluation criteria. Therefore, COTS evaluation involves determining the spatial viability of COTS products. That is to say, tradeoffs associated with COTS evaluation can be defined across several Categories of evaluation criteria for a single stakeholder perspective and can also be defined over several stakeholder perspectives for one or more categories of the evaluation criteria. For complete COTS evaluation, both cases of tradeoffs are required to be carried out. We achieved this requirement by using a Qualitative Reasoning Model presented in the paper titled, "Qualitative Reasoning Model for Tradeoff Analysis" (Wanyama et al., 2005b). The input to the model is the various COTS alternatives preferred by the different stakeholders, and the output is a set of graphs indicating to the concern stakeholder what he/she loses and/or gain by switching from his/her preferred COTS to another.

Capability for Information Sharing. It is common for the COTS selection stakeholders not to be co-located both in time and in space. Therefore, a framework for COTS selection should possess a communication component through which stakeholders can share information and through which they can access expert knowledge. The DSS that we developed is web-based, and it has a repository for storing discussion.

Component that Enables Learning from Previous COTS Selection Processes. COTS selection is characterized by large quantities of information. Therefore, CSDSS should have an intelligent component for intelligent knowledge retrieval, knowledge discovery and approximate reasoning that facilitates learning from historical data. Models based on techniques such as Neural Net (CorMac Technologies, 2004) can be integrated into the CSDSS to learn from examples (historical data), and generate predictions of the target variables. For the CSDSS that we implemented, this requirement was addressed with respect to the elicitation of user preferences. It can, however, be extended to the other steps of the framework for COTS selection.

Provision of a COTS Repository. Morisio et al suggest that organizations, which develop software from COTS products, should have COTS teams that act as consultants whose main responsibility is to gather and store information about COTS products. This information should be stored in repositories in a format based on predefined models to facilitate quick investigation of different scenarios involving various COTS selection aspects and views, for the final selection of COTS products. Moreover, presenting COTS information that is formatted according to predefined models enables the COTS selection problem to be represented in the

same manner for all stakeholders, which in turn enables information sharing and negotiation. For our DSS, we used the COTS selection repositories presented in the paper titled, "Repositories for COTS Selection" (Wanyama et al., 2006).

Requirements 9, 10, and 11 in Table 1. The following are the requirements identified in this subheading:

- Support for individual and group processes.
- Support for the integration of CSDSS into the existing information systems of the various stakeholders.
- Provision of a component for process modeling and simulation.

The above requirements require that an appropriate CSDSS has, among others, the following capabilities:

- Distributed problem-solving capabilities so as to offer decision support to the individual stakeholders as they develop their views and preferences.
- Handling iterative decision making both at individual and group level (Hall et al., 2002).
- Evaluation of products without views of some stakeholders and inclusion of those views (views of previously absent stakeholders) when they are available (flexibility and modularity), – support for asynchronous decision-making (Cao et al., 1999), and support for reasoning with incomplete information.
- Distributed learning in order to offer efficient and customized decision support to different stakeholders (customization of the DSS to stakeholder needs and background).
- Automated negotiation and tradeoff component that enables describing, monitoring, controlling and simulating ('what-if' analysis) the underlying characteristics of the COTS selection processes and track changes in parameters and dependences.

Software agents have characteristics that make it easy to develop and use a DSS that offers the above capabilities (Jennings et al., 1996). For a review of the characteristics of agents with respect to using software agents in decision support systems, refer to Jennings et al. When developing our DSS, we took into account the fact that Carlsson and Turban report that DSS have not been embraced fully because of the following problems:

- People find it difficult to adopt intelligent systems.
- People are unable to understand the support they get, thus they ignore DSS and opt for past experiences and visions.
- People are not able to handle large amounts of information and knowledge.
- People do not like working with theories that they do not understand.

N

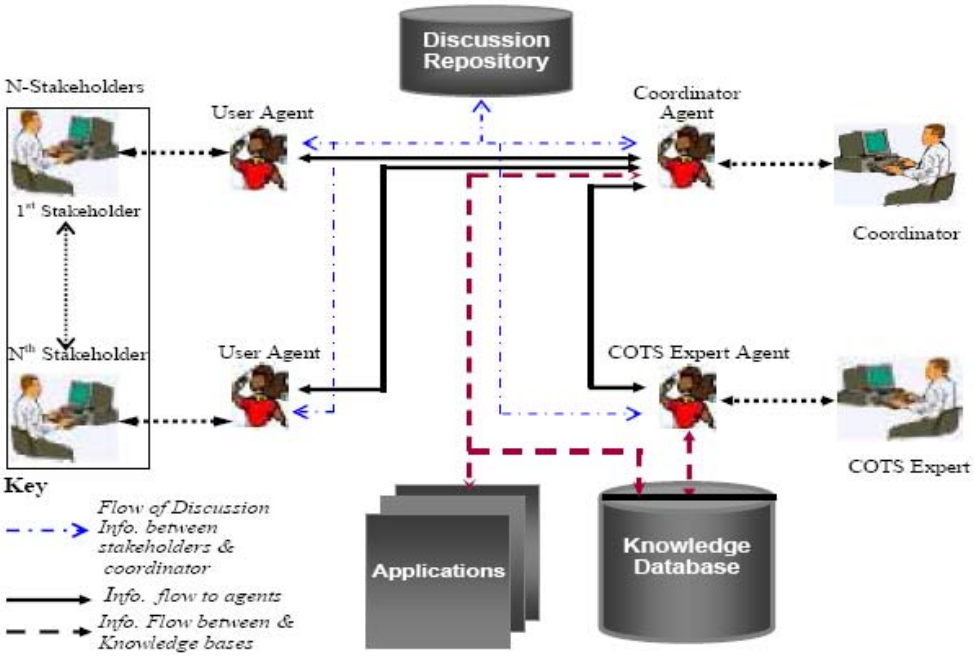
- People are more comfortable with get support from other people, than getting it from DSS.

Agent technology enabled us to address the third and fourth problems, because the agents handle most of the information, and agent technology has an inherent characteristic of masking the techniques used to provide decision support, as well as their underlying theories. Moreover, we addressed the second problem by providing GUIs that are tailored to the roles of the various COTS selection stakeholders, and that display the results both numerically and graphically. Finally, we addressed the first and fifth problems by integrate the capabilities for the following activities, in our DSS: Information Sharing, discussion, access to expert knowledge, access to COTS experts and access to lessons-learned from previous COTS selections

An Integrated Agent-Based DSS for COTS selection. Figure 3 shows a high level architecture of the DSS for the selection of COTS products. We separated the applications that provide the required techniques from the internal structure of the agents in order to increase scalability of the system. Moreover, the separation enabled using the already available applications (DSA) whenever possible.

The DSS has user agents that access the applications and the repositories to assist each stakeholder in the multicriteria decision making process. After evaluating alternative solutions according to the concerns of their individual clients (stakeholders), the user agents cooperate to identify the agreement options for the stakeholders, as well as a single ranking of alternatives that reflects the combined concerns of all stakeholders. Moreover, each user agent analyzes the agreement options to identify the tradeoff criteria and which way to adjust the criteria weights. The coordinator agent assists user agents during negotiation. It is responsible for identifying the agreement options by determining the fitness factors (see Wanyama et al., 2004) of the alternative COTS product for various stakeholder coalitions. On the other hand, the human coordinator modulates discussions among the stakeholders, and controls the behavior of the coordinator agent (negotiation support) by setting relative importance of the stakeholders. The COTS expert or the COTS team (Morisio et al., 2000) is responsible for maintaining the knowledge repository through the expert agent.

Fig 3: Multi-Agent System for Facilitation of the NEMO-COSE Process



Conclusions and Future Work

This paper reports the results of an evaluation of the state of the art COTS selection methods, identifying their strength and weaknesses. Moreover, the paper presents the following: comprehensive set of the requirements for the framework for COTS selection, a COTS selection process model that supports carrying out individual stakeholder activities and group activities concurrently, and an agent-based CSDSS that facilitates the COTS selection process model.

In the future we would like to integrate the support for requirements elicitation process with the support for COTS selection, because these two interdependent processes are normally handled concurrently in CBSD. This work may lead to modifying existing requirement elicitation methods for CBSD, or engendering a completely new method.

Furthermore, we would like to develop an ontology or representation language that would facilitate COTS vendors to state the functional features of their products in a way that enables COTS-Based Software Developers to assess and compare products. At the present, such an ontology or representation language is not available. Finally, we would like to extend the framework for COTS selection to address the issue of selecting multiple COTS products of a COTS-based software system, concurrently.

References

Alves C. (2003). "COTS-Based Requirements Engineering" Chapter of the Book Component-Based Software Quality – Methods and Techniques. Lecture Notes in Computer Science. Springer

Alves C. and J. Castro (2003. "CRE: A Systematic Method for Components Selection", available at URL: http://www.cs.ucl.ac.uk

Burgues X, C. Estay, X. Franch, J. A. Pastor, and C. Quer (2002). "Combined Selection of COTS Components, Procurement of COTS Software Components", Proceedings of ICCBSS, pp.54-64

Cao P. P., F. C. Burstein (1999). "An Asynchronous Group Decision Support System Study for Intelligent Multicriteria

Decision making", Proceedings of the 32nd Hawaii International Conference on System Sciences

Carlsson C. and E. Turban (2002). "DSS: directions for the next decade", Decision Support Systems, Vol. 33, No. 2, pp. 105-110

Cavanaugh B. P. and S. M. Polen (2002). "Add Decision Analysis to Your COTS Selection Process", The Journal of Defense Software Engineering

Chung L. and K. Cooper (2002). " A knowledge-based COTS-aware requirements engineering approach", SEKE 2002, pp. 175-182

Comella-Dorda S., J. C. Dean, and E. Morris, P. Oberndorf (2002). "A Process for COTS Software Product Evaluation", Proceedings of ICCBSS, pp. 86- 96

CorMac Technologies (2004). "What is a Neural Net?", Available at URL: http://www.cormactech.com

Davis J., W. Hands, and U. Maki (1997). "Handbook of Economic Methodology" Edward Elgar, London, pp. 342- 350

Faltings B., P. Pu, and M. Torrens (2004). "Design Example-critiquing Interaction", IUI, Funchal, January 13th -16th, 2004.

Hall D., Y. Guo, and R. A. Davis (2002). "Developing a Value-Based Decision-Making Model for inquiring Organizations", Proceedings of the 36th Hawaii International Conference on System Sciences

Hansen W. J. A. (2003). "Generic Process and Terminology", available at URL: http://www.sei.cmu.edu accessed

Holmes L. A. (2000). "Evaluating COTS using the Function Fit Analysis", CROSSTALK: Journal of Defense Software Engineering, Available at URL: http://www.stsc.hill.af.mil

Jennings N.R., P. Faratin, M.J. Johnson, P. O'Brien, M.E. Wiegand: Using Intelligent Agents to Manage Business Processes. Proceedings of PAAM (1996) 345-360

Kontio J., S. F. Chen, and K. Limperos (1995). "A COTS Selection Method and Experiences of its Use", Twentieth Annual Software Engineering Workshop, NASA Goddard Space Flight Center, Greenbelt, Maryland

Kunda D. and L. Brooks (1999). "Applying Socio-Technical Approach for COTS Selection", Proceedings of 4th UKAIS Conference, University of York, McGraw Hill

Morisio M., C. B. Seaman, A.T. Parra, V.R. Basilli, S.E. Kraft, and S.E. Condon (2000). "Investigating and Improving a COTS-Based Software Development", Process. ICSE 2000. Limmerick, Ireland.

Morisio M., C. B. Seaman, V. R. Basili, A. T. Parra, S. E. Kraft, and S. E. Condon (2002). "COTS-Based Software Development: Processes and Open Issues", Journal of Systems and Software, Vol. 61 No. 3, pp. 189- 199

Ncube C. and J. C. Dean, "The Limitations of Current Decision-Making Techniques in the Procurement of COTS Software Components", Proceedings of ICCBSS, pp.176-187,

Ncube C. and N. A. M. Maiden (2003). "PORE: Procurement-Oriented Requirements Engineering Method for the Component-Based Systems Engineering Development Paradigm", available at URL: http://www.soi.city.ac.uk,

Ochs M. and G. Chrobok-Diening (2000). "A COTS Acquisition Process: Definition and Application Experience", ISERN Report

Pu P., B. Faltings, and M. Torrens, "User-Involved Preference Elicitation", In workshop notes, workshop on Configuration, the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), 2003.

Saaty T. L. (1980). "The Analytic Hierarch Process", Wiley, New York

Wanyama T. and B. Far (2004). " Multi-Agent System for Group-Choice Negotiation and Decision Support", The 6th International workshop on Agent-Oriented Information Systems at AAMAS2004; New York – USA

Wanyama T. and B. Far (2005)a. "Using Prediction to Provide Decision Support for the Elicitation of User Preferences", The proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2005, Saskatoon, Saskatchewan, Canada

Wanyama T. and B. Far (2005)b. "Qualitative Reasoning Model for Tradeoff Analysis", MDAI2005, Lecture Notes in Artificial Intelligence (LNAI 3558), pp.99-109, Springer-Verlag, Berlin Heidelberg

Wanyama T. and B. Far (2005)c. "Towards Providing Decision Support for COTS Selection", The proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2005, Saskatoon Saskatchewan, Canada

Wanyama T. and B. Far (2006) "Repositories for COTS Selection", The proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2006, May 1-3, 2006, Ottawa, Ontario, Canada

Xue D. and H. Yang (2004). "A Concurrent Engineering-Oriented Design Database Representation Model", Computer Aided Design, Vol. 36, pp. 947-965

Yager R. and J. Kacprzyk (1997). "The Ordered Weighted Averaging Operators, Theory and Applications", Kluwer Academic Publishers, Boston

34
Not all Visualizations are Useful: The Need to Target User Needs when Visualizing Object Oriented Software

Mariam Sensalire and Patrick Ogao

"A picture is worth a thousand words". In the software field, this is justified by the increasing research into software visualization. Pictures are increasingly being used to represent software code with many studies establishing that they improve comprehension. This paper discusses results from observing expert programmers use 3 visualization tools. The results show that if tools are developed without user consultation, they may fail to be useful for the users. The necessity for developers of tools to target the needs of the users for whom the tools are aimed is further discussed.

1. Introduction

Understanding a program is usually of great importance to a software project. Whether a programmer is maintaining or developing a program, it is paramount that they understand the program being worked on [12]. In many cases, when trying to understand software, many programmers have to read the lines of code of a particular program. This approach is however not feasible as the program size increases [19]. As a supplement method, visualization has been used to increase program understanding [8]. Visualization is the process of transforming data into insight [26]. It can also be looked at as the technique for creating images, diagrams, or animations to communicate a message [6]. Software is complex, multifaceted, large, and contains many relationships between its component parts.

Therefore there are many aspects of software that may be appropriate for visualization [9]. Previously, different forms of visual presentations were used in understanding software, among which were the UML diagrams. While these mechanisms are good for representing a small component of the software, they are not scalable with regard to large software systems [24]. Software visualization aims to address these shortcomings and many designers have therefore embarked on developing software visualization tools to meet these demands. Despite this, not enough work is put into determining the desired features for such tools [27]. As such many of the designed tools are not used in practice. To be able to design a software visualization tool that is effective however, a proper needs assessment of the target users may need to be carried out.

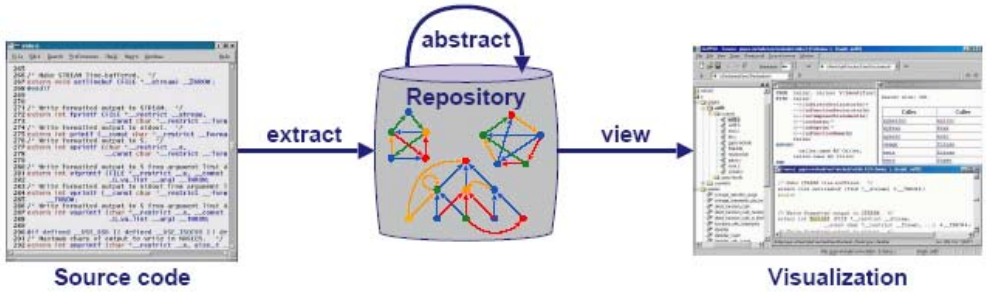
2. Motivation

It has been noted in the past that very few of the software visualization systems developed are systematically evaluated to ascertain their effectiveness [18]. This means that many of the developed software visualization systems may not even be appropriate for their aims. Price [18], in his taxonomy, creates a scope for effectiveness of software visualization systems. It looks at the purpose of the system, the appropriateness and clarity, as well as empirical evaluation. It is further noted by [24] that a useful software visualization tool can only be produced if it is based on a study of program understanding.

It is important for software visualization tools to actually be able to increase understanding. The images that are generated have to be easily interpreted by the user in relation to the structure of the software being visualized [1] as effective visualizations can be of assistance even to expert programmers. In many cases, expert programmers join projects in whose initiation they were not involved [2], as a result, they are faced with the challenge of contributing substantially to a programming project whose low level and high-level structure they may not understand. Such programmers can use visualizations as a starting point to understanding the software.

There are other cases where programmers are recruited and given access to the program documentation with the aim of making small changes to already implemented projects. These programmers also face challenges. This is due to the fact that documentation for software is usually not updated in parallel with the software changes over time [2]. A result is documentation that may not reflect the actual status of the program being worked on. Due to this, many programmers consider source code to be the most trusted form of documentation making it critical to be understood [25].

Fig1: Visualizing form source code



An example of visualizing from source code is shown in figure 1 (adapted from [4]). The first part represents the source code, which can be received in its raw form or exported to a format that the tool expects. Once the tool receives the right format of the code, the visualizations are then generated.

A programmer who does not understand a programs source code can easily make changes which have negative effects in other parts of the software leading to

undesirable effects for the whole project ([2], [24]). If programmers had the ability to know how the different components in a program relate, such problems would be overcome. Even when programmers have access to the source code as well as UML diagrams of the system, it is still not easy to understand the system since these notations do not scale up well with respect to comprehension [11].

Source code becomes incomprehensible beyond hundreds of lines of code while UML decreases information density, which in turn limits scalability [1]. With both these systems, it is quite difficult to see an entire system. Visualization has been shown to improve the productivity and effectiveness of programmers, especially when working with complex systems, which can span the work of several people over extended periods of time [5]. Diagrams of a system are much easier to understand and they convey a lot of information in a standardized way [2]. It is also important to visualize from source code because in many situations, it may be the only source of information available about a given program [25].

3. Previous Work

In the past, there has been a lot of research on visualizing Object Oriented Software. Jerding and Stasko [8], were among the first to advocate for the use of visualization to foster the understanding of O-O software. Those events that need to be visualized to enable comprehension of O-O software were specified. Jerding and Stasko [8] however acknowledged the difficulty in gathering the necessary information to construct useful visualizations. Based on these difficulties, a framework was proposed which recommends that: Visualizations should require no programmer intervention once developed and they should present the aspects of a program that are important and will be of use to programmers. The framework also specifies that visualizations should be able to cater for the programmers' needs in a timely manner, and lastly should be able to handle real world problems. Jerding and Stasko [8] however did not mention carrying out a user study to justify their conclusion despite the fact that a user study is critical for producing useful visualization tools [8]. Malectic et al, [11] also made a case for visualizing O-O software in virtual reality. The subsequent Imsovision tool proposed was based on Shneiderman's [22] Task by Data Type for Information Visualization which specified a visual design guideline of overview first, zoom and filter, details on demand, relate, history and extract. Imsovision was however not based on any existing software comprehension models nor did it have a previous validation procedure to prove that virtual reality enables comprehension better than 2D or 3D methods.

In terms of proximity, this work closely relates to that of Pacione [17] who supported the use of a multifaceted three-dimensional abstraction model for software visualization. In that study, effective presentation techniques for visualization were presented. These were specified as the use of diagrams for describing software as well as the use of views for software comprehension. Diagrams were noted to be more effective if used in an interrelated hierarchical manner that addresses all levels

of abstraction. It was also specified that the use of multiple interdependent views was the best arrangement in relation to comprehension. Pacione however noted that identifying the views that were appropriate for particular comprehension tasks was still a challenge. It was also stated that the usefulness of the multifaceted three-dimensional model in software comprehension would be evaluated in future. This therefore means that the procedure proposed was yet to be proven as effective since it was not yet evaluated. Ihantola [7] on the other hand looked at algorithm visualization and specified a taxonomy for effortless creation of algorithm visualizations. The justification for this taxonomy was the identified lack of use of algorithm visualization tools beyond the labs, a problem that also applies to software visualization tools.

The view taken by that study was that the difficulty of creating visualizations using the developed tools led to their lack of reception in the classrooms where they were needed the most. This view was arrived at after consulting the different teachers of algorithms and getting their views on the source of the problem. Ihantola [7] also noted that many developers create systems based on their own needs or beliefs about others' needs.

This problem was also noted to be prevalent in software visualization tools ([28],[21]) thus justifying the need for user consultation before tool development. Maletic at al. [11] also build the case for identifying the most appropriate visualization techniques for given software visualization tasks. It is proposed that this is done after considering why the visualization is needed, the people that will be using it, the kind of information it is going to be representing as well as the medium of representation. This is the same procedure that will be used in this research with specialization put on O-O software.

4. The experiment

This section looks at the design and conduct of an experiment observing five expert programmers use three visualization tools. The following section discusses the programmers' use of the tools, in an effort to show what useful visualizations should be like as well as the process of generating them.

4.1 Tools and Source Code

Three tools were used for the study. These were Code Crawler ([20], [10]), Creole ([3], [14]), and Source Navigator [23]. These tools can all visualize Object oriented software. In particular, they can all visualize Java, which was the language that the source code used was in. The tools are all freely available and use different techniques for visualizing.

Fig 2: Display from Code-Crawler

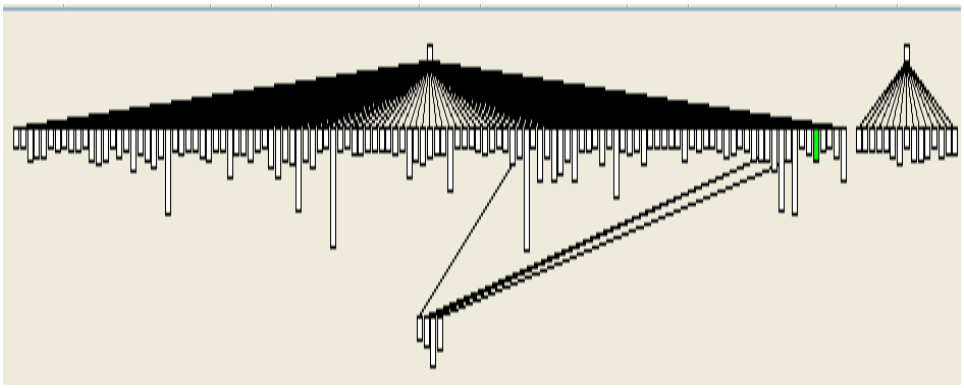


Figure 2 above shows the system complexity diagram display from Code-crawler It is the default diagram generated by Code-Crawler when visualizing code.

Fig 3: Display from Source Navigator

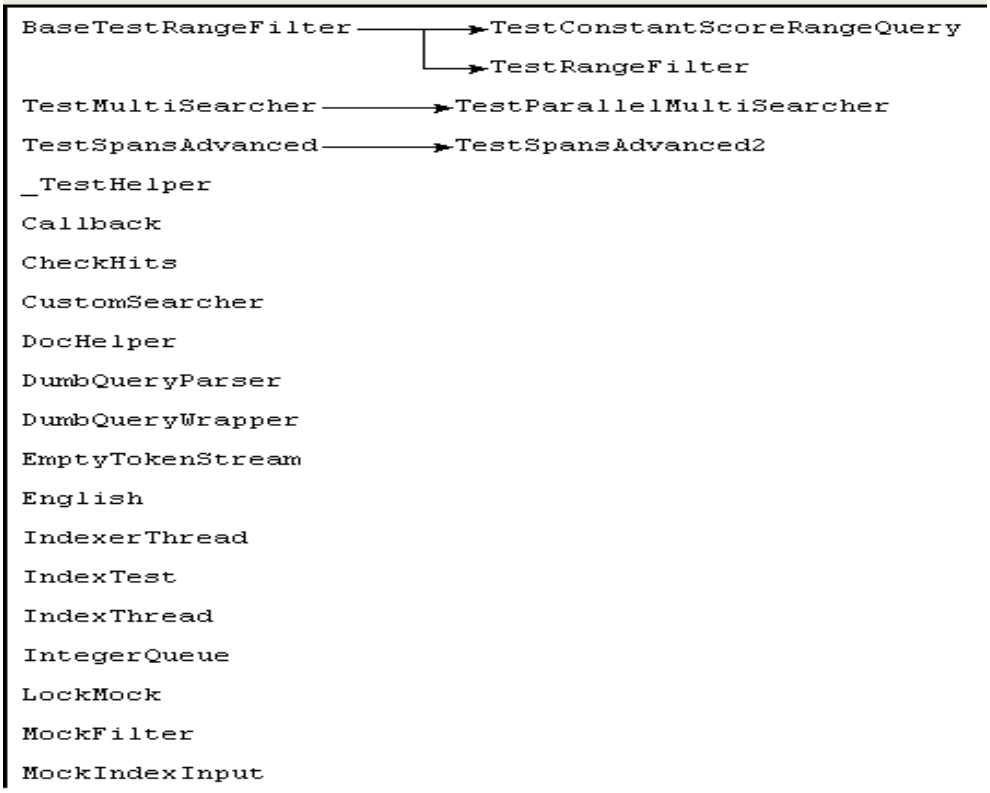
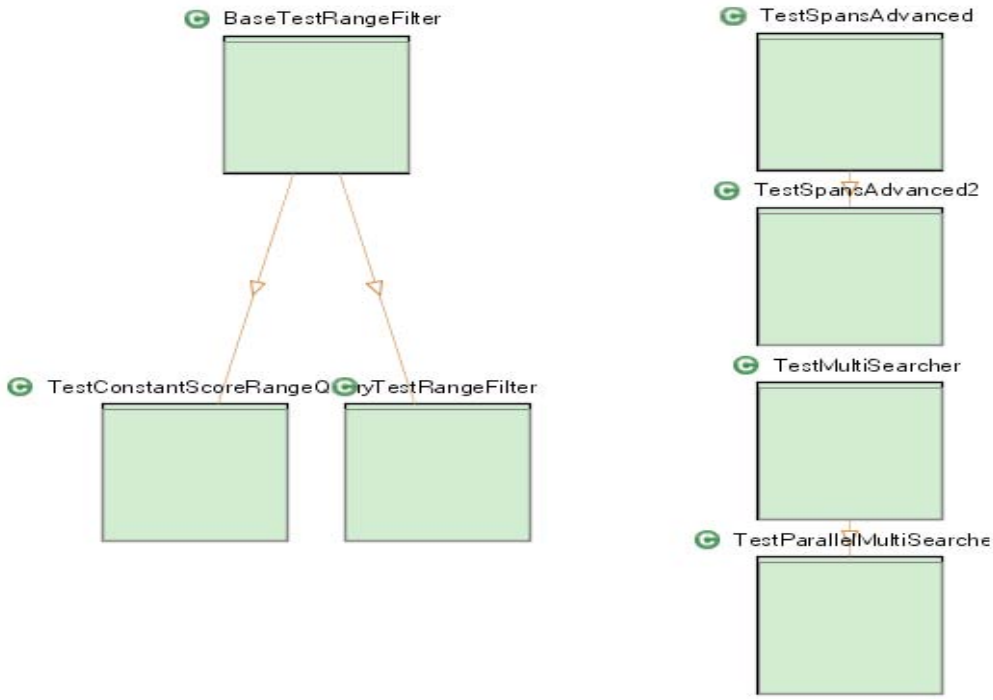


Fig 4: Display from Creole



while figure 3 and figure 4 display the sample class diagrams from Source Navigator and Creole respectively.

The source code displayed represents one package from the total packages included in the Lucene search engine. That package has 147 classes with 687 methods.

Three different sets of source code were used for the three different tools. Code Crawler was evaluated using the Lucene search engine source code; Creole was tested using Apache beehive code whereas Source Navigator was tried using Apache tomcat code. The source code was in the same range in terms of size.

4.2 Participants

Five expert programmers participated in the study. They were all male with over ten years experience both in programming and computer usage. They were experienced with the object oriented paradigm with knowledge of at least two object oriented languages as shown in table 1.

Table 1. Summary of participants

User	Computer Usage	Langs	Programming
1.	>10 Yrs	Java C++ SmallTalk	>10 Yrs
2.	>10 Yrs	Java C++ Python	>10 Yrs
3.	>10 Yrs	Java C++ Python	>10 Yrs
4.	>10 Yrs	Java C++ Ruby	>10 Yrs
5.	>10 Yrs	Java C++	>10 Yrs

Java was the language that was frequently used by all the expert programmers. The experiments were therefore carried out using Java source code.

Creole's default display, shown in figure 4, is that of packages. A double click into a particular package shows the classes and methods inside the package as well as any links between each other.

Each participant was given a 5 minutes introduction to a tool, after which they had 5 extra minutes within which to familiarize with the tools themselves or seek any extra information.

After the familiarization stage, 2 tasks were given to the participant for each tool, one task at a time. The tasks given out were:

i) Describe the static structure of the system, ie the main classes and their relationships.

ii) What would be the effect of deleting the Hook class?

Those tasks were replicated for all the three tools, however the second task was modified according to the source code being analyzed.

i) For Creole, the second task was changed to "What would be the effect of deleting the org.apache.beehive.netui.tags"

ii) Source Navigator's second task was naming the effect of deleting the DbStoreTest class.

There was a 1-minute break between the completion of tasks for each tool after which the second tool was evaluated and the third accordingly. The planned time for the experiment was 1 and half hours calculated as shown by figure 5.

Fig 5: Timing of the tasks

Task	Times Task Carried Out	Duration Per Task In Minutes	Total Duration In Minutes
Pre-study questionnaire	1	5	5
Introduction to tool	3	5	15
Familiarize with tool	3	5	15
2 assignments	3	7	42
Post-study questionnaires	3	5	15
			92 min

4.3 Experimental Analysis

The details of the experiment were analyzed as follows: The pre-study questionnaire aimed at establishing the computer knowledge of the experts in order to know how it would affect their performance. It also sought confirmation on whether the users had used a software visualization tool before and if indeed they were expert programmers.

Tool usage instructions were also availed to the candidates before the actual experiment and sample tasks given to them. These tasks were similar to the ones in the experiment only that the code used was different. This stage helped in clearing any questions that the users had about the tools and also ensuring that working knowledge of the tools had been established before the actual experiment.

During the actual experiment, Tasks about the makeup of the code that was being analyzed were given and answers written down for all the three tools. After this stage, the post study questionnaire was availed. This questionnaire aimed at establishing the shortcomings of the tools that were used during the experiment as well as their positive points. Questions also sought to know what extra features the users needed but felt were lacking in the tools. A combination of the answers given in this questionnaire as well as observations during the experiment led to the results for the study.

5. Results

In this section, the results from the study are discussed. The importance of targeting user needs when developing visualization tools is also shown based on these results.

Summary of Results

Tool	Best Features	Additional Features needed
Creole	The Zooming Visualization Interface Eclipse Integration Search Abilities	Reduced Crowding Better Speed
Code-Crawler	Code-Complexity View	Direct Code Access Less Steps for tasks IDE Integration Package views
Source Navigator	Simple Interface Speed Ease of Use Ability to view code	Better Lay out IDE Integration

Based on the results of the study, there was a lot that was still lacking from all the visualization displays as well as the tools used for the experiment. Some of the queries from the expert programmers are summarized below.

The displays were too crowded. The visualizations tended to display too much information in a small area rendering the display unhelpful for the programmers. The ability to search the visualization was another desired component that the users felt could have been addressed better. Even a good visualization that cannot be manipulated was not found to be useful.

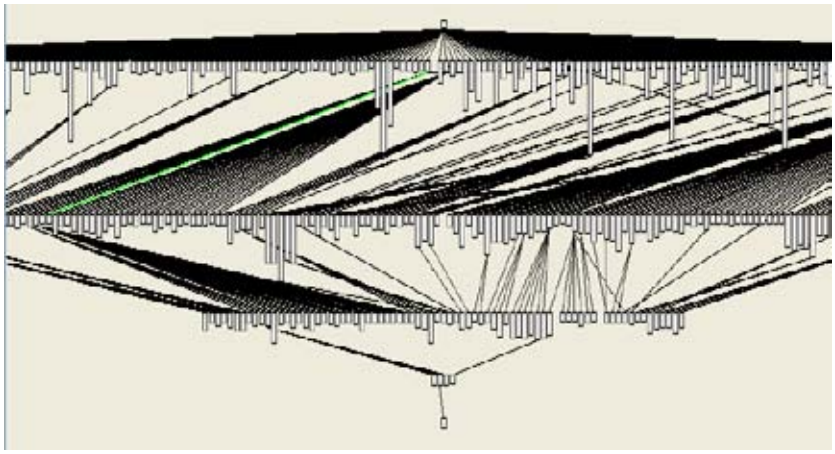
The speed of achieving the visualization was also highly complained about. More than two users stopped the generation of call graphs because they felt it was taking too long. So even after the visualization has been generated if the timing is not right then it is not considered useful.

There was also concern about integrating the visualization tools with an IDE. The reasoning was that when one visualizes, it is usually for a purpose. If the desire were to add more code to existing software, then it would be too much effort to switch between the visualization tool and the environment that is being used to program. So even if a tool is able to generate amazing visualization, the effort and time spent switching between the two environments may have on effect

on the knowledge for programming. The expert programmers appreciated tools that were able to generate visualizations in minimal steps.

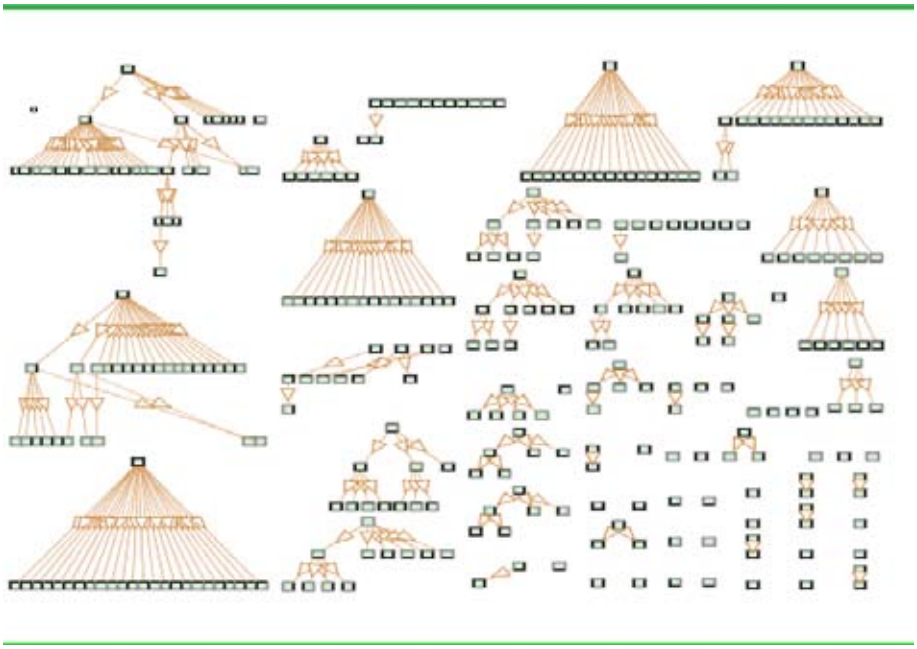
5.2: Interpretation

Fig 6: System complexity diagram of Lucene source code in CodeCrawler.



The results however showed that more time than had been planned was spent on the tasks. Figure 7. Class interface hierarchy displayed in Creole. As shown in figure 6 by the partial visualization from Lucene above, too much information in a single display may be complex to understand.

Fig 7: Shows Lucene source code as displayed by Creole.



The creole display was appreciated by the participants of the study due to the ability to zoom in and expand particular components as well. Speed was however still an issue.

Fig 8: Lucene displayed by Source Navigator

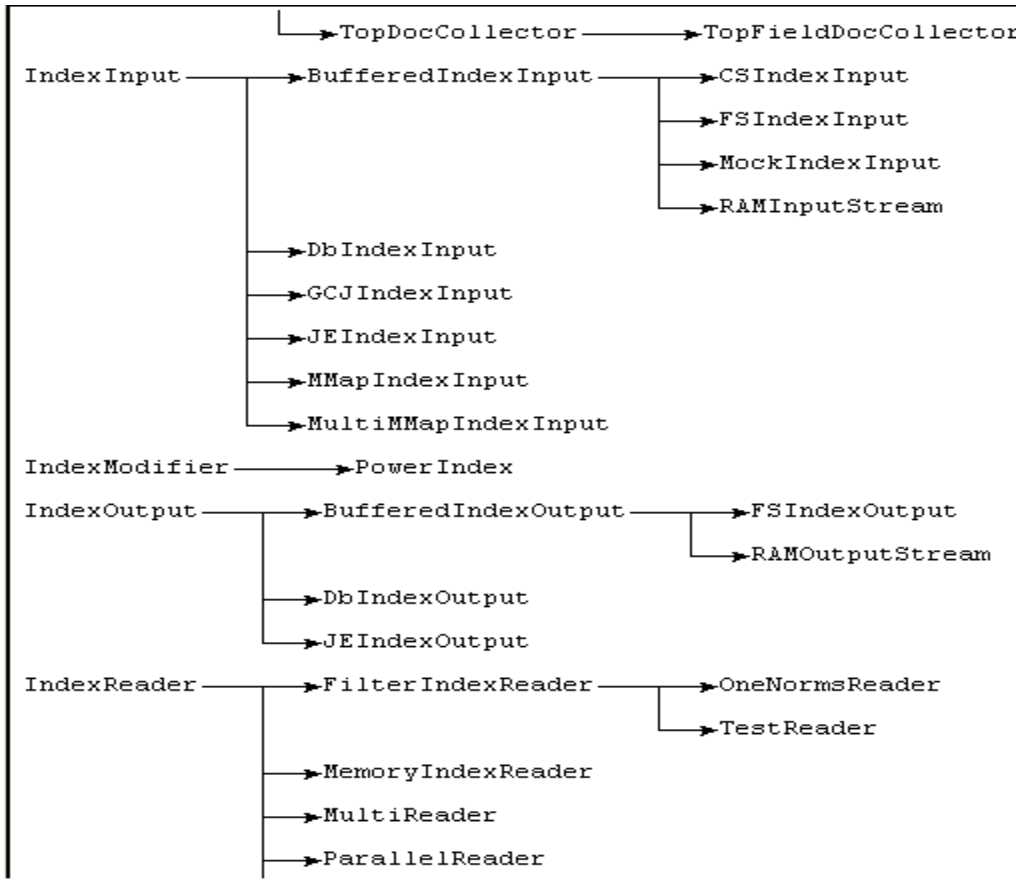


Figure 8 shows the same source code that was displayed in Creole as displayed in Source Navigator.

5.3. Effective Visualization

There are also different views from the literature about what an ideal visualization must be like. According to Knight [9] a visualization can only be considered effective if results can be achieved by users that use it without incurring any extra cost. It has to be able to reduce complexity while increasing understanding and must be able to capture the viewers attention [15]. Staples, [24] adds that a visualization which provides information that can be readily got by inspection of code, is not effective.

The presentation techniques also have to be chosen carefully as the style in which a visualization is specified affects its usability (Price [24], Schafer [21],Tory

[28]). If the visualization tool however requires too much effort to use, it would not be adopted in the real world regardless of the advancements of the techniques [2]. There is therefore need to know the kind of presentation that is suitable for the users and the task to be visualized [15]. This re-emphasizes the need to consult the group for which the visualizations are aimed before developing the tools.

5.4. Conclusions and Future work

From the results of the study carried out combined with literature from previous studies, it can be concluded that better visualization tools and techniques can be achieved if the views of the target group are sought before hand. Future work will re-visit program comprehension models in an effort to supplement the current work with the aim of better visualizations. It is not feasible to enforce processes on programmers when they are not supported by validated cognition models [13]. Despite this, some cognition models developed in the past were established after carrying out experiments using very few lines Of code [13]. Future work would include carrying out a comprehensive study on object oriented software comprehension to supplement the existing comprehension models.

References

Balzer, M. and Noack, A. Software landscapes: Visualizing the structure of large software systems. Joint EUROGRAPHICS-IEEE TVGD Symposium on Visualization, 2004.

Chan, A. and Holmes, R. Prawn: An interactivetool for softwarevisualization. University of British Columbia, Vancouver, Canada, 2000.

CHISEL. http://www.thechiselgroup.org/?q=creole. University of Victoria, Canada, 2006.

Ebert J, Kullbach B, and Riediger W. Gupro generic understanding of programs an overview. Universitt Koblenz- Landau, Institut fr Softwaretechnik , 2002.

Eng, D. Combining static and dynamic data in code visualization. PASTE02 Chalseton, SC, USA, 2002.

Gomez, H. Softwarevisualization: An overview. UPGRADE, II:12, 2001.

Ihantola, P. and Karavirta, V. Taxonomy of effortless creation of algorithm visualizations. Department of computer science and Technology Helsinki University of Technology,Finland., 2004.

Jerding, D. and Stasko, J. Using visualization to foster object oriented program understanding. Georgia Institute of Technology, 1994.

Knight, C. Software visualization conundrums. .Research Institute in Software Evolution University of Durman ,UK. , 2001.

Lanza, M. Codecrawler lessons learned in building a software visualization tool michele lanza. Software Composition Group - University of Berne, Switzerland, 2002.

Maletic J, Leigh J, and Marcus A. Visualizing object oriented software in virtual reality. Division of Computer Science, The University of Mephis TN., 2000.

Marcus, A, and Feng, L. Comprehension of software analysis data using 3d visualization. 2001.

Mayrhauser, A. and Vans M. Comprehension process during large scale maintenance. Department of Computer Science Colorado State University, Fort Collins CO.,, 1994.

Michaud R, Storey M, and Wu X. Plugging-in visualization: Experiences integrating a visualization tool with eclipse,. ACM Symp. on Software Visualization, (Softvis' 2003),, 2003.

Nevalainen S. and Sajaniemi J. Short-term effects of graphical versus textual visualization of variables on program perception. 17thWorkshop of the Psychology of Programming Interest Group, Sussex University,, 2005.

Pacione, M. A Book He Wrote. His Publisher, Erewhon, NC, 1999.

Pacione M, Roper, and Wood M. A comparative evaluation of dynamic visualization tools. Proceedings of the 10th Working Conference on Reverse Engineering (WCRE),Victoria, BC,Los Alamitos, pp. 80-89, CA: IEEE CS Press,, 2003.

Price A, Baecker R, and Small I. A principled taxonomy of software visualization. Journal of Visual Languages and Computing 4(3):211-266, 1993.

Rajlich V and Wilde N. The role of concepts in program comprehension. The InternationalWireless Industry Consortium, pages 271–278, 2002.

SCG. www.iam.unibe.ch/˜scg/research/moose/index.html. Software Composition Group - University of Berne, Switzerland, 2006.

Schafer T. Towards more flexibility in software visualization tools. Department of Computer Science Darmstadt University of Technology., 2005.

Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. Department of Computer Science University of Maryland, College Park Maryland., 1996.

Source-Navigator-Team. http://sourcenav.sourceforge.net,. GNU,, 2006.

Staples S. and Bieman J. 3-d visualization of software structure. School of Computing Colorado State University, BC, Canada., 1999.

Storey M. On integrating visualization techniques for effective software explorations. School of Computing Simon Fraser University, Burnaby, BC, Canada, 1997.

Storey M. Theories, tools and research methods in program comprehension: past, present and future. SOFTWARE QUALITY JOURNAL, 2006.

Storey M. and D. Frachiaa. Cognitive design elements to support the construction of a mental model during software. School of Computing Simon Fraser University,Burnaby, BC, Canada, 1997.

Tory M. and Moller T. Human factors in visualization research. IEEE Transactions on Visualization and Computer Graphics Vol 10, No.1, 2004.

35
Towards Compositional Support for a Heterogeneous Repository of Software Components

Agnes F. N. Lumala and Ezra K. Mugisa

Composition of Software Components is a key activity in a component-based software development lifecycle. In this paper we highlight the importance of having development environments for component composition. Currently, software components are developed under different standards, therefore introducing the challenge of heterogeneity at the time of composition. We propose research to define composition within an environment of heterogeneous software artifacts. We then propose a strategy to handle the research. As a first step towards solving the research problem, this paper concentrates on defining a software component and outlines our future work.

Introduction

The ever growing need to produce larger complex systems in the shortest possible time, at the lowest cost, has promoted a shift to paradigms such as component-based software engineering. It is important that all software engineering stakeholders promote such paradigms in a bid to give the customer the best of service. This widely accepted fact serves as a foundation for our research.

Motivations

Component-based software development (CBD) advocates the use of existing artifacts of the software development process to build systems. This does not exclude any artifact of the software development process. In fact, reuse in the early lifecycle is claimed to have higher potential pay off in terms of quality and productivity (Ali and Du, 2003). However, most current efforts are towards the use of existing artifacts of the implementation activity of the software development process. This claim is supported by results of a survey we did of a number of software repositories (Guo and Luqi, 2000; Kung-Kiu and Zheng, 2005; Lee et al., 2003; MacroVista; LogicLibrary; SourceBook; SourceForge; Tarvar) to find out the type of artifacts which they store. We randomly sampled repositories which store one or more software artifacts irrespective of what type of artifacts they store. We studied the types of artifacts each repository stores. The types of artifacts have been generalized based on the generic activities of the software development lifecycle. Table 1, shows the results of our survey.

Table 1: Types of Artifacts Stored in Different Repositories

	Stored Software Artifact								
Repository	**Requirements**	**Designs**	**Patterns**	**Architecture**	**Source Code**	**Test Suites**	**Applications**	**Error & Trouble Shooting Routines**	**Documentation**
+1 Reuse		*			*				*
SALMS	*	*	*	*	*	*	*	*	*
AIRS					*				
DSRS	*	*	*	*	*	*	*	*	*
WSRD	*	*	*	*	*	*	*	*	*
PAL					*				
Logidex	*	*	*	*	*	*	*	*	*
AFDSRS	*	*	*	*	*	*	*	*	*
Visual SourceBook					*				
SourceForge					*		*		
CodeVista					*				
VB Code Library					*				
PHP Code Snippet Library					*				
ToolBox in BDK					*		*		
J2EE Server							*		
COM Server							*		
CCM Server							*		
SOFA Template Repository		*	*	*					
KobrA components file system	*	*	*	*					

AFDSRS: AirForce Defense Software Repository System SALMS: Software Asset Library Management System DSRS: Defense Sofwtare repository system WSRD: Worldwide Software Reuse Discovery Library PAL: Public Ada Library BDK: Beans Development Kit

From table 1, it is clear that only 5 of the 19 repositories surveyed, store artifacts of the whole software development process. Only 2 of the surveyed repositories do not store artifacts of the implementation activities of the software development process.

Early reuse has various challenges as advanced by (Ali and Du, 2003; Cybulski et a.l, 1998; Li and Van Katwijk, 1992; Rubin, 1990) One key challenge of early reuse is the lack of tools and reuse-friendly environments needed to foster reuse of the early artifacts. This motivates our research to create a component based development environment.

The Software Supermarket (Sos)

The SoS is a bigger research aiming at providing a “one stop shop” for component

based software development. That is a place to obtain components, compose them and/or obtain systems assembled from various components. It is therefore an appropriate case for validating and verifying the results of our proposed research on composition. We therefore introduce it in this paper before we discuss the proposed research.

As a result of the activities to build the SoS, a heterogeneous repository to house the various components has been developed (Pyne and Mugisa, 2004). We look at heterogeneity in this case, from two perspectives. In the first perspective, the repository houses all artifacts of the software development process. Secondly, the repository houses artifacts which have been built under different standards. Essential elements that a development environment needs to support the said repository, were identified by Pyne and Mugisa , (2004) as:

- A facility for component composition.
- A mechanism for ensuring that each component used is properly self-described.
- A facility for component adaptation.
- Support for heterogeneous software repositories.
- Support for all life cycle activities.
- Mechanism for retrieval of components from the repository based on behavior specification.
- Platform independent target environment.
- Collaboration with third party repositories.
- Cooperation with other development environments.
- Scalability motivated by changing user requirements.
- Support for developers to create new components.
- Support for application testing.
- Support for access to remote repository.
- A formal specification tool.
- A simulation environment.
- A code generation facility.
- Graphical user interfacing (GUI).

It was noted that no current environment satisfies all the identified elements. One of the elements is a facility for composition i.e. a facility to enable the composition of deployable components in assembling a new application or maintaining an existing one. Most of the existing environments have a facility for component composition. However, most of the environments support homogeneous repositories. This research therefore generally aims to define composition within an environment of heterogeneous software artifacts.

Approach to the Research

Composition of software components can occur both at the design and deployment phases of the life cycle of a component (Kung-Kiu and Zheng, 2006). Therefore we need to formally define how components can be composed into composites and deposited back into our repository for future retrieval. On the other hand we need to formally define how component binaries can be composed in order to realize an executable system. In order to address composition challenges, we need to address static and dynamic aspects of composition as well (Kiziltan et al., 2000). In both these aspects we are talking about connection-oriented activities. We therefore intend to use a model that supports connectors for component composition. We find the Abstract Behavior Type (ABT) model (Arbab, 2005; 2006) appropriate to define connectors that will assist in:

- Binding provided and required interfaces of components to be composed,
- Handling non-functional requirements that result from interaction of components.

The ABT model supports exogenous coordination of components which in turn reduces coupling problems, therefore being more suitable for composition of our heterogeneous components. Before we embark on the task of formal definition of software composition, we have to define a software component. Our main reason for this action is that Component-Based Software Engineering lacks universal definitions for some of its core concepts like ' a software component'.

What is a Software Component?

There are as many definitions of a software component as there are component users. An interesting discussion of various definitions is given in Broy et.al: (1998). A review of relevant literature gives some of the numerous definitions as below:

1. A component is a unit of composition with contractually specified interfaces and explicit context dependencies only. A software component can be deployed independently and is subject to composition by third parties. This definition is widely adopted (Szyperski et al., 2002).
2. A component is a software element that conforms to a component model and can be independently deployed and composed without modification according to a composition standard (Heinman and Councill, 2001).
3. A component is a software element (modular unit) satisfying the following conditions (Meyer, 2003):
   (a) It can be used by other software elements, its 'clients'.
   (b) It possesses an official usage description, which is sufficient for a client author to use it.
   (c) It is not tied to any fixed set of clients.
4. A component is a"static abstraction with plugs". "Static" because a software component is a long-lived entity that can be stored in a software base,

independently of the applications in which it has been used. "Abstraction" in the definition means that a component puts more or less an opaque boundary around the software it encapsulates. "With plugs" means that there are well defined ways to interact and communicate with the component (parameters, ports, messages, e.t.c) Nierstranz and Dami, (1995).

5. A component is a deployable, independent unit of software that is completely defined and accessed through a set of interfaces (Sommerville, 2004).
6. A component is a language neutral independently implemented package of software services, delivered in an encapsulated and replaceable container, accessed via one or more published interfaces (Sparling, 2000).
7. A software component is a physical packaging of executable software with a well-defined and published interface (Hopkins, 2000).
8. Brereton and Budgen, (2000) describes software components as units of independent production, acquisition and deployment that interact to form a functional system.
9. D'Souza and Wills, (1999) define a component as a reusable part of software, which is independently developed, and can be brought together with other components to build larger units. It maybe adapted but may not be modified.

The definitions above are clearly different. All the definitions except D'Souza and Wills (1999)'s definition focus on the implementation activity of the software development process. Since reuse should cut across the software development lifecycle, it is important that we define an all encompassing component. We propose a definition for the software component that is closer to definitions by Szyperski et al. (2002), D'souza and Wills (1999). It accommodates all artifacts of the software development process but emphasizes essential characteristics of a component that enable a component to successfully participate in the process of component assembly to produce a system. Before we give the proposed definition, it is important to note that the definitions above agree that a component:

i) Is a software element. That is, a component contains sequences of abstract program statements that describe computations that are performed by a machine (a von Neumann computing device) Heinman and Councill, (2001). In the past, it was not clear whether requirements specifications and documentation qualified to be software elements especially since they were only informally specified. The current trend towards formal specifications has solved the said challenge. Further still documentation is now a break down of the system reference, system guide, technical reference and technical guide, which can all be provided for as sequences of abstract program statements that describe computations that are performed by a machine.

ii) Exports and imports services. A component requests for services from its environment and in turn provides services to its environment.

iii) Is able to interact and communicate with other components. That is, a component has interfaces which clearly define how it can interact with other components

iv) Is a unit of composition. That is it is designed with the ability of being used as a part of a whole. It is designed to be used in a compositional way with other components.

Definition:

A software component is a reusable artifact of the software development process that provides part of the services that are required to build a software system. The software element should:

(a) Export and/or import services.

(b) Have well defined interfaces

(c) Be a unit of composition.

Examples include: requirements specification, designs, patterns, architectures, test data, test plans, source code and documentation, a component that provides accounting services, a word processing component, a graphical diagram editor, a calculator component e.t.c.

Conclusions And Future Work

Conclusion

We have stated the need to support composition of components from a heterogeneous repository of software components. We have gone on to make a plan for addressing the composition challenge. Before dealing with composition, we have defined a software component. We have noticed (from literature) that there is more effort on reusing artifacts of the implementation activity of the software development process. Even as definitions for a component continue to emerge, the focus is more on composition and deployment. We have thus proposed a definition that shifts emphasis back to the general aim of CBSE which is to reuse artifacts of the whole of the software development process.

Future Work

A component in our context is now clearly defined. However in order to make the SoS a one stop shop, we expect to define composition of our components. Before we define composition, we shall define an abstraction for our components so that they all qualify for exogenous coordination at assembly time. One group of components to pay particular attention to, are the components developed using object-oriented techniques. Object-orientation supports endogenous coordination which we argue is appropriate for intra-component communication but not flexible for inter-component communication (not especially if the components come from different sources). Therefore the next step is to first address abstraction of object-oriented components.

References

Ali F. M. and Du W.(2003). Toward reuse of object-oriented software design models. Information and Software Technology, Vol.46, pp 499-517.

Arbab F. (2005). Abstract Behavior Types: a foundation model for components and their composition.Science of Computer Programming. Vol. 55, pp 3-52.

Arbab F. (2006). Coordination for Component Composition. Electronic Notes in Theoretical Computer Science. Vol. 160, pp 15-40.

Bachmann F., Bass L., Buhman C., Comella-Dorda, S., Long F., Robert J., Seacord and Wallnau K. (2000). Technical

Concepts of Component-based Software Engineering. second edition. Technical Report CMU/SEI-2000-TR- 008, Carnegie Mellon Software Engineering Institute.

Brereton, P. and Budgen, D. (2000). Component-Based System: A classification of Issues. Research Feature, IEEE.

Broy, M. et al. (1998). What characterizes a software component? Software -Concepts and Tools, Vol.19, No.1, pp 49- 59.

Cybulski, J. L., Neal, R. D., Kram, A. and Allen, J.C.(1998). Reuse of early life-cycle artifacts: work products, methods and tools. Annals of Software Engineering, Vol. 5, pp227-251. Springer-Verlag.

D'Souza D.Fand WillsA.C. (1999). Objects, Components and Frameworks with UML. The Catalysis Approach. Addison-Wesley.

Guo J. and Luqi (2000). A survey of Software Reuse Repositories. In:Proceedings of the 7th IEEE Symposium on Engineering of Computer-Based Systems.

Heineman, G., T. and Councill, W., T.(2001). "Component-Based Software Engineering", Putting the Pieces Together. Addison-Wesley.

Hopkins, J. (2000): Component Primer. Communications of the ACM, Vol. 43, No.10, pp 27-30.

Kiziltan, Z. Johnsson, T. and Hnich, B. (2000). On the Definition of Concepts in Component Based Software Development. Report on Component-Based Software Engineering -State of the Art. Malardalen University, Sweden. pp 19-34.

Kung-Kiu L.and Zheng, W.(2005). A taxonomy of software Component Models. In: Proceedings of the 31st EUROMICRO conference on Software Engineering and Advanced Applications (EUROMICRO-SEAA'05). IEEE Computer Society.

Kung-Kiu L. and Zheng, W.(2006). A Survey of Software Component Models, 2nd ed. School of Computer Science, The University of Manchester. Preprint Series CSPP-38.

Lee J., Kim J. and Shin G. (2003). Facilitating Reuse of Software Components using Repository Technology. In: Proceedings of the 10th Asia-Pacific Software Engineering Conference (APSEC'03). IEEE.

Li H. and Van Katwijk, J.(1992). Issues Concerning Software Reuse in the Large. IEEE. LogicLibrary. "Asset Reuse Library." http://colab.cim3.net/cgi-bin/wiki.pl?action brow se&id=AssetReuseLibrary&revision=6 – Accessed July 2006

MacroVista Software. "CodeVistaII." http://www.macrovista.biz/CodeVista/ Intro.asp. – Accessed July 2006

Meyer B.(2003). The grand challenge of trusted components. In: Proceedings of ICSE 2003, IEEE. pp 660-667.

Nierstranz O. and Dami L.(1995). "Component-Oriented Software Technology", Object-Oriented Software Composition, Nierstranz,O and Tsichritzis (Eds.), Prentice Hall, 3-28.

Olson M. and Ogbuji U. (2001) "The Python Web services developer: Web services software repository" http://www- 128.ibm.com/developerworks/ webservices/library/ws-pyth4/ - Accessed May 2006

Pyne R.A. and Mugisa E.K (2004).Essential Elements of a Component-Based Development Environment for the Software Supermarket. In: Proceedings of the IEEE SouthEastcon, March 2004, Greensboro, North Carolina, 173-180, IEEE 2004.

Rubin K. S. (1990). Reuse in Software Engineering: An Object-oriented Perspective. IEEE Sommerville I.(2004).

"Software Engineering". Seventh edition , Addison-Wesley.

SourceBook. "Total Visual SourceBook." http://www.fmsinc.com/products/sourcebook/. – Accessed July 2006

SourceForge. DocumentA01 –What is SourceForge.net? http://sourceforge.net/docs/ about. – Accessed July 2006

Sparling M. (2000). Lessons learned through six years of component-based development. Communications of the ACM, ` Vol. 43, No.10.

SPR software repository. http://www.sprweb.org/repository/index.html. - Accessed May 2006.

Szyperski C. Gruntz D. and Murer S.(2002).Component Software: Beyond Object-Programming. second edition, Addison-Wesley,.

Tarvar K. "Air Force Defense Software Repository System." http://www.stsc.hill.af.mil/ crosstalk/1994/05/xt94d05i.asp. – Accessed July 2006

The NetBeans Metadata Repository. http://www.netbeans.org/about/press/8.html - Accessed May 2006.

36
A User-Centered Approach for Testing Spreadsheets

Yirsaw Ayalew

Spreadsheets are a special form of computer program, which are widely used in areas such as accounting, finance, business management, science and engineering. The wide use of spreadsheets can be attributed to the fact that they appeal to end-user programmers because they are easy to use and require no formal training on designing and programming techniques. However, as the literature indicates, a significant proportion of spreadsheets contain faults. This paper presents an approach for checking spreadsheets on the premises that their developers are not software professionals. It takes inherent characteristics of spreadsheets as well as the conceptual models of spreadsheet programmers into account and incorporates ideas from symbolic testing and interval analysis. To evaluate the methodology proposed, a prototype-tool preserving the look-and-feel spreadsheets developers are accustomed to has been developed.

Introduction

In business, many important decisions are based on the results of spreadsheet computations. Spreadsheet systems are also used for a variety of other important tasks such as mathematical modeling, scientific computation, tabular and graphical data presentation, data analysis and decision-making. The computational model of spreadsheets is even adapted to areas such as information visualization (Chi 1999), concurrent computation (Yoder and Cohn 1993, 1994), or user interface specifications (Hudson 1994) to name just a few. There was even a proposal to use the spreadsheet model as a general model for end-user programming (Nardi and Miller 1990).

In light of this popularity, it seems surprising that a significant proportion of spreadsheets have severe quality problems. In recent years, there has been increased awareness of the potential impact of faulty spreadsheets on business and decision-making. A number of experimental studies and field audits (Brown and Gould 1987; Panko and Halverson 1996; Panko and Sprague 1998; Panko 1998, 1999; Panko 2000) have shown the serious impact spreadsheet errors have on business and other spreadsheet-based decisions. In (Clermont et al. 2002), it is indicated that even in environments where great care is taken for the correctness of "numbers", spreadsheet quality is assessed on a value-based surface structure only. The deeper layer, i.e., the formula structure, is not sufficiently considered. Hence, users show undue reliance on accidental correctness of specific spreadsheet instances, instances that might even after moderate evolution show wrong and highly misleading results.

In trying to address the issue of spreadsheet quality, one has to recognize that spreadsheets are programming artifacts developed by application experts who are not professional programmers. The perceived simplicity of writing simple spreadsheets leads to write sheets of a size and/or complexity where formal quality assurance would be needed. However, conventional testing, as used with conventional software, is – for good reasons – not part of spreadsheet development methodologies. One might even question - whether there is any methodology comparable to classical software development methodologies, where a mix of methods and techniques for testing consumes a substantial portion of the overall life-cycle cost (Beizer 1990; Myers 1979; Sommerville 1992).

To address this issue, there are some attempts tackling the problem from different perspectives. Faults, which are not prevented during the earlier stages of development, can only be identified through detective mechanisms. Since with spreadsheet development, no well defined and user accepted development process can be postulated, many of the research works focus on detective approaches looking for (symptoms of) faults. This can be achieved through visualization and testing tools. Visualization approaches (Chen and Chan 2000; Clermont et al. 2002; Clermont 2003; Davis 1996; Igarashi et al. 1998; Mittermeir et al. 2000; Sajaniemi 2000; Shiozawa et al. 1999) are used to examine the validity of spreadsheet programs by tracing the interactions among cells to uncover unintended or missing connections. They indicate irregularities or mismatches between the physical structure and the invisible dataflow structure. However, visualization approaches do not focus on faults within the formulas; they rather try to highlight anomalies by investigating relationships among cells showing signs of potential errors. To our knowledge there have been only few attempts so far to tackle the problem of spreadsheet quality from a testing perspective. The most accessible systematic testing approach is the one proposed by (Rothermel et al. 1998; Rothermel et al. 2001). They introduced the idea of the *all-uses dataflow adequacy* criterion to spreadsheets.

This paper presents an approach for checking spreadsheets based on the premises that their developers are not software professionals. The approach takes inherent characteristics of spreadsheets as well as the conceptual models of spreadsheet programmers into account. In the broader context of this research, we distinguish between spreadsheet developers having rather a spatial model in mind – considering the sheet as an arrangement of values (and hence the formulas for computing them) in a plane – and developers having rather a numerical model in mind. The latter are assumed to consider a sheet as a network of formulas to compute certain (mutually related) values. This paper focuses on the second category of users. Hence, we assume this category of spreadsheet developers/users builds their sheet based on some numerical domain model of the problem to be solved. Thus, this category of users could at least for critical parts of the sheet indicate the ranges into which expected values of given cells might fall. These ranges are used for interval-calculations as explained in the sections to follow.

Conventional Vs. Spreadsheet Programs

To assure not only theoretical feasibility but also practical applicability of a quality assurance methodology, it is important to consider the similarities and differences between spreadsheets and conventional software.

(Rothermel et al. 1998), identified *order of evaluation, interactivity*, and *expertise of users* as key-differences between spreadsheets and procedural programs. We noted some further aspects, which may have impact on testing spreadsheets.

- *Structure of code:* The structure of a spreadsheet is two-dimensional while the structure of a procedural code is represented linearly. In spreadsheet programming, placement of code is guided by the tabular layout of the results. Hence, the geometrical layout plays a major role.
- *Separation of input/program/output*: In procedural programs, there is a clear separation between input, program, and output. Spreadsheets, on the contrary, at least from the user's point of view, do not have this explicit separation. A sheet's cells contain both the input and the program (formula) while their visible part constitutes the spreadsheet's output. The main part of the spreadsheet program, its formulas, is hidden below the surface. Thus, the computational structure is not readily available to the user
- *Declarativeness:* The ease of use of spreadsheet languages rests on their declarative nature (Miller 1990). The value of a cell is computed by the formula associated with the cell. The detailed procedure of computation is transferred to the system and hence no longer the programmer's assumed responsibility. Formulas describe relations, which specify what is to be computed. Users' understanding of the dependencies is at a higher semantic level.

Since spreadsheet writers would never classify themselves as programmers and since most of them might not even consider the artifacts they produced to be programs, they are not ready to listen to conventional teachings of software development methodology. As application experts, however, they have a gut feeling as to what a reasonable result of their sheets might be – at least for certain critical portions of the sheet. If this "gut feeling" can be systematically made known to the sheet, a tool can check it for consistency not only with computed values but also with the system of formulas constituting the better part of the sheet, a kind of indirect cross-validation can be obtained that will improve the overall quality of the sheet. The next section will show how this can be done systematically.

Interval-based Spreadsheet Testing

Spreadsheets shield users from low-level details of traditional programming. They allow users to think in terms of tabular layouts of adequately arranged and textually designated numbers, with formulas being the crucial constitutive element of spreadsheet programs. These formulas are mathematical expressions containing

references to the values of other cells. It is fair to assume that most spreadsheet errors are introduced while creating formulas. Indeed, studies in (Brown and Gould 1987; Chadwick et al. 2000; Saariluoma and Sajaniemi 1994) have shown that the number of errors in formulas is higher than other errors found in spreadsheets. The proposed interval-analysis focuses specifically on the correctness of formulas in numerical computations based on previous work in (Ayalew et al. 2000; Ayalew 2001; Ayalew and Mittermeir 2002, 2003).

Rationale: A testing methodology for spreadsheets should take into consideration at least three points: It should be *user-centered*; it has to be *spreadsheet-based*; and it should take care of *the situational characteristics of spreadsheets*.

- A *user-centered* approach refers to a methodology, which takes into account both the expertise of users and the conceptual models users have in mind about their programs. Users just specify formulas to establish dependencies so that values of other cells can be used in a computation. Users of spreadsheet systems often do not know the far-reaching dependencies (full data-flow) relating computations in their sheet. Hence, they will not base their testing on the details of interrelated computations. The declarative nature of spreadsheet languages allows them to reason about their program just on the basis of short sections (adjacencies) in the full data-flow network connecting the sheet implicitly. Nevertheless, these (movable) limited windows of awareness allow them to state bounds on the expected values of outcomes of elementary computations based on the intended functionality of a given formula rather than on the structure of the sheet.
- A *spreadsheet-based* approach has to consider the coupling of input/program/output. In conventional testing, code and data are distinct entities. Hence, varying data has no effect on the integrity of the code. Since in spreadsheets code (formulas) and data are tightly interwoven, it is hard to assure that testers varying input are not sub-consciously also destroying figurative numeric constants or even full-fledged formulas. An approach assessing the sheet without directly manipulating the cells themselves would be preferable.
- *Situational characteristics of spreadsheets*: Users prefer spreadsheet systems because they are easy to use and enable them to develop working applications rapidly. Thus, they may not have the time and patience to check their programs by varying inputs (i.e., in essence generating test cases) for each formula cell in all environmental situations this cell might be evaluated. They might appreciate possibilities of validating groups of cells together, if this block of cells follows a common pattern. Assuming that the user/developer of the sheet is an application expert knowing "his/her" numbers, it is fair to ask him/her, what can be expected as reasonable results (expected minima and maxima), even if not all intermediate figures

can be correctly assessed. However, with this implicit arithmetic model, an evaluation of the sheet can be performed.

Based on these considerations, we refrain from asking users to specify voluminous test suites. Instead, spreadsheet developers are asked to indicate the expected range of values a particular cell might assume. These intervals need not necessarily be given for each cell. They have to be given for critical cells though, i.e. for cells serving as sources in the (implicit) data flow – per default, one might say, for cells assuming the role of input cells – and for cells that are considered to be result cells, i.e. cells containing values that would be printed output with conventional programs or cells leading to important intermediate results.

Model: The previous section established that spreadsheets could be seen on several levels. For the purpose of this exposition, we need three levels:

- *Value level:* The values actually presented on the fully displayed sheet as desired by the application expert.
- *Formula level:* The formulas residing in those cells that are neither meant to be pure input cells nor void. For the sake of consistency, constants and input-cells are treated as formula containing just this value as constant.
- *Data-flow level:* It represents the network-structure of data-flow dependencies expressed by the references in the individual formulas. The computational requirements of spreadsheets ensure that this network has the form of a lattice for each well-defined sheet.

To accomplish our quality assurance goals, we add a fourth layer:

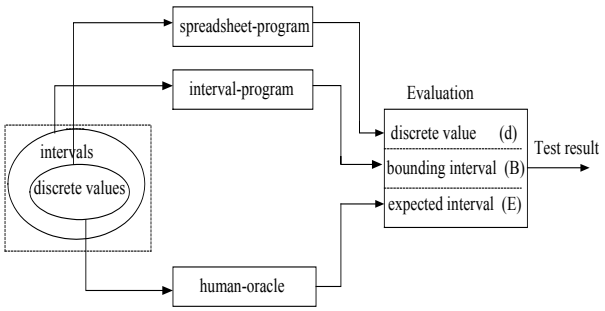
- *Interval level:* Here the user can specify for desired cells the range of expected values.

This additional level can be interpreted in two ways and thus opens up the opportunity to perceive the sheet from three perspectives (see Fig. 1).

- The actual sheet of *discrete values* computed by the regular *spreadsheet program* the user has written. This sheet contains the results of the computations the user wanted to have.
- The *intervals* in intermediate or source cells, interpreted as range specified by the *human oracle*.
- The *interval program* resulting from evaluating the formulas of the individual cells not by means of conventional arithmetic but by means of interval arithmetic over all those subsections of the data-flow network, where, due to interval-input in the source, interval computations can be performed. (Note that constants are treated as intervals of length 0). As will be shown later this allows not only to check numerical results for their reasonableness, it also allows to check whether the users' expectations are consistent among themselves and consistent with respect to the model of the application expressed by the various interrelated spreadsheet formulas.

The reader will note that this design, shown in Fig. 1, has little in common with the conventional notion of repeatedly executing a piece of software on a carefully chosen test suite. It allows rather execution of a program on two levels: the level of user data and the level of quality assurance data in the form of intervals. Quality assessment is done by assessing the consistency between actual computation, interval computation, and directly voiced expectations of result values given by the application expert.

Fig 1: Spreadsheet program test process



Comparing this to established testing approaches, symbolic testing seems closest. Interval-testing can be considered as a special case of symbolic testing. Symbolic testing is used to verify a formula (i.e., symbolic output) by expressing the formula only in terms of input variables instead of using actual values. In contrast to symbolic testing, interval-based testing uses intermediate variables (cells taken as variables) during the interval computation. Another difference is that symbolic testing assumes any type-correct value for the input variables while in this approach the values of input variables are restricted to a range of reasonable values indicated by the tester/user. The intervals specified for cells on intermediate positions in the data-flow might further narrow down the computed interval.

Spreadsheet Testing Process: Interval-based spreadsheet testing is a comparison of the users' computational model (goal) with the actual spreadsheet computation. Users usually have a computational model derived from the domain knowledge of the application they try to solve. Finally, the goal of computation which is generated by domain knowledge will be described in terms of language constructs (plans) of some programming tool (in this case a spreadsheet system) which helps to achieve the specified goal. Errors occur when the language constructs chosen do not match the desired goal or model of computation. Goal and plan as they exist in conventional programming are also used in spreadsheet systems. (Sajaniemi et al. 1999) conducted an experimental study and found that spreadsheet users have a set of basic programming goals and plans describing spreadsheet-programming knowledge. A detailed discussion of goals and plans in spreadsheet systems can be found in (Tukiainen and Sajaniemi 1996; Tukiainen 2000).

In the interval-based model, the intervals specified by the application expert might be considered as goal, while the formulas are the expressions of the plan. However, since each operator in these formulas can be interpreted with conventional point arithmetic semantic as well as with interval arithmetic semantic, two versions of the plan exist. Both plans can be executed and this execution offers three chances of comparison (cf. Fig. 1). The discrete values **d**, resulting from executing the formulas with conventional point semantic, have to be within both, the intervals expected by the human oracle (application expert), **E**, as well as within the intervals resulting from executing the sheet with interval semantics, **B**. Since interval arithmetic yields the overall minimum and maximum, the bounding intervals **B**, resulting from the execution with interval semantics, have to contain the intervals **E** expected by the human oracle. Thus, one assumes that the expert's deeper insight in the specific application semantic should lead to narrower ranges. If one of these comparisons fails, there is either a mismatch between the conceptual model and its realization in the formulas written into the cells or there is a mismatch in the expectations as to what ranges might actually be covered by the sheet. Either of these indicates a symptom of inconsistency or of a fault.

Thus, the comparison $\mathbf{d} \in \mathbf{E} \wedge \mathbf{E} \subseteq \mathbf{B}$ reports no symptom of fault, whereas $\mathbf{d} \notin \mathbf{E} \vee \mathbf{E} \not\subset \mathbf{B}$ indicates a symptom of fault. Currently, we are investigating the issue of decision making when there is some intersection between E and B but one is not a subset of the other. For those formula cells for which a symptom of fault is generated, further investigation is carried out using a fault tracing strategy. A discussion of the fault tracing strategy can be found in (Ayalew and Mittermeir 2003).

To summarize, one might note that these considerations have the following assumptions and consequences.

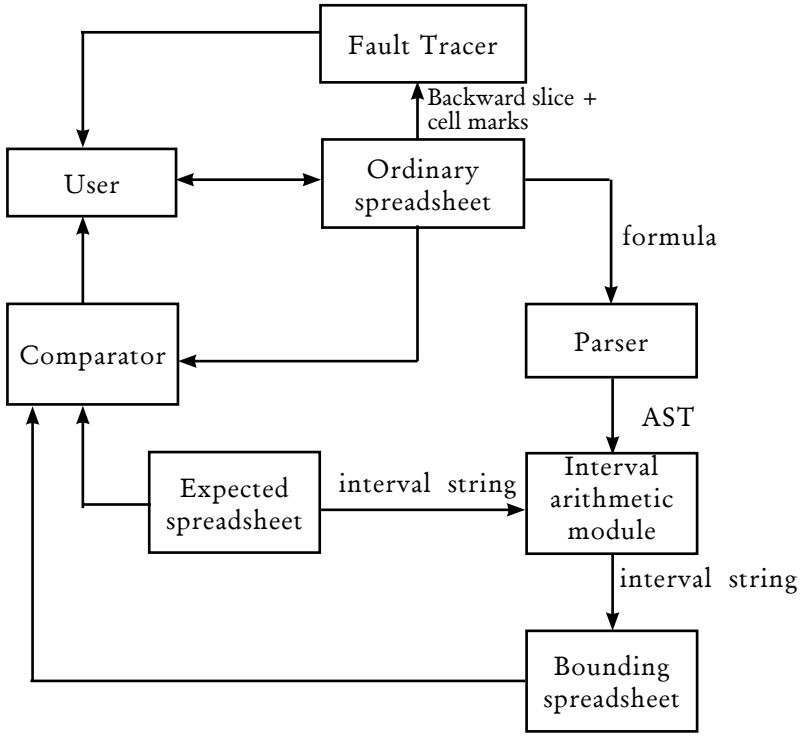
- Users do not have in mind exact values about the results of formulas, but they have some range of reasonable values, which can be expressed as intervals.
- Quality assessment can be conducted without touching any value of the original spreadsheet.
- During the comparison of spreadsheet computation, user expectation, and interval computation, boundary violations are only indicators of symptoms of faults.
- Errors within the boundaries of expected intervals can go unnoticed but might not harm dramatically.
- Ripple effects might lead to noticeable boundary violations.

System Architecture

In order to demonstrate the effectiveness of the interval-based testing methodology, a prototype tool was built. The tool incorporates a parser, an interval arithmetic

module, a comparator, and a fault tracer. Lead by its widespread use, we chose MS-Excel as pilot environment for demonstrating interval-based testing. For integration of this add-on, MS-Excel's object model was used. The extensions were coded in Visual Basic. Fig. 2 depicts the general architecture chosen to integrate the interval-based testing prototype on top of MS Excel. This architecture indicates the dataflow between the different components of the system.

Fig 2: Architecture of the interval-based testing

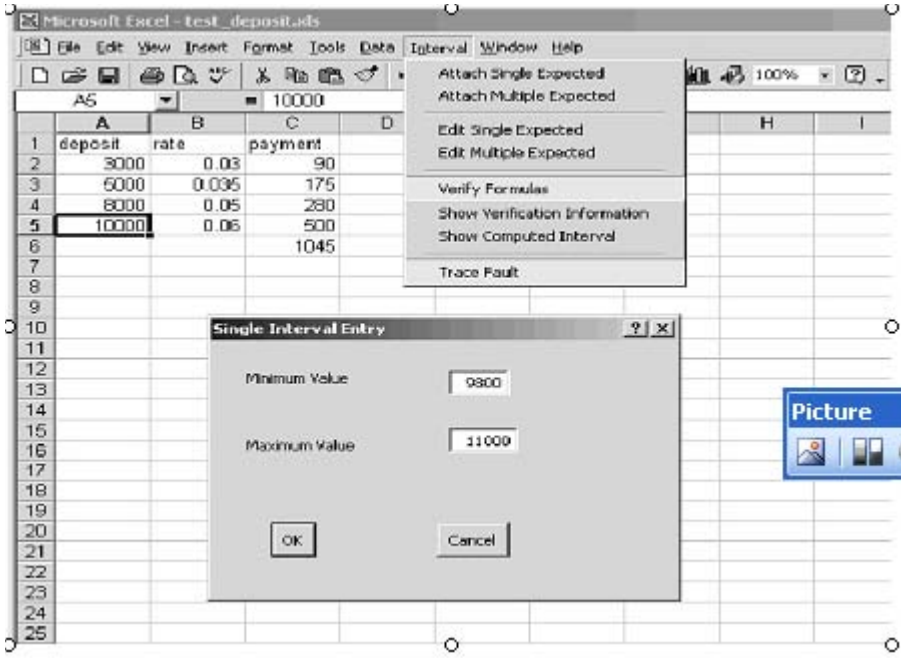


As can be seen from Fig. 2, the user interacts only with the ordinary spreadsheet. In other words, development of a spreadsheet program is carried out in the usual way. When the user wants to attach an interval to a given cell, the user selects the cell and chooses a command from the **Interval** menu through which interval-based testing features are available. The attached interval is then stored in the expected spreadsheet as a string with the same coordinate as the corresponding cell selected in the ordinary spreadsheet (see Fig. 3). This can be done also for a group of cells when they are intended to have the same interval.

Whenever a user is interested to verify a particular formula cell or group of formula cells, the desired cells are selected and a **Verify Formulas** command is issued. Subsequently, the parser analyzes each formula and generates the abstract syntax tree. Then, the formulas are evaluated based on interval arithmetic semantic of the operators used. Finally, the resulting value of the interval computation, i.e., the bounding interval for the respective cell, is stored in the bounding

spreadsheet as a string since Excel neither supports interval data types nor allows user-defined data types. Once the necessary values are available from the three sources, i.e. spreadsheet computation, user expectation, and interval computation, the comparator determines the existence of symptoms of faults. Among those cells with symptoms of faults that contribute to a given faulty cell, the fault tracer identifies the most influential faulty cell.

Fig 3: Attaching an interval to a numeric cell



In a typical spreadsheet system, the majority of formulas are non-conditional. In those cases, applying interval arithmetic is straightforward. In conditional formulas, different computations are performed based on the computed truth-value of a condition. This implies that the result of executing a cell holding a conditional is not a single interval but a pair of intervals and the user is expected to attach intervals corresponding to each branch of the decision that will be executed. For example, for the formula IF(A1 > 5, A1*B1, A1-B1), the expectations are different based on whether the comparison A1 > 5 is true or false. For nested IFs the same procedure is applied. However, the representation becomes a set of intervals. To generalize the number of expected intervals, which are needed for nested IFs, let N be the number of nested IFs used in a formula. Then, N + 1 expected intervals are required. However, for a formula which involves N IFs without being nested (e.g. IF(...) + IF(...)), 2^N expected intervals are needed. Since deep nesting of alternatives is rather exceptional, the current implementation supports only ordinary conditionals but not nested or multiple IFs.

Discussion

Some initial experiments with the prototype showed that this approach is promising especially for cases where sheets go beyond a threshold of structural complexity while remaining (at least in the respective portion) small enough so that the users' conceptual numeric model comes into play. An important observation in those cases was that actual deviations of bounding and expected intervals or of discrete values falling outside of the permissible interval were only part of the cases where faults were identified. Of about the same importance was the fact that spreadsheet-writers saw the need to systematically walk through their sheets and reflect on what actual permissible intervals could be. Doing so, they themselves made some top-of-the head interval calculations and doing so was in many cases sufficient to identify a problem in the sheet itself. Thus, we conclude that although we would not dare to recommend spreadsheet-writers a formal review of their work, the proposed methodology induces them to perform such a review in a systematic way along various dataflow-branches through the sheet.

Another aspect worth reporting is the way users define their expected intervals varies. Some are using their domain knowledge and propose the tight bounds expected by the requirement $E \subseteq B$. Others, however, reason with "easy numbers", i.e. with those suitable for doing mental calculations. This might lead to situations where E yields wider intervals than B and the system will point out this inconsistency. Situations like these require some additional explanations. However, even these users realize the value of the exercise since the reported inconsistency between the expected and computed intervals allows them to reflect again on what they have actually specified by the respective formula.

Conclusion

The rationale for interval-based testing is that testing approaches for spreadsheets should take into account the inherent characteristics of spreadsheets as well as the conceptual models of spreadsheet developers. Therefore, interval-based testing is proposed based on the premises that spreadsheets are mainly used for numerical computations and spreadsheet developers are end-users who are not expected to follow the formal process of software development.

Interval-based testing tries to establish a connection between the users' numeric expectations relative to individual cells and how these expectations lead to intervals of potential values. Discrepancy between computed intervals and expected intervals will inform the user about divergences between the conceptual model and the specified model. Despite the requirement to attach intervals, the proposed approach does not require any knowledge of conventional software testing. As such, it is a user-centered approach. Further, the interval-definition phase requires the user to check and think about the functionality of the particular formula cell once again. Hence, this phase also serves as a kind of manual review process giving the methodology a combined feature of automatic testing and manual reviews.

References

Ayalew, Y. (2001), 'Spreadsheet Testing Using Interval Analysis', (Klagenfurt University).

Ayalew, Y. and Mittermeir, R. (2002), 'Interval-based Testing for Spreadsheets', *International Arab Conference on Information Technology* (University of Qatar, Doha), 414 – 22.

--- (2003), 'Spreadsheet Debugging', *Fourth International EuSpRIG Conference* (Trinity College, Dublin, Ireland), 67 – 79.

Ayalew, Y., Clermont, M., and Mittermeir, R. (2000), 'Detecting Errors in Spreadsheets', *EuSpRIG Symposium: Spreadsheet Risks, Audit & Development Methods* (London: University of Greenwich), 51-62.

Beizer, B. (1990), *Software Testing Techniques* (2 edn.; New York: Van Nostrand Reinhold).

Brown, P. and Gould, J. (1987), 'An Experimental Study of People Creating Spreadsheets', *ACM Transactions on Office Information Systems,* 5 (3), 258-72.

Chadwick, D., Knight, B., and Rajalingham, K. (2000), 'Quality Control in Spreadsheets: A Visual Approach using Color Codings to Reduce Errors in Formula', . *8th International Conference on Software Quality Management*.

Chen, Y. and Chan, H. (2000), 'Visual Checking of Spreadsheet', *EuSpRIG 2000 Symposium: Spreadsheet Risks, Audit & Development Methods* (London), 75 – 85.

Chi, E. (1999), 'A Framework for Information Visualization Spreadsheets', (University of Minnesota).

Clermont, M. (2003), 'A Scalable Approach to Spreadsheet Visualization', (Klagenfurt University).

Clermont, M., Hanin, C., and Mittermeir, R. (2002), 'A Spreadsheet Auditing Tool Evaluated in an Industrial Context', *EuSpRIG*, 35 - 46.

Davis, J. (1996), 'Tools for Spreadsheet Auditing', *International Journal of Human-Computer Studies,* 45 (4), 429-42.

Hudson, S. (1994), 'User Interface Specification Using an Enhanced Spreadsheet Model', *ACM Transactions on Graphics*, 209 – 39.

Igarashi, T., et al. (1998), 'Fluid Visualization of Spreadsheet Structures', *IEEE Symposium on Visual Languages*, 118 – 25

Miller, R. (1990), *Computer-Aided Financial Analysis* (Addison-Wesley).

Mittermeir, R., Clermont, M., and Ayalew, Y. (2000), 'User-Centered Approaches for Improving Spreadsheet Quality', (Klagenfurt: Klagenfurt University), 12.

Myers, G.J. (1979), *The Art of Software Testing* (Wiley-Interscience).

Nardi, B. and Miller, J. (1990), 'The Spreadsheet Interface: A Basis for End User Programming', (Hewlett-Packard Software Technology Laboratory).

Panko, R. R. (2000), 'Two Corpuses of Spreadsheet Errors', *33rd Hawaii International Conference on System Sciences*.

Panko, R. R. and Halverson, R. P. (1996), 'Spreadsheets on Trial: A Survey of Research on Spreadsheet Risks', *29th Hawaii International Conference on System Sciences*, 326-35.

Panko, R. R. and Sprague, R. (1998), 'Hitting the wall: Errors in developing and code inspecting a 'simple' spreadsheet model', *Decision Support Systems,* 22 (4), 337-53.

Panko, R.R. (1998), 'What We Know About Spreadsheet Errors', *Journal of End User Computing: Special Issue on Scaling up End User Development,* 10 (2), 15-21.

--- (1999), 'Applying Code Inspection to Spreadsheet Testing', *Journal of Management Information Systems,* 16 (2), 159-76.

Rothermel, G., et al. (1998), 'What You See Is What You Test: A Methodology for Testing Form-based Visual programs', *20th International Conference on Software Engineering*, 198-207.

--- (2001), 'A Methodology for Testing Spreadsheets', *ACM Transactions on Software Engineering and Methodology,* 10 (1), 110 – 47.

Saariluoma, P. and Sajaniemi, J. (1994), 'Transforming Verbal Descriptions into Mathematical Formulas in Spreadsheet Calculation', *International Journal of Human-Computer Studies,* 41 (6), 915-48.

Sajaniemi, J. (2000), 'Modeling Spreadsheet Audit: A Rigorous Approach to Automatic Visualization', *Journal of Visual Languages and Computing,* 11 (1), 49-82.

Sajaniemi, J., Tukiainen, M., and Vaisanen, J. (1999), 'Goals and Plans In Spreadsheet Calculation', (Department of Computer Science, University of Joensuu).

Shiozawa, H., Okada, K., and Matsushita, Y. (1999), '3D Interactive Visualization for Inter-Cell Dependencies of Spreadsheets', *IEEE Symposium on Information Visualization*, 79-82.

Sommerville, I. (1992), *Software Engineering* (4th edn.: Addison-Wesley).

Tukiainen, M. (2000), 'Uncovering Effects of Programming Paradigms: Errors in Two Spreadsheet Systems', *12th Workshop of the Psychology of Programming Interest Group*, 247-66.

Tukiainen, M. and Sajaniemi, J. (1996), 'Spreadsheet Goal and Plan Catalog: Additive and Multiplicative Computational Goals and Plans in Spreadsheet Calculation', (Department of Computer Science, University of Joensuu, Finland).

Yoder, A. and Cohn, D. (1993), 'Architectural Issues in Spreadsheet Languages', *International Conference on Programming Languages and System Architectures*.

--- (1994), 'Real Spreadsheets for Real Programmers', *International Conference on Computer Languages*, 20-30.

Part 6

Sustainable Development

37
ICT as an Engine for Uganda's Economic Growth: The Role of and Opportunities for Makerere University

Venansius Baryamureeba,

The use of Information and Communications Technologies (ICT) to improve how goods are produced and services are delivered is a feature of everyday life in developed countries. If ICT is used appropriately, it has the potential to vastly improve productivity. Thus the issue for developing and least developed countries is how best to use ICT to achieve development objectives, given the operating constraints in these countries. The constraints are mainly lack of infrastructure and human capacity. It is now a known fact that ICT infrastructure readiness without adequate skilled ICT human capacity cannot lead to economic growth. In this paper we discuss the role of Makerere University and suggest opportunities for Makerere University in this area of ICT led –economic growth of Uganda.

1. Introduction

Information and communications technologies, broadly defined, facilitate by electronic means the creation, storage, management and dissemination of information. The emphasis in this paper is on both ICT as a vehicle for communication and as a means of processing information. The communication vehicles range from: radio (analogue, digital and high frequency two-way), television, telephone, fax, computers and the Internet. Newspapers are also included; as they also often now have an electronic form on the World Wide Web.

The old types of ICT i.e. the newspapers and as well as radio and television have the advantages of low cost, requiring little skill to operate and the potential to be highly relevant to the needs of the users in terms of local information delivered in local languages. Their downsides are to do with the often one-sided nature of the communication and potential for censure by governments.

The new, more advanced forms of ICT include networked computers, satellite-sourced communication, wireless technology and the Internet. A feature of these technologies is their capacity to be networked and interlinked to form a 'massive infrastructure of interconnected telephone services, standardized computing hardware, the Internet, radio and television, which reach into every corner of the globe'. Four interconnected characteristics of the new, advanced ICTs are worth noting (Curtain, 2004)[1]. The first is their capacity for interactivity: the

new forms of ICTs offer effective two-way communication on a one-to-one or one-to-many basis. Second, the new ICTs are available 24 hours a day on real time, synchronous or delayed, asynchronous basis. Third, ICT through its interconnected infrastructure now has a reach over geographic distances not possible even in the recent past. The fourth feature of the new ICT that is also highly significant is the continuing reduction in the relative costs of communicating, although this differs by location.

In Uganda there is a high incidence of radio ownership in low-income communities, which indicates that it is a low cost communications technology that many people can afford. There is also a relatively high incidence of mobile phone ownership in low-income communities in Uganda. TV is also important as a means of communication where people of low-incomes have access to electricity. Community ICT facilities such as community radio, and community television exist and should be encouraged as they play a significant role in preserving and providing access to cultural information and other resources. They can promote the traditions and heritage of ethnic and marginalized groups and help to keep their language, indigenous knowledge and way of life and livelihood alive and active.

The key factors responsible for the different ICT take-up rates in Africa are: per capita income, language, levels of education (illiteracy), internal digital divide within the African continent, restrictive regulatory framework, poverty and the lack of infrastructure and the rural concentration and dispersed nature of a country's population. In general, the lower a country's per capita income, the less likely its population is to have access to both old and new information and communication technologies (Curtain, 2004) [1].

There are trade offs for low-income countries in terms of devoting scarce resources to ICT and therefore there is a need to identify which kinds of ICT access deliver the best value for money in developing countries, and how the limited resources that can be spent on it can be made to best suit the particular needs of the poor (Caspary, 2002) [2]. Maximizing the use from ICT for developing countries requires an understanding not only of the opportunities ICT present, but also of the trade-offs involved – and of the particular ways in which ICT access has to be tailored if any developmental benefits are to be reaped.

Universities should provide a vision, strategy and an enabling environment that promotes the use of information and communications technologies (ICT) in universities in particular and society in general. Through access to information and freedom of expression, citizens are able to gain civic competence, air their views, engage in discussions and deliberations, and learn from one another, all of which provides the citizen with an enlightened understanding of government action. Governance is the way power is exercised in managing a country's economic and social resources for development. ICT present opportunities for African countries to implement e-governance/ e-government. ICT can support transparency, create a public space for citizens as well as offer a readily available

consultation mechanism. The Internet, distance-learning opportunities, online (electronic) learning, computerized library packages and strategic databases must be brought nearer to the isolated and poor African nations unable to integrate their economies and intellects with the powerful and respected community of states.

1.1. Development

Development means improvement of a country's economic and social conditions. More specifically it refers to improvements in ways of managing an area's natural and human resources in order to create wealth and improve people's lives. Development is sustainable when it meets the needs of the present without compromising the ability of future generations to meet their own needs. Thus, sustainable development is defined as maintaining a delicate balance between the human need to improve lifestyles and feeling of well-being on one hand, and preserving natural resources and ecosystems, on which we and future generations depend.

1.1.1. Economic Development

Economic development is a measure of how wealthy a country is and of how this wealth is generated (for example agriculture is considered less economically advanced than banking). Economic development is any effort or undertaking which aids in the growth of the economy. Economic growth is the increase in value of the goods and services produced by an economy. It is conventionally measured as the percentage rate of increase in real gross domestic product. Growth means moving towards wealth, which means the same thing as moving away from non-wealth. Growth is usually calculated in real terms, i.e. inflation-adjusted terms, in order to net out the effect of inflation on the price of goods and services produced.

The relation between economic growth and human capital (skilled labor) is illustrated by the two figures below adopted from Soubbotina and Seram (2003):

Fig 1: Economic growth and human development

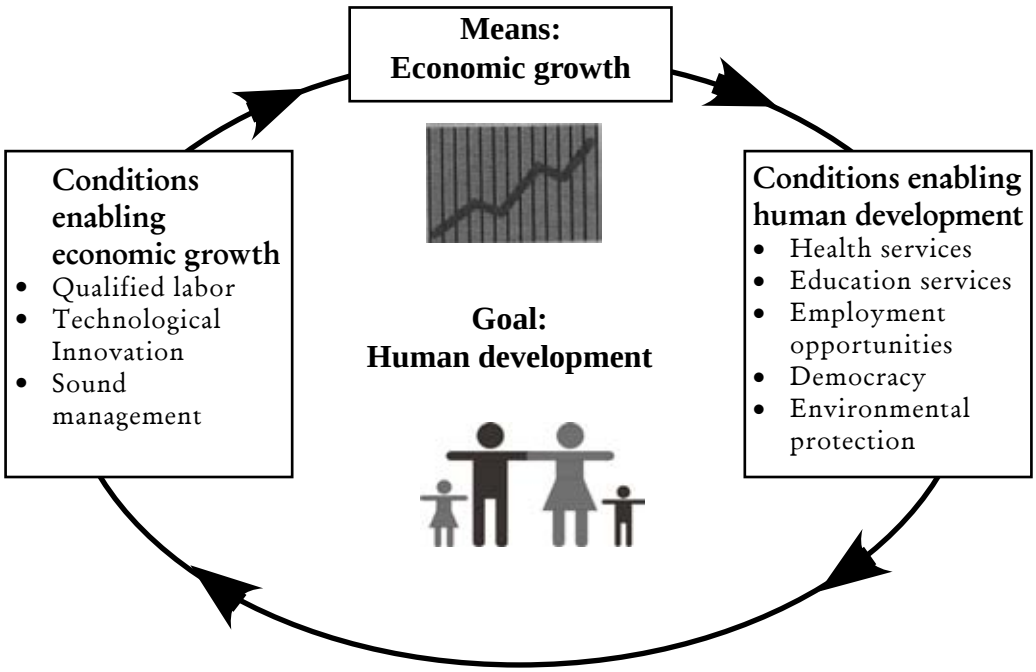
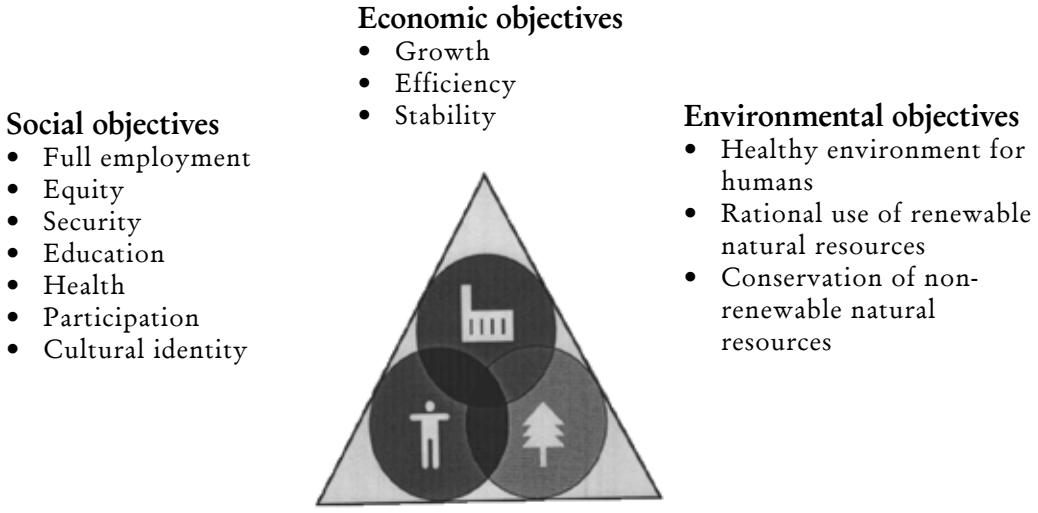


Fig 1: Objectives of sustainable development



Different countries or regions can be grouped under More Economically Developed Countries (MEDCs) and Less Economically Developed Countries (LEDCs). Development can be considered in terms of either economic or human development. Indicators are used to judge a countries level of development. The economic indicators include:

a) Gross Domestic Product (GDP) measures the wealth or income of a country. GDP is the total value of goods and services produced by a country in a year.
b) Gross National Product (GNP) is another measure of a country's wealth or income. GNP measures the total economic output of a country, including earnings from foreign investments which are not included in GDP.
c) GNP per capita is a country's GNP divided by its population. (*Per capita* means *per person*.)
d) Economic growth measures the annual increase in GDP, GNP, GDP per capita, or GNP per capita.
e) Inequality of wealth is an indication of the gap in wealth and income between a country's richest and poorest people. It can be measured in many ways (e.g., the proportion of a country's wealth owned by the richest 10% of the population, compared with the proportion owed by the remaining 90%).
f) Inflation measures how much the prices of goods, services and wages are increasing each year. High inflation (above a few percent) is believed by many to be a bad thing, and suggests a government's lack of control over the economy.
g) Unemployment is measured by the number of people who cannot find work.
h) Economic structure shows how a country's economy is divided between primary, secondary and tertiary industries.
i) Demographics studies population growth and population structure. It compares birth rates to death rates, shows average ages, and compares numbers of people living in towns with numbers living in the countryside. (Many LEDCs have a younger, faster-growing population than MEDCs, with more people living in the countryside than in towns.)

1.1.1. Human Development

Human development measures the extent to which people have access to wealth, jobs, knowledge, nutrition, health, leisure and safety as well as political and cultural freedom. The more material elements in this list such as wealth ad nutrition are often grouped as together under the heading standard of living and the less material elements such as health and leisure under quality of life. Human development indicators measure the non economic aspects of a countries development and include:

a) Life expectancy is the average age to which a person lives.
b) Infant mortality rate counts the number of babies, per 1,000 live births, who die under the age of one year.

c) Poverty indices count the percentage of people living below the poverty level, or on very small incomes (e.g. under £1 per day).

d) Access to basic services measures the availability of services necessary for a healthy life, such as clean water and sanitation.

e) Access to healthcare takes into account statistics such as how many doctors there are for every patient.

f) Risk of disease calculates the percentage of people with dangerous diseases such as AIDs, malaria, tuberculosis, etc.

g) Access to education measures how many people attend primary school, secondary school and higher education.

h) Literacy rate is the percentage of adults who can read and write.

i) Access to technology, includes statistics such as the percentage of people with access to phones, mobile phones, television and the internet.

j) Male/female equality compares statistics such as the literacy rates and employment between the sexes.

k) Government spending priorities compares health and education expenditure with military expenditure and paying off debts.

1.2. Two Approaches to use of ICT for Development

1.2.1. ICT as tool to promote economic growth

This focuses on ICT as a driver of the development process. This usually focuses on providing the poor with opportunities to receive up-to-date information or the ability to communicate more easily or achieve an enhanced ability to communicate with others. The explicit or implicit objective of an ICT-led development project such as Telecentres is often on promoting economic growth through access to better opportunities to generate income to reduce poverty.

The ICT-driven approach is often underpinned by the economic assumption that better information improves how economic resources are allocated. It is a fundamental axiom of orthodox economics that the capacity of an economy to operate efficiently depends on how well markets work. Markets operate through the adjustment of supply and demand of goods and services through prices, which send signals about the balance between these two sides of the equation. In practice, prices do vary widely not only over time but from region to region, particularly where information flows are limited or non-existent (Eggleston et al., 2002)[3].

In poor countries, the coordination of economic activity rarely works well. In isolated rural villages in most developing countries, there are virtually no sources of information regarding market prices and other production related information. For them, 'information is poor, scarce, maldistributed, inefficiently communicated and intensely valued'. ICT and village knowledge centers offer the possibility of improving the life and well-being of rural communities; not only by enhancing markets and generating knowledge-based livelihoods, but also by

furthering healthcare, education, government entitlements, social cohesion and societal reform.

The economic case for the contribution of ICT to the reduction of poverty through economic growth is summarized in Figure 1 below. The postulated relationship between access to ICT and economic growth is spelt out through a five-step process starting at the bottom of the pyramid.

Fig 3: Five-step pyramidal process

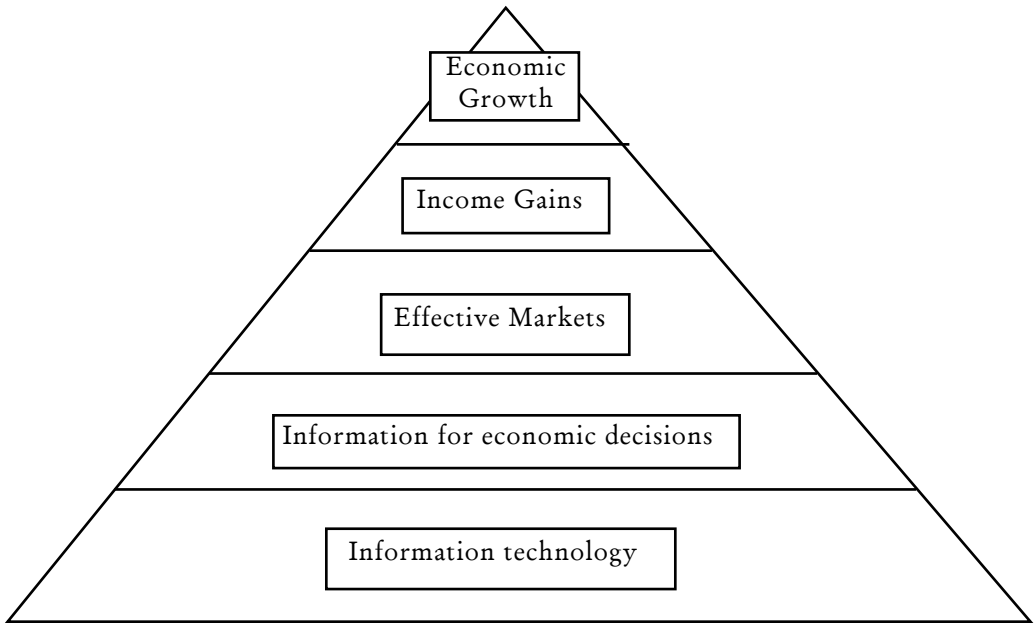


Fig3. Source: Eggleston, K; Jensen, R; and Zeckhauser, R; 2002, 'Information and communication technologies, markets and economic development', in G. Kirkman et al. the Global Information Technology Report: Readiness for the Networked World. Oxford University Press, New York, p 71.

The economic justification for giving the poor better access to ICT is that up-to-date and reliable information about prices and availability of resources can be more easily disseminated to areas where the poor are more likely to be concentrated. The poor receiving the information are then better able, as both producers and consumers, to participate in effective markets (Eggleston et al., 2002)[3]. The immediate consequence should be income gains for participants, and the ability to better spend their incomes. Over the long term, enhanced access to information should enable producers to significantly improve their practices. Such improvement lays the path to economic growth.

The ICT-driven approach to development is more likely to emphasise communication as a good outcome in itself. ICT-based projects such as telecentres offering access to e-mail or the setting up of a web site as a marketing tool are

favoured because they provide better access to markets through current and reliable information on prices, and offer the opportunity to promote goods. There are many publicized stories of how small traders or poor communities in low and middle income countries have gained access to wider markets through the Internet. These range from a small venture in a slum area of Nairobi selling sandals made from car tyres to the United States to use of the web pages to encourage pro poor tourism in Nepal [4]. The poor benefit through increased demand for their products.

1.1.2. Use of ICT in support of development

This focuses on the uses of ICT in a supplementary role in development projects. This approach places a more specific development objective to the fore and seeks to use ICT to support that objective. Here ICT plays a supporting or supplementary role to meeting a primary objective. This approach first clarifies the development goal the project is addressing; works out what the information and communication needs of the target group are and then looks to a cost effective way that ICT and other media can play in managing information and providing channels of communication (Heeks, 2002) [5]. This approach starts with a more multidimensional perspective on poverty reduction, acknowledging the importance of better access to services such as education and health. Access to government services in a transparent way with low transaction costs is another way in which ICT can play a key supporting role in development. The role of the poor themselves in defining their own information needs is a key characteristic of this approach.

2. Role of Makerere University

2.1. Human Capital Development

For any meaningful development to take place, a critical mass of human capital and in this particular case ICT human capital is mandatory. People with ICT skills and knowledge now drive the service industry world over. As of 30th October 2006, the Faculty of Computing and IT (CIT) had a total of 5560 students (5000 undergraduate and 560 postgraduate students including 54 PhD students) distributed in four academic departments of Computer Science, Information Technology, Information Systems and Networks. There are other students undertaking short courses. By 2008 CIT will have about 10,000 diploma and degree students in this area alone. In preparation for high student enrollment in this area, Makerere University is currently putting up the largest computing centre in Africa. Once completed it will accommodate over 12,000 students at any one time. CIT is posed to serve as the number one ICT human capital producer and sustainer not only in Uganda but also in the neighboring regions. The one advantage of the graduates from CIT is that they follow internationally recognized curricula (Computing Curricula 2005, 2005) [7].

The Faculty of Computing and IT at Makerere University has designed and implemented programs in computing disciplines (Baryamureeba, 2007) [6]. For example:

- The programs under the Department of Computer Science include a PhD in Computer Science, M.Sc. in Computer Science, Postgraduate Diploma (PGD) in Computer Science, B.Sc. in Computer Science, Diploma in Computer Science and Information Technology, and Computer Science as a subprogram (as part of the B.Sc. Degree Program) in collaboration with Faculty of Science.
- The programs under the Department of Information Technology include a PhD in Information Technology; Master of Information Technology with options in Information Technology Management, Information Security, Internet and Service Delivery, and Internet and Web Computing; PGD in Information Technology; and Bachelor of Information Technology.
- The programs under the Department of Information Systems include a PhD in Information Systems; M.Sc. in Information Systems with options in Computer Information Systems, Management Information Systems, Information Systems Management, and Internet and Database Systems; PGD in Information Systems; Bachelor of Information Systems (*proposed*) to begin in 2007/2008 academic year.
- The programs under the Department of Networks include a PhD in Software Engineering; M.Sc. in Data Communications & Software Engineering with options in Data Communications Engineering, Network and System Administration, and Software Engineering; PGD in Data Communications & Software Engineering; PGD in ICT Policy and Regulation; is run under Nettel@Africa program [11]; B.Sc. in Computer Engineering (*proposed*) to begin in 2007/2008 academic year; and B.Sc. in Software Engineering (*proposed*) to begin in 2007/2008 academic year. The Department of Networks will soon be split into three departments: Department of Software Engineering, Department of Computer Engineering and Department of Data Communications and Computer Networks.
- The Faculty of Computing and IT also runs several short courses of duration 1-8 months which include Certificate in Computer Applications (CCA); International Computer Driving License (ICDL); Oracle Certified Associate (OCA); Oracle Certified Professional (OCP); Cisco Certified Network Professional (CCNP); Cisco Certified Network Associate (CCNA); IT Essentials I & II; and Microsoft Certification: MOS, MCDBA, MCSE, MCSA, and MCSD.

Many African countries (e.g. South Africa, Ghana, Rwanda, Uganda, Kenya) have developed ICT-led socio-economic development strategies. However, without a critical mass of ICT professionals in these countries these strategies will only remain on paper. It is the critical mass that will lead to several innovations and patents that are a preliquisite for development. Also the critical mass will ensure stable workforce, easy retention of staff and generally affordable skilled human capital. This in the long term will lead to low costs of production, which will attract multi-national companies and lead to high tax collections, as is the case in China and India.

2.2. Research Development

No country or company can grow without investing heavily in research. For example Microsoft Inc. and Cisco Systems Inc budget for millions of dollars for research every year and through their foundations invest a lot of money in private sector-university partnerships. For example Cisco Systems Inc. runs Cisco Systems University Research Program worth millions of dollars and twice every year researchers from Universities around the world compete for research funds in priority areas to Cisco Systems Inc. research strategy. Countries like USA, Canada, Finland and China to a name a few, are investing heavily in research. For example, in 2005 China accelerated Science and Technology (S&T) expenditure by 25% increase compared to 2004, at the same time Research and Development (R&D) expenditure alone grew by 20% and R&D workforce surpassed a million. Makerere University Faculty of Computing and IT is posed to spearhead research in Computing and ICT since its advantaged by high graduate student population, hosts both an Annual International Conference on Computing and ICT Research and an International Journal of Computing and ICT Research (Baryamureeba, 2007) [6]. The cutting edge (multidisciplinary) research will lead to innovations/ patents, which could foster private sector competitiveness in the long run.

2.3. Contribute towards a literate community

Most of the rural communities have a wealth of indigenous knowledge but because they cannot read and write they are normally termed as illiterate. Makerere University will work towards a literate society especially in ICT by customizing search engines and operating systems in local languages and designing user-friendly computer interfaces. This will strengthen the community outreach function of the university. In addition Makerere University will ensure that all graduates leaving the University are computer literate and have both basic and advanced ICT skills.

2.4. Local content development.

The area of generation of local content in Uganda is still untapped. There is need to generate local content that can be put on the web and also used in e-learning. There is need to preserve and disseminate indigenous knowledge through creation of online databases.

2.5. Incubation and Innovation Centre

Makerere University Faculty of Computing and IT is one of the largest computing faculties in Africa. Universities have and will always serve as the best and cheapest incubation centres for many products. Makerere University in partnership with the Government of Uganda is taking advantage of the graduates from this faculty and other faculties to incubate and rollout ICT outsourcing in areas like data and call centres, software development and customization, and e-service delivery. The faculty is about to complete a 12,000 sq metres computing building (Baryamureeba,

2007) [6] that will provide space for the centre activities. Former students will be given a fully equipped computing facility of at least 1000 computers within this building to explore ideas and incubate them under the supervision of Makerere University staff and later start spin off companies. Some of the advisors to the students will come from the Department of Software Development and Innovations, an autonomous unit with in the Faculty of Computing and IT and ICT Consults Ltd, the consulting arm of the Faculty of Computing and IT at Makerere University (Baryamureeba, 2007) [6]. The Faculty of Computing and IT will also partner with the private sector on some of the initiatives and this is expected to facilitate two-way knowledge transfer. It will also open up an avenue for development of commercial products from the prototypes developed by the students. Transformation of staff and student research and prototypes into commercial products adds value to research and also provides an opportunity for the University to generate income.

3. Key Opportunities For Makerere University

3.1. Software Industry

The term East Asian Tigers refers to the economies of Hong Kong, Taiwan, Singapore, and South Korea. They are also known as Asia's Four Little Dragons. These countries and territories were noted for maintaining high growth rates and rapid industrialization between the early 1960s and 1990s. The growth of these economies was mainly from exports. In the early 21st century, with the original four Tigers at or near to fully developed status, attention has increasingly shifted to other Asian economies which are experiencing rapid economic transformation at the present time. The four Tigers share a range of characteristics with other Asian economies, such as China and Japan, and pioneered what has come to be seen as a particularly "Asian" approach to economic development. Key differences include initial levels of education and physical access to world markets (in terms of transport infrastructure and access to coasts and navigable rivers, which are essential for cheap shipping).

Africa and in particular Uganda has the opportunity to accelearate development using ICT as an engine of economic growth. One strategic area is software development. Microsoft Inc., Google and Yahoo for example are among the richest companies in the world and contribute a lot of taxes to governments around the world. What is interesting is that most of the world's successful companies like Microsoft Inc., Cisco Systems, IBM, Sun Microsystems, Yahoo and Google trace their roots at education institutions. Makerere University being a premier University in Uganda and in the region holds key to spur the software industry and lead to spin off companies that could accelarate economic growth. Luckily by 2008 Uganda will have the highest ICT Human Capital Index in Africa making it an attrative investment destination for ICT companies. This is on the assumption that the other factors like political stability will be favourable to foreign direct

investment and local investment. However, it must be noted that as much as countries like Uganda would like to attract foreign investors, no country in the world has ever developed on the basis of foreign investors alone; there need for a strong local private sector. Makerere University will work towards a stronger local private sector industry focussing on software solutions.

3.2. Education Industry

ICT should enable universities to operate 24-7 via online tutors who could operate from their homes, thus providing employment from home. Many universities across the globe provide education to thousands of students off campus (online) for several awards ranging from certificates to degrees in different disciplines. Phoenix University [10] is one example of a University that offers online diploma and degree programs. ICT present opportunities for lifelong learning in Africa. ICT present students with an opportunity to register and receive information online. The largest repository of information is the Internet and as of today thousands of journals and books are available online. Governments are looking at education as an Industry and many universities are aggressively recruiting international students on the online degree/diploma programs. Makerere University will take advantage of ICT to run online academic (diploma and degree) programs and expand distance education by introducing tele-education (tele-education is the application of telecommunication systems (the use of ICT) to provide distance education). Makerere University will rollout tele-education in Eastern Africa under the auspices of the African Union.

In addition to running core programs in computing and ICT, Makerere University will run crosscutting computing courses for all the students in the University to enable them integrate ICT in their disciplines.

There is also the opportunity to use the modern computing facilities (Baryamureeba, 2007) [6] being put up to conduct ICT skills training and awareness to local governments, students in pre-University education institutions (primary, secondary and tertiary institutions), the public and private sectors. Makerere University also plans to encourage other academic institutions in Uganda including Universities to outsource computing training from Makerere University since many universities in Uganda do not have either sufficient computing facilities or human capacity to impart the right skills to the students. This way the University expects to generate income from these occasional or short -term or part-time or modular students.

3.2.1. Midnight University

In a bid to provide education for all especially in ICT, Makerere University Faculty of Computing and IT started midnight classes in professional courses like Cisco Certified Network Associate (CCNA), International Computer Driving License (ICDL) and Certificate in Computer Applications (CCA). All these courses are available online. With the growing emphasis of tele-education and e-learning

Makerere University is exploring the opportunity of becoming a 24/7 University and open its doors to thousands of people across the globe. In light of this Makerere University is currently training several online tutors in all the disciplines offered at Makerere University. Makerere University Faculty of Computing and IT is involved in setting up government owned e-learning centres countrywide.

The Midnight University will give an opportunity to students to study from their homes if they have a computer connected to the internet or use the facilities at the university at night and be able to be tutored by lecturers from either their homes or on campus at night. This flexibility will make it possible for many Ugandans and others to access education. The Midnight University will also benefit from the African Diaspora especially the Ugandan Diaspora since when it is midnight in Uganda its in the afternoon in North America and morning in Asia. It is possible to exploit the time differences and use the Diaspora to participate in the educational activities in Uganda.

3.2.2. Engage the African Diaspora

More than 10,000 Africans are senior experts in Science and Technology and innovations in developed countries. There is need for Makerere University in particular and African Universities in general to tap into this skilled human resource with the aim of transferring high-end skills and knowledge to the local experts on the African continent. With ICT it is possible to turn brain drain into brain gain and in the end have brain circulation. The African Diaspora in the area of ICT/ Science and Technology hold key to Africa's development and should team up with the local scientists in the area of innovations, e-supervision of graduate students, e-tutoring, e-learning, tele-education and tele-medicine. Thus African governments must put incentives such as dual citizenship, centres of excellence and attractive salary packages in place so as to attract the Africa's best brains just like China and India did and continue to do. The advantage is that with the use of ICT they do not have to be physically in Uganda.

3.3. Consultancy

The Faculty of Computing and IT at Makerere University has the cream of the ICT Consultants in the country and has been providing consultancy services to the different sectors through its consultancy arm ICT Consults Ltd [8]. As both multinational and national companies setup business in Uganda, the demand for local consultants is going to increase. There will be need to outsource Information Technology services like network and system administration services from outside the companies.

Universities world over get most of their finances from the private sector and a case in point is Stanford and Harvard. So it is in the best interest of Makerere University to place a key role in nurturing private sector growth by providing supportive human resource and creating an environment for spin off companies from University incubation and innovations centres. ICT Consults Ltd has been

giving on the job consultancy training to continuing and former graduates of Makerere University Faculty of Computing and IT and as a result these trainees have established some successful spin off ICT companies within Uganda and neighboring countries.

Policy development. In most African countries including Uganda, enabling policies in the ICT sector, telecommunications sector and other related sectors are still lacking or restrictive. Makerere University has developed a knowledge base in this area and is ready to provide these services to the Government of Uganda and the region through consultancy and research.

3.4. Private Sector-University partnerships

ICT provide opportunities for Universities to engage in business with the private sector and increase on the tax base. Makerere University has already put in place an investment policy to enable such partnerships and joint ventures.

3.5. Outsourcing Services

The other opportunity for universities in the area of outsourcing is data and call centres business, software development and customization, customer care and support. The Government of Uganda has approved Makerere University Faculty of Computing and IT as the lead institution in incubating and rolling out data and call centres in Uganda. Makerere University in partnership with the Government of Uganda and the private sector is going to set up commercial data centres and call centres at Makerere University and across the country. The data and call centres at Makerere University will benefit from the computing facilities and bandwidth at night when most of the students are not utilizing the resources. This may lead to optimal utilization of resources.

Makerere University plans to offer remote (offsite) and onsite ICT services like network and system administration, information/ data processing and backup data centres to both local and international companies. The specialized human resource now exists in the Department of ICT Services and the Department of Software Development and Innovations within the Faculty of Computing and IT, ICT Consults Ltd and the Directorate for ICT Support at Makerere University.

3.6. Telemedicine

Telemedicine means using various forms of telecommunications/ ICTs to deliver health services across a distance, re-creating a clinical environment to provide patients with basic information and specialists with clinical advice enabling them to operate on the patient. Makerere University in liaison with Mulago National Referral Hospital in Uganda will undertake telemedicine rollout in Eastern Africa under the auspices of the African Union. Makerere University Faculty of Computing and IT is partnering with the College of Heath Sciences at Makerere University to provide commercial telemedicine services to the Uganda community, especially those in low-income communities but with access to ICT.

3.7. University Management

Makerere University sees information systems offering an opportunity to down size administrative staff and make the administrative services faster and more efficient. Makerere University commissioned an integrated information system comprising of the Academic Records Information System, Finance Information System, Human Records Information System and the Library Information System in January 2007. We hope that this integrated system will help the University to cut on administrative costs and provide timely and efficient services. Other academic institutions and government departments that have not yet adopted ICT in management may learn from the local good practices of using ICT in management.

4. Concluding Remarks

In this paper we have discussed the relation between ICT and economic growth. The role of Makerere University in enabling ICT as an engine for Uganda's economic growth has been discussed. We have stressed the importance of University-private sector partnerships as the major formal channel for two-way knowledge transfer. Lastly we have proposed opportunities Makerere University can exploit in this new area of ICT-led economic growth and generate income.

References

Baryamureeba, V. (2007), On curricula for creating human capital needed for ICT-led economic growth: Case of Faculty of Computing and IT. Published by the Partnership for Higher Education in Africa in the Proceedings for Frontiers of knowledge for Science and Technology in Africa: University Leaders Forum, University of Cape Town, South Africa.

Caspary, G. (2002), 'Information Technologies to Serve the Poor: How Rural Areas Can Benefit from the Communications Revolution', D+C Development and Cooperation No. 1, January/February 2002, pp. 4.

Computing Curricula 2005, @2005, held Jointly by the ACM and IEEE Computer Society, www.acm.org

Curtain, R. (2004), Information and Communications Technologies and Development: Help or Hindrance? Report Commissioned by the Australian Agency for International Development.

Eggleston, K; Jensen, R; and Zeckhauser, R (2002), 'Information and communication technologies, markets and economic development', in G.Kirkman et al. The Global Information Technology Report: Readiness for the Networked World. Oxford University Press, New York, pp 62-63.

Heeks, R; 2002, 'I-Development not e-Development: special issue on ICTs and Development', Journal of International Development Vol. 14, p 7.

http://www.foundation-partnership.org/pubs/leaders/assets/papers/BaryaSession5Part2.pdf

ICT Consults Ltd website, www.ict.co.ug

NetTel@Africa website, www.nettelafrica.org

Phoenix University (www.phoenix.edu)

Soubbotina P.T., Sheram, K., (2003) Beyond economic growth: Meeting the Challenges of Global Development, ISBN: 0-8213-4853, The World Bank, Copyright © 2003 National Council on Economic Education.

The Dutch development agency SNV works with local communities to set up specific enterprises and communities along a trekking trail...- http://www.propoortourism.org.uk/nepal_sum.html.

39
The Role of Academia in Fostering Private Sector Competitiveness in ICT Development

Tom Wanyama and Venansius Baryamureeba

This paper presents the prerequisite to producing computing graduates who have the skills required to fostering private sector competitiveness in information and communications technology development. Furthermore, the paper discusses the steps the Faculty of Computing and Information Technology at Makerere University, has taken to ensure that our graduates are of high quality and have the computing skills needed by the private sector and other potential employers. Finally, the paper presents the issues that need to be addressed, so to ensure sustainable private sector competitiveness in information and communications technology development.

Introduction

Increased adoption of eCommerce and Information and Communications Technology (ICT) in the private sector often leads to expanded economic growth by opening new markets, increasing access to market information, and improving efficiency (The Asia Foundation, 2001). For example, the use of eCommerce has the potential to expand the operations of Small and Medium Enterprises, and increase their competitiveness in the global supply chain network. Moreover, for many developing countries, the exportoriented information services, or "eServices" sector provides tremendous opportunities for developing economies, even for those countries that have limited economic potential due to scarce capital, poor infrastructure, and limited natural resources. This sector is unique in that it allows a local company to provide high-value services, such as software development and data processing that can be delivered to the recipient country across the Internet. However, for any country to attract, maintain and benefit from private sector investment in ICT, it must have a critical mass of highly skilled people. Therefore, it is no surprise that China and India that have a large number of Computing and Technology institutions that have taken the lead in benefiting from private sector investment in ICT among the developing countries.

Like in all other science fields, in the computing field, it is very important that institutions of higher learning adhere to discipline definitions that have national, regional and international recognition. This assures the private sector and especially the foreign investors that people produced by the academia have at least the minimum skills expected of someone in their disciplines. Inline with this

reasoning, the Government of India ensures that institutions of higher learning take into account the concept of globalization during discipline definitions and curricula development. This is done through the country's agency for quality assurance in higher education (Gopal et al., 2006). It should be noted that assuming all the other factors constant, an investor views a country where he/she does not have to retrain workers as a better investment destination than a country where the workers need to be trained so to attain the minimum skill-standards that the investors is familiar with. In the worst-case scenario, if the discipline definitions in a country do not adhere to any standard, foreign investors may feel obliged to recruit workers from other countries, which reduces the competitiveness of their investments, and makes the country a less attractive investment destination.

Having clear and recognized discipline definitions is a necessary, but not a sufficient requirement, for higher institutions of learning to produce people with relevant skills, to foster private sector competitiveness in the ICT development. Instead, it is necessary to complement clear and recognized discipline definitions with curricula that is relevant to the country and/or region where the institution is located, and that provide students with at least the minimum skills which are internationally expected of people in the various defined disciplines. In order to ensure that appropriate curricula are developed, the academia has to consult with the following entities during curriculum development and review: professional bodies, quality assurance agencies, and the private sector. In some cases, governments can actively encourage curricula development or improvement in particular areas of national interest through scholarships and grants (Taylor et al., 2006).

Provision of appropriate skills for fostering private sector competitiveness in ICT development requires supplementation of formal diploma and degree programs that are associated with recognized discipline definitions, and that have appropriate curricula; with short courses that provide skills in specific areas, technologies, equipment, and/or software. Furthermore, it is necessary that all forms of training be complemented with academia-industry collaboration through joint activities such as research, industrial training, consultancies, workshops, and conferences.

Well defined computing disciplines, as well as good curricula for formal programs and short courses cannot lead to producing highly skilled computing graduates, without highly motivated and well trained academic human resource. Therefore, it is imperative that higher institutions of learning acquire and retain this type of human resource so to produce graduates who foster private sector competitiveness in ICT development.

The outline of this paper is as follows. The definition of disciplines is discussed in Section 2, and Section 3 deals with curricula development. Section 4 addresses the role of short courses, Section 5 addresses the role of the academia-industry collaboration, and Section 6 addresses the role of academic human resources in fostering private sector competitiveness in ICT. Section 7 deals with efforts being made by the Faculty of Computing and Information technology (FCIT) to foster

private sector competitiveness in ICT development. Section 8 addresses the future of the FCIT in fostering private sector competitiveness in ICT development, and Section 9 presents key conclusions.

Definition of Disciplines

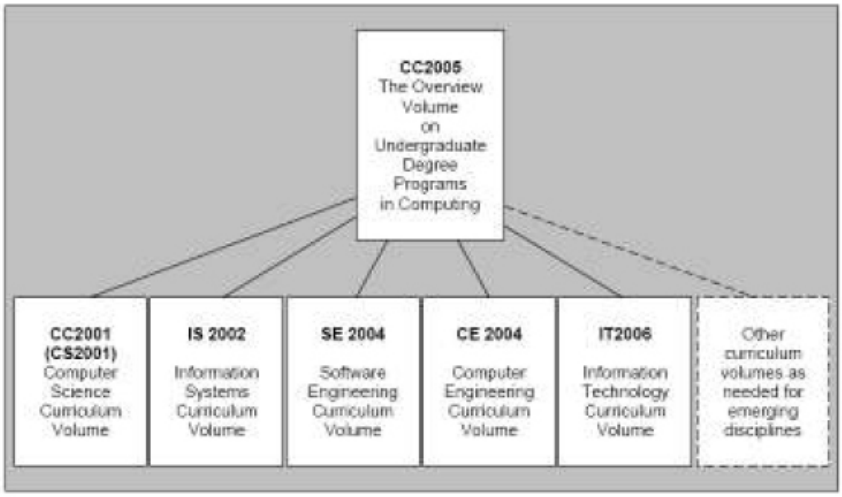
Professional bodies and quality assurance agencies play a very important role in the definition of disciplines because of the following reasons:

- Professional bodies have membership of academicians with varied teaching and research experiences, as well as professionals who have industry experience and are well versed with the interaction of various fields and disciplines in industry. Knowledge of these two types of professionals makes an important input to the definition of disciplines.
- Quality assurance agencies normally employ auditors who have extensive knowledge in their fields of expertise.

As a matter of fact, professional bodies and quality assurance agencies have a responsibility to define disciplines and to develop benchmarks and/or frameworks for evaluating discipline programs developed by academic institutions. Unfortunately, professional bodies in Uganda, such as the Uganda Institution of Professional Engineers (UIPE), Uganda Computer Society (UCS) and the quality assurance agency, the National Council for Higher Education (NCHE) have not developed any discipline definitions for the computing field. Moreover, these bodies and agency are not affiliated to any international organization so as to ensure that any discipline definitions that they make are internationally accepted. However, at the international level, a number of professional bodies are involved in discipline definitions of the computing field.

The computing field has grown rapidly and in many dimensions so that the many degree programs in the field have left many students, parents, and potential employers of computing graduates confused. To dispel this confusion, Association for Computing Machinery (ACM), Institution of Electrical and Electronic Engineers – Computer Society (IEEE-CS), and the Association for Information Systems (AIS) have made efforts to clearly define the various computing disciplines (The Joint Task Force for Computing Curricula, 2006). These efforts have resulted in defining five computing disciplines as shown in Figure 1. Moreover, Figure 1 reveals that computing professionals are aware that in the future, other new computing disciplines shall have to be defined (upcoming computing discipline include Bioinformatics and Computer Engineering Technology). Please note that the years shown against each computing program in Figure 1, stands for the year when the curriculum for the discipline was last updated.

Fig 1: Computing Disciplines [adopted from Computing Curricula 2005, @2005, held Jointly by the ACM, AIS and IEEE Computer Society]



Curriculum Development

It is a role of the academia to ensure that computing curricula foster private sector competitiveness in ICT development. To do this, the academia needs to work closely with professional bodies (The Joint Task Force for Computing Curricula, 2006), quality assurance agencies (QAA, 2006), and the private sector (Rizvi et al., 2005), during curriculum development and review. In addition, it is important that professional bodies and quality assurance agencies develop their own generic curricula, which the academia can use as benchmarks when developing curricula for their institutions. Like in the case of discipline definitions, Uganda's professional bodies and quality assurance agency do not have any such curricula in place. But as revealed in the following examples, the case is different in other countries:

- In the United Kingdom the Quality Assurance Agency (QAA) was established to provide an integrated quality assurance service. The main activities of the QAA are to ensure the quality of education delivered in UK Universities and other institutions of higher education. This is done at a subject level, and also at an institutional level, using periodic reviews. These reviews involve the production of self-evaluation documents by the institutions, and audit visits of the institution by QAA auditors (QAA, 2006).
- The Government of Bermuda participates in the development of curricula through its Ministry of Education (Barron et al, 2001).
- The Government of the United States of American influences curricula development and improvement through scholarships and grants (Taylor et al., 2006).

Moreover, ACM, AIS, and IEEE-CS have developed generic computing curricula that can be customized to the needs of a particular country and/or region while keeping an international outlook of the discipline in the curricula (The Joint Task Force for Computing Curricula, 2005).

Short Courses

In order to ensure that the private sector obtains all the computing skills required to keep it competitive, it is necessary that the academia supplement the conventional programs in the defined disciplines with a range of short courses. Such courses cover specific skills that are known to be lacking in the conventional programs. They include courses that provide skills for implementing and maintaining vendor-specific technologies, hardware, and software (vendor-specific certification programs). In the computing field, vendor-specific certification programs include the following courses:

- Certificate in Computer Applications (CCA)
- International Computer Driving License (ICDL)
- Oracle Certified Associate (OCA)
- Oracle Certified professional (OCP)
- Cisco Certified Network Professional (CCNP)
- Cisco Certified Network Associate (CCNA)
- IT Essentials I & II
- Security Plus
- IP Telephony
- Wireless LAN
- Microsoft Certification programs, namely: Microsoft Certified Professional (MCP), Microsoft Office Specialist (MOS), Microsoft Certified Database Administrator (MCDBA), Microsoft Certified Systems Engineer (MCSE), Microsoft Certified Systems Administrator (MCSA), and Microsoft Certified Solution Developer (MCSD)

Besides covering specific skills that are known to be lacking in the conventional programs, short courses such as those identified above enable professionals to keep up with the ever-changing technology in various fields. Keeping this in mind, it is imperative that short courses are developed and ran in closed collaboration with the industry (Siegel, 2006).

Academia-industry Collaboration

Institutions of higher learning not only provide the private sector with skilled human resources, but also support the sector in many other ways, including research and development (Rizvi et al., 2005). Therefore, collaboration between the academia and industry plays a crucial role in fostering private sector

competitiveness in ICT development through both indirect and direct knowledge transfer. The indirect knowledge transfer is achieved through such activities as industrial training, using members of the academia as consultants in the private sector, holding joint workshops and conferences, and journal publications. On the other hand direct knowledge transfer is achieved through collaborative research and/or purchases of patents.

It should be noted that in both types of knowledge transfer, knowledge flow is in both directions. That is, knowledge flows from the academia to industry and from industry to the academia. The reason being, the academia provides expertise in ICT while industry provides knowledge, in the application of ICT, project design and implementation, data analysis, and regulatory affairs. Knowledge from the academia to the industry is used to improve products and services, while the knowledge from industry to the academia is used to define disciplines, develop curricula, design short courses, and improve pedagogy. In reality the ICT industry is a fast-expanding major source of funding for research in academia and the principle source of applied ICT research. In return, industry gains valuable insight into its products from ICT experts in academia as well as leaders in the basic sciences. Ultimately, of course, the private sector benefits from a better understanding of ICT as an enabler, and from better ways of applying the technology.

Academic Human Resource

The creation of human capital starts with the definition of disciplines and the design of curricula. The curricula must address the skills set required of the graduate and at the end of each course the students must have acquired the required skills that will contribute to the overall skills set of the program. But for the students to acquire the skills set out for each course the lecturer himself/ herself must possess the same skills before he or she can lecture a given course. Therefore recruiting, training, and retaining academic staff in the Computing field is crucial to fostering private sector competitiveness in ICT.

Efforts by FCIT to foster Private Sector Competitiveness in ICT development

The Faculty of Computing and Information Technology (FCIT) at Makerere University has taken a multidimensional approach to fostering private sector competitiveness in ICT development. This approach has the following five main dimensions: discipline definition, curriculum development and review, short courses and academia-industry collaboration, and academic human resource.

The Discipline Definition Dimension. There are no national benchmarks for defining computing programs in Uganda. Therefore, FCIT defines its computing programs based on internationally accepted definitions, proposed by a joint task force of ACM, AIS, and IEEE-CS; through consultations with the professional bodies, and through reviewing published materials jointly produced by the professional bodies on computing discipline definitions. Moreover, the faculty

takes into account the level of development of ICT in Uganda as well as the development goals of the Government of Uganda when defining its programs, through consultancies and academia-industry collaboration. Consequently, FCIT has four academic departments, namely: Computer Science, Information System, Information Technology, and Networks (Please note that the FCIT is split in departments based on the discipline definitions proposed by ACM, AIS and IEEE-CS). Within these departments, the faculty runs computing undergraduate programs in computer science (Department of Computer Science), and in information technology (Department of Information Technology). Moreover, the faculty has proposed programs in Information Systems to begin in 2007/2008 academic year (Department of Information Systems), Computer Engineering to begin in 2007/2008 academic year, and Software Engineering to begin in 2007/2008 academic year. When these new programs begin, the Department of Networks shall be split into three departments: Department of Software Engineering, Department of Computer Engineering and Department of Data Communications and Computer Networks.

By having internationally recognized discipline definitions, FCIT ensures the following:

- Organizations in the private sector that employ our graduate's can easily compare their skills endowment with international organizations. Such comparisons normally arise in such cases as when organizations are seeking international collaboration and/or accreditation.
- Foreign investors can easily identify available skill based on discipline definition that he/she is familiar with.

The Curriculum Development and Review Dimension. In the absence of national curricula guides, FCIT follows internationally recognized curricula guides. However, through our links with the private sector and other employers of our graduates, we know that graduates from purely computing programs (computer engineering, computer science, information systems, information technology, and software engineering) would lack some of the skills required by the employers in Uganda in particular and the region in general. These skills are mainly professional and vocational. In order to address this, the Faculty of Computing and IT has integrated Cisco Certified Network Associate (CCNA), Cisco Certified Network Professional (CCNP), and IT Essentials in the computing degree programs.

By providing internationally recognized curricula that are customized to the needs of employers in Uganda and in the region, FCIT achieves the following:

- Reduction of the cost of computing/ICT training for the private sector hence fostering private sector competitiveness in ICT development
- Benefits associated with internationally recognized discipline definitions (see Section 7.1)

The Short Courses Dimension. It is not possible to have all the skills required by the employers especially the private sector integrated in one degree programme. Therefore, in addition to CCNA/CNNP and IT Essentials, all the other professional programs like Microsoft Certified System Engineer (MCSE), Oracle

Certified Network Associate (OCA), Oracle Certified Network Professional (OCP), Security Plus, IP Telephony and IT Essentials to mention but a few are conducted during the semester breaks and students are free to take them as optional courses. Similarly, all the other courses that provide skills needed by the employers (mainly the private sector) which have not been integrated into the program curricula are run during the semester breaks and the students depending on their interests in career development are allowed to take any course being conducted. This ensures that by graduation time, the graduates have gone through both academic and professional/vocational training and as a result have acquired most of the skills needed in the workplace; hence reducing the cost of training for the private sector.

The Academia-Industry Collaboration Dimension. The Faculty of Computing and IT, Makerere University has used its collaborations with the industry to help connect students with potential employers through the partnerships it has developed with a number of organisations. These organisations include Uganda Investment Authority, True African, Seven Seas, Public Service, Inspector General of Government (IGG), Uganda Wildlife Education Centre (UWEC), MFI Office Solutions and AFSAT Communications Uganda to mention but a few. These partnerships have exposed students to work-site tours, talk sessions and internship placements. Some of the students have also had the opportunity to be permanently employed.

The Faculty has also established research partnerships with organisations outside Uganda and used these to collaborate with the local industry. Two examples of this effort are the Quality of Service research that was funded by one of the partners, Cisco Systems and had its research carried out on the networks of Uganda Telecom, MTN Uganda and Makerere University. Another example is the research partnership with Radbord Nijmegen University in the Netherlands, which sent its students and the Faculty got them to visit the industry partners.

The Faculty's Work-Force Development Program (WDP) arranges the various partnerships. On average, the industry partners take on students from the faculty every 3 to 6 months, depending on the available projects and/or areas of study. The main challenge faced here is that the number of students is a lot higher than that which the industry has demand for. But the success is, all the students that go through the WDP get equipped with soft skills like writing a good CV and preparing for job interviews. Moreover, the issue of having more students than the private sector can absorb is being addressed by providing the students with entrepreneur skills so that they can be job creators other than job seekers.

The Academic Human Resource Dimension. Computing is a new discipline and universities in Uganda in particular and Africa in general have very few PhD holders in any of the computing subfields. At Makerere University - FCIT, the undergraduate programmes in computing are run by the fulltime local staff (mainly M.Sc. holders), that use both online (e-learning) and face-to-face classroom instruction. Due to lack of sufficient local human resource in the area of computing, some departments of the FCIT have started with only postgraduate programs so to generate people who can teach at undergraduate level, before starting the undergraduate programs. This is possible because our postgraduate programs do not only depend on local staff, but also depend on African Diaspora especially Ugandan Diaspora, PhD holders on projects supported by the development partners, occasional (short visits) Professors and visiting fellows on sabbatical leave. Most of the postgraduate students are supervised online by academic staff from institutions around the world. Furthermore, FCIT has devised the following activities as part of the strategies to boost the postgraduate programs:

- **Organizing The Annual International Conference on Computing and ICT Research**. (http://www.srec.cit.ac.ug/). This is an annual event that brings together scholars from all over the world every August of every year. This series of conferences started in 2005. Most scholars who come to the conference stay much longer at Makerere University working with local researchers and postgraduate students on several research projects. Some scholars give a series of seminars on different topics in computing. There is always a PhD colloquium that gives the PhD students an opportunity to get advice on their research from several experienced researchers. Postgraduate students especially PhD students have been able to get a 2nd or 3rd supervisor from either the scholars at the conference or scholars at the home institution of the conference participants as a result of networking with the scholars at the conference. Also this conference gives an opportunity to young scholars especially PhD students to have their research peer reviewed.
- **Hosting the International Journal of Computing and ICT Research (IJCIR).** (http://www.ijcir.org). This is a peer reviewed International journal with an objective of providing a medium for academics to publish original cutting edge research in the field of computing and ICT. IJCIR publishes two issues per year. The Journal publishes papers in computing and ICT and other related areas. This journal is hosted by Makerere University in the Faculty of Computing and IT. It has encouraged local researchers and postgraduate students especially PhD students to send in their papers for peer reviews. This journal only publishes papers that meet its high standards but also receives papers from young researchers, which are not necessary published but takes them through the review process so that the young researchers can gain from the reviews to improve their papers and resubmit or submit to other journals. In a way it acts as an

incubation facility for young researchers. On a good note most of those younger researchers who submitted papers in this journal have in the process gained confidence and are now submitting good papers to this journal and other international journals for possible publication.

- **Establishing a Program for Visiting Scholars and African Diaspora.** The Netherlands organization for cooperation in higher education (Nuffic) and Makerere University put in place a modest fund to support visiting staff on sabbatical leave to spend 3-12 months at Makerere University. This offer was also extended to African Diaspora of which the Ugandan Diaspora tremendously responded. Of all these categories the Ugandan Diaspora tends to stay much longer and even on return to their home institutions, they continue to dedicate a substantial amount of their time to the activities at Makerere University such as supervision, research, and online instruction with the help of ICT and digital learning environments like blackboard.

Future of The FCIT in fostering Private Sector Competitiveness in ICT development

As the largest producer of skilled personnel in ICT in Uganda, the FCIT plays a major role in fostering private sector competitiveness in ICT development. Therefore, the faculty is putting in effort to ensure that it produces computing graduates who have the relevant skills for the social and economic development of Uganda. However, a number of issues remain to be addressed so as to foster ICT development; some of these are discussed in the following subsections.

Quality of Computing Programs. Currently, there is not any quality assurance frameworks/ subject benchmarks at both national and University level to guide the process of establishing academic programs and developing curricula. Therefore, quite often there are rifts between faculties, on which faculty should host a program and at times faculties have put in place programs, which cannot measure up to international benchmarks. This in the end renders graduates from such programs unacceptable internationally as a result of following curricula that does not provide the relevant skills or being taught by staff that lack the necessary skills. Fortunately, the National Council for Higher Education is in the process of putting in place a quality assurance framework and subject benchmarks, and Makerere University is also in the process of putting in place a quality assurance framework. We look forward to receiving these frameworks and to reviewing our programs and curricular accordingly. We also hope that professional bodies, such as the UIPE and UCS shall follow the example of the National Council for Higher Education and Makerere University, and development curricula development guidelines of their own. Otherwise, the FCIT shall continue to develop its programs based on the economic and development needs of Uganda, and on internationally accepted frameworks.

The computing field is undergoing tremendous developments day by day as a result of technological advances. As a result, many students, potential employers, and scholars find it difficult to comprehend what computing is all about. For example, it is common for many academicians to imagine that software engineering is part of computer science and that computer engineering is not a computing discipline. To address this issue, the FCIT and other stakeholders need to sensitize University leaders and Quality Assurance Agencies about the computing disciplines. Information such as Table 1 below is crucial in the sensitization work. The table presents eleven skill areas (column 1), the associated skills (column 2) and the level (0-5) to which those skills are expected of a graduate of a given discipline (column 3). Using such information in discipline definitions shall assist Makerere University and other institutions to optimally utilize the scare resources by housing all the computing programs under one faculty.

Please note the following expectations from graduates of each of the computing disciplines:

- Computer engineers should be able to design and implement systems that involve the integration of software and hardware devices.
- Computer scientists should be prepared to work in a broad range of positions involving tasks from theoretical work to software development.
- Information systems specialists should be able to analyze information requirements and business processes and be able specify and design systems that are aligned with organizational goals.
- Information technology professionals should be able to work effectively at planning, implementation, configuration, and maintenance of an organization's computing infrastructure.
- Software engineers should be able to properly perform and manage activities at every stage of the life cycle of large-scale software systems.

Having a sensitized public and university administrations in particular on the disciplines of the computing field will enable putting the core lecturers required on computing programs under one faculty.

Finally, it should be noted that no single student can have all the computing skills set, but students from different computing programs can undertake joint projects (this is easy when all the students fall under the same faculty). Since most employers normally require graduates in multiple computing disciplines (computer science, computer engineering, software engineering, information systems, information technology), joint multi-disciplinary projects provide students with teamwork skills that are crucial in today's workplace.

Academia-Industry Partnership. Currently, Makerere University does not have a Technology Transfer Office, which would track innovations at the university so as to market them to the private sector. On the other hand, the FCIT is working towards setting up such an office, and we hope that the University shall also setup

one in the near future. Furthermore, the FCIT would like to increase collaboration with the industry to a level where organizations in the private sector shall be able to sponsor long term and/or permanent research chairs in the faculty.

Availability of Academic Human Resource. The FCIT has to continue training people at undergraduate, masters and PhD level so to work towards providing adequate academic human resources. But we also request Government of Uganda to put in place a framework to involve the Ugandan Diaspora in Science and Technology, and Innovations in the country. With ICT it is possible for them to still contribute from wherever there are based. Also centres of excellence with good terms and conditions of service must be put in place to attract skilled Ugandan Diaspora especially in the area of Science and Technology/ ICT to return home. Ugandan Diaspora have gained knowledge if tapped by our Universities and Government could boost Research and Development and innovations, which in turn would foster private sector competitiveness in ICT development.

Table 1: Relative Performance Capabilities of Computing Graduates by Discipline [adopted from Computing Curricula 2005, @2005, held Jointly by the ACM and IEEE Computer Society]

Area	*Performance Capability*	*CE*	*CS*	*IS*	*IT*	*SE*
Algorithms	Prove theoretical results	3	5	1	0	3
	Develop solutions to programming problems	3	5	1	1	3
	Develop proof-of-concept programs	3	5	3	1	3
	Determine if faster solutions possible	3	5	1	1	3
Application programs	Design a word processor program	3	4	1	0	4
	Use word processor features well	3	3	5	5	3
	Train and support word processor users	2	2	4	5	2
	Design a spreadsheet program (e.g., Excel)	3	4	1	0	4
	Use spreadsheet features well	2	2	5	5	3
	Train and support spreadsheet users	2	2	4	5	2
Computer programming	Do small-scale programming	5	5	3	3	5
	Do large-scale programming	3	4	2	2	5
	Do systems programming	4	4	1	1	4
	Develop new software systems	3	4	3	1	5
	Create safety-critical systems	4	3	0	0	5
	Manage safety-critical projects	3	2	0	0	5
Hardware and devices	Design embedded systems	5	1	0	0	1
	Implement embedded systems	5	2	1	1	3
	Design computer peripherals	5	1	0	0	1
	Design complex sensor systems	5	1	0	0	1
	Design a chip	5	1	0	0	1
	Program a chip	5	1	0	0	1
	Design a computer	5	1	0	0	1
Human-computer interface	Create a software user interface	3	4	4	5	4
	Produce graphics or game software	2	5	0	0	5
	Design a human-friendly device	4	2	0	1	3
Information systems	Define information system requirements	2	2	5	3	4
	Design information systems	2	3	5	3	3
	Implement information systems	3	3	4	3	5
	Train users to use information systems	1	1	4	5	1
	Maintain and modify information systems	3	3	5	4	3
Information management (Database)	Design a database mgt system (e.g., Oracle)	2	5	1	0	4
	Model and design a database	2	2	5	5	2
	Implement information retrieval software	1	5	3	3	4
	Select database products	1	3	5	5	3
	Configure database products	1	2	5	5	2
	Manage databases	1	2	5	5	2
	Train and support database users	2	2	5	5	2

IT resource planning	Develop corporate information plan	0	0	5	3	0
	Develop computer resource plan	2	2	5	5	2
	Schedule/budget resource upgrades	2	2	5	5	2
	Install/upgrade computers	4	3	3	5	3
	Install/upgrade computer software	3	3	3	5	3
Intelligent systems	Design auto-reasoning systems	2	4	0	0	2
	Implement intelligent systems	2	4	0	0	4
Networking and communications	Design network configuration	3	3	3	4	2
	Select network components	2	2	4	5	2
	Install computer network	2	1	3	5	2
	Manage computer networks	3	3	3	5	3
	Implement communication software	5	4	1	1	4
	Manage communication resources	1	0	3	5	0
	Implement mobile computing system	5	3	0	1	3
	Manage mobile computing resources	3	2	2	4	2
Systems Development Through Integration	Manage an organization's web presence	2	2	4	5	2
	Configure & integrate e-commerce software	2	3	4	5	4
	Develop multimedia solutions	2	3	4	5	3
	Configure & integrate e-learning systems	1	2	5	5	3
	Develop business solutions	1	2	5	3	2
	Evaluate new forms of search engine	2	4	4	4	4

Conclusions

It is necessary that the academia and the private sector build strong collaborative relationships. Such relationships should not be limited to industrial training, and research and development, but should also include other important areas, such as curricula development, business proposal writing, improvement of business processes, and continuous training of private sector workers. Since it is wrong to think that a single academia-private sector model is beneficial to all units of academic institutions, as well as the private sector, it is essential that each unit of the academic institutions identifies areas where they can build effective academia-private sector relationships. In addition, these units have to identify the endowments they have that can benefit the private sector. The Faculty of Computing and Information Technology at Makerere University has put a lot of mechanisms to foster private sector competitiveness in ICT development. These include the following: development of internationally recognized curricula that take into account the local and regional needs of the private sector, running short courses to fill the gaps left by formal degree and diploma programs, and providing continuous training to computing professionals, and providing adequate academic human resource to ensure that the quality of graduates of the faculty is high.

References

Barron B. J., C. K. Martin, E. S. Roberts, A. Osipovich, and M. Ross (2001). "Developing a Secondary School Computing Curriculum for Bermuda Public Schools", Available at http://bermuda.stanford.edu, accessed on 10th November 2006.

Gopal R., and M. Bhattacharya (2006). "Globalization and Quality Assurance in Higher Education", Available at www.caluniv.ac.in , and accessed on 11th November 2006.

QAA - The Quality Assurance Agency for Higher Education (2006). "QAA around the UK and QAA internationally", Available at www.qaa.ac.uk, and accessed on 12th November, 2006

Rizvi I. A., and A. Aggarwal (2005). "Enhancing Student Employability: Higher Education and workforce Development", Proceedings of the 9th Quality in Higher Education Seminar, Birmingham, UK

Siegel D. (2006). "Industry, Academia Build Education Partnerships", In Focus Continuing Education – National Society of Professional Engineers, pp. 34 - 36

Taylor C., R. Shumba, and J. Walden (2006). "Computer Security Education: Past, Present, and Future", Computer Security Education, pp. 67 - 78

The Asia Foundation (2001). "Information and Communication Technology in Asia", Available at www.asiafoundation.org, and accessed on 15th November, 2006

The Joint Task Force for Computing Curricula (2006). " The Computing Curricula 2005: The Overview Report", A Cooperative Project of the Association for Computing Machinery (ACM), the Association for Information Systems (AIS), and the Computer Society (IEEE-CS, A New Volume of the Computing Curricula Series

40
Conceptual ICT tool for Sustainable Development: The Community Development Index (CDI)

Joseph Muliaro Wafula, Anthony J Rodrigues and Nick G Wanjohi

As part of the shift from society to the community as the object of rule, disembodiment of expertise knowledge especially into ICT tools is crucial in empowering people to manage their lives. It also enables people to adopt a prudent and calculative approach to self-governance through being enabled to take appropriate economic, social and political decisions. The belief that there is a technological silver bullet that can 'solve' illiteracy, ill health or economic failure reflects scant understanding of real poverty. It is on this basis that new indicators monitored throuch ICT, and based on sources of poverty in our society, need to be developed to help understand poverty from a bottom-up/community-led approach. Hence, a preventive approach rather that a curative approach is recommended for poverty alleviation. By monitoring sources of poverty indicators, a community or government can be able tell whether it is on the right track of alleviating poverty or applying appropriate strategies.Therefore, there is need to develop ICT-based tools such as Community Development Index (CDI), for policy makers and citizens alike to be able to measure their own strengths and weaknesses, as well as recognize their opportunities. This paper presents the CDI concept that can be used to develope a decision support tool for policy makers, leaders and their people.

Introduction

People all over the world have high hope that new technologies will improve health, increase social freedom, increase knowledge and productive livelihoods. New technology policies can spur progress of attaining Millennium Development Goals (MDGs) especially having known from history that technology has been a powerful tool for human development and poverty reduction. Technology works well as a powerful tool for human development and poverty alleviation when good and appropriate policies, regulators and a high degree of transparency in its deployment are exhibited. Developing countries lack policies and institutions needed to manage risks associated with technology. Policy not charity, will determine whether new technologies will become a tool for human development everywhere in the world (UNDP, 2001). It is important now that developing countries come up with ICT policies that support and enable monitoring and evaluation of poverty alleviation programs.

Developing countries came to political independence with governments that had formal structures that were somehow representative in nature. Their political leaders in their bid to consolidate political power opted for highly centralized modes of governance. This centralized mode of governance in the developing countries was reinforced by a culture of politics of patrimony in which all powers and resources flow from the head of state to the citizens. This pattern of power and resource distribution was strongly supported by both domestic and external actors until the late 1980s. The reasons adduced for adopting this approach included: Rapid economic and social development actualized through centralized planning; unity and national integration; containment of corruption and political stability. A monocentric governance model was adopted and this affected the manner in which decentralization was approached. In monocentric governance model, an administrative decentralization or deconcentration rather than political or democratic approaches was used. Democratic decentralization includes not only the transfer of responsibilities but also of financial and human resources to semi-autonomous entities with their own decision-making powers. In the last decade, however, many countries changed course dramatically. They have abandoned the monocentric political model and sought to replace it with its exact opposite, which is the polycentric governance model (Olowu, 2003). The Polycentric structure of governance accepts the idea of multiple centers of power within a state. This involves devolution to local governance organs that enmesh both state and society institutions at regional and community levels.

Sustainable development requires appropriate philosophy besides appropriate technology.

Community-led rural development is widely regarded in the US and Europe and Australia as the key to improving the sustainability of disadvantaged regions and providing local people with the capacities to respond positively to change. While at face value such development increases local autonomy and control, a number of scholars have recently located community-led development as part of a broader shift from government to governance (Herbert-Cheshire Lynda and Higgins Vaughan, 2004). Here, new institutional and administrative arrangements and actors extending beyond formal state authorities play an increasingly significant role in ensuring that communities have the capacities to take a more active role in their development. This shift to governance implies that in order for communities to successfully take charge of their own development, they must first become enmeshed in a network of relations that assists them in acquiring the capacities to govern themselves responsibly.

The emergence of *'the community'* as an object of knowledge in public policy, formal political discourse and development initiatives is indicative of a fundamental shift in the spatialisation of government. Unlike previous forms of government that sought to achieve national security through state-based socialised forms of intervention and responsibility, this is a new type of ruling —governance through community. Such governing through community is advanced liberal in that it seeks to desocialise and individualise risk, with subjects encouraged to *'shape*

their lives according to a moral code of individual responsibility and community obligation' (Herbert-Cheshire Lynda and Higgins Vaughan, 2004). That is to say, instead of citizenship constituted in terms of social obligations and collectivised risk, it becomes individualised based on one's capacities to conduct oneself in an entrepreneurial and responsible manner. Such entrepreneurialism forms the basis for governing through community. As part of the shift from society to the community as the object of rule, expertise knowledge becomes crucial in empowering people to manage their lives and adopting a prudent and calculative approach to self-governance through appropriate social decision making. This expertise translates the political concerns of the people into outputs in form of government for instance: efficiency; industrial productivity; law and order; accountability of political leadership and political stability among others. Armed with techniques that promise improved financial management, a better lifestyle, efficient work practices or, in the case of rural population, empowerment to improve community economic well-being, these expert knowledge is directed at enhancing self-regulatory capacities, thereby aligning political objectives with broader community goals. An expertise is considered to be crucial on account of neutrality and mechanism for enabling community members to be aware of their capabilities and develop positive entrepreneurial attitudes through which they can build their community and leadership capacities, reduce their dependency on the government and create sustainable development in the form of technological enhancement, wealth creation and rising standards of living.

The contemporary advanced liberal emphasis on 'bottom-up' or community-led development is indicative of the fact that the management of resources and the attendant risks are no longer an activity of the state alone, but the responsibility of citizens and their communities wherever they may be (Herbert-Cheshire Lynda and Higgins Vaughan. (2004).

In general, ICT has a big role to play in promoting good governance (Odendaal, 2003).

CDI Importance

The belief that there is a technological silver bullet that can 'solve' illiteracy, ill health or economic failure reflects scant understanding of real poverty. It is on this basis that new indicators based on sources of poverty in our society need to be developed to help understand poverty from a bottom-up/community-led approach. So as to have a preventive approach rather that a curative approach that so far has failed in helping alleviate poverty. Poverty is complex, and without identifying uniquely what are its sources whose compound effect lead to a state called poverty, and destroy them, poverty will continue to exist in different forms. It is reasonable to postulate that there exist a direct relationship between poverty sources and sustainable development: as sources of poverty increases, sustainable development capacity decreases. By monitoring sources of poverty indicators, a community, government or development partner can be able tell whether it is on the right track of alleviating poverty or applying appropriate strategies.

CDI supports self-reliance development model that enables communities to be confident enough to base their development on their own thinking and value systems without being defensive or apologetic. CDI is intended to enable communities become agents and not just beneficiaries of development. Self-reliance enables nations or regions to assume fuller responsibility for their own development within a framework of enlarged political and economic independence. It builds development around individuals and groups rather than people around development and it attempts to achieve this through the deployment of local resources and indigenous efforts. It mobilizes the creative energies of the people themselves. This contributes directly to the formulation of new values systems, to the direct attack of poverty, alienation and frustration, and to the more creative utilization of the productive factors. Self-reliant development, with its reliance on local rather than imported institutions and technologies, is a means whereby a nation can reduce its vulnerability to the decisions and events which fall outside its control: a self-reliant community will be more resilient in times of crisis. This self-reliance model has been partly adopted in the Economic Recovery Strategy (Government of Kenya, 2003) and Poverty Reduction Strategy Paper (Ministry of Finance and Planning, 2001) of Kenya. It has been known to solves to a good extend the unemployment problem as observed in China (McGinis, 1979). There are two methods that are commonly used to eliminate absolute poverty. The first one is by government enabling creation of jobs with incomes that are sufficient to meet basic needs to those who suffer from poverty. The second one is by the government forcefully through taxation, redistributing the income or corresponding real resources from those who have jobs and incomes to those who do not have jobs or income. Any combination of them is acceptable and possible (Howard and Gunner). However, in Kenya, the taxation option cannot solve the poverty problem since the ratio of the employed to unemployed is too small. Already Kenyans are among the most highly taxed people in the world. The option left is to create jobs and this is where ICT comes in handy particularly in service industry.

Donor case studies undertaken by OECD highlight the fact that identification and targeting of the poor is often broad-brush, and an assessment of European aid to Zambia showed that few, if any, donors came close to having the insights needed (Hjorth, 2003). There was neither much systematic analysis of the causes to poverty nor any in-depth discussion/identification of poor groups. Also a study in Zambia showed a distressing lack of knowledge about what aid does, particularly in terms of its poverty reduction effects. Adoption and use of ICT can enable better monitoring of parameters shown in Fig 1 and the understanding of poverty at its roots. A web-based application computing and displaying CDI for the targeted communities, would act as mirror to the communities, government and their development partners.

Policy, economic and technical factors initiated the development and deployment of the Internet. Government policies like the national information infrastructure (NII) initiative in the US played a pivotal role in Internet's widespread diffusion

as did a great deal of educational and research activities related to and utilizing the Internet. Such explanatory frameworks miss fundamental historical, social and cultural contexts associated with the diffusion of the Internet (Kim Pyungho, 2003). Supply-side variables alone cannot achieve a rapid and widespread diffusion of the Internet. In other words, there are certain economic, social and cultural conditions that intrigue people to enthusiastically embrace the Internet. An analogy can be drawn from this case for sustainable development. Most donors and governments have mostly emphasized the supply (push) side of development and ignored the demand (recipient) side of it. The development of the CDI would present more information on the demand side.

Radio and television have served as a social and cultural regulator to standardize our life by inculcating homogenized, mass consumer culture so that they become an effective means of 'control and coordination of production, distribution and consumption' in the industrialized modern society. This technology has not been used exhaustively in the developing countries. More so in the area of passing relevant information to communities in the form and languages they understand. By choosing to limit this technology, communities have been denied the opportunity to coordinate their production, distribution and consumption of what they produce. Although few, entrepreneurs have always been there but due to political reasons and fear, they are never allowed to provide these services freely, especially in Africa, where people live in communities within the context of mass cultures. The poor countries have lost on the opportunity to use radio and TV or any technology targeting masses in helping to alleviate poverty through provision of appropriate, timely and relevant information (see Fig.1). Most parts of Africa currently have access to radios, but do not use these tools of communication to drive development effectively.

We live in an era of shrinking states and expanding markets (Bratton, 2003). Whereas in many countries market liberalization has achieved overall efficiency gains and has often promoted sustainable practices, there is increasing evidence that these benefits have bypassed most rural areas in developing countries (Kuyvenhoven, 2004). In many remote areas, the disappearance of marketing boards and parastatals for Input provision have not been followed by private-sector initiatives to provide market services, and many farmers are further from the market than before. New, low-cost communication technologies (ICT) can play a major role in speeding-up the process of information diffusion and improving market efficiency, especially in the fields of input provision and marketing outlets.

As part of modernization on local government and support of reform agenda for a strong local government, the development of CDI would promote:

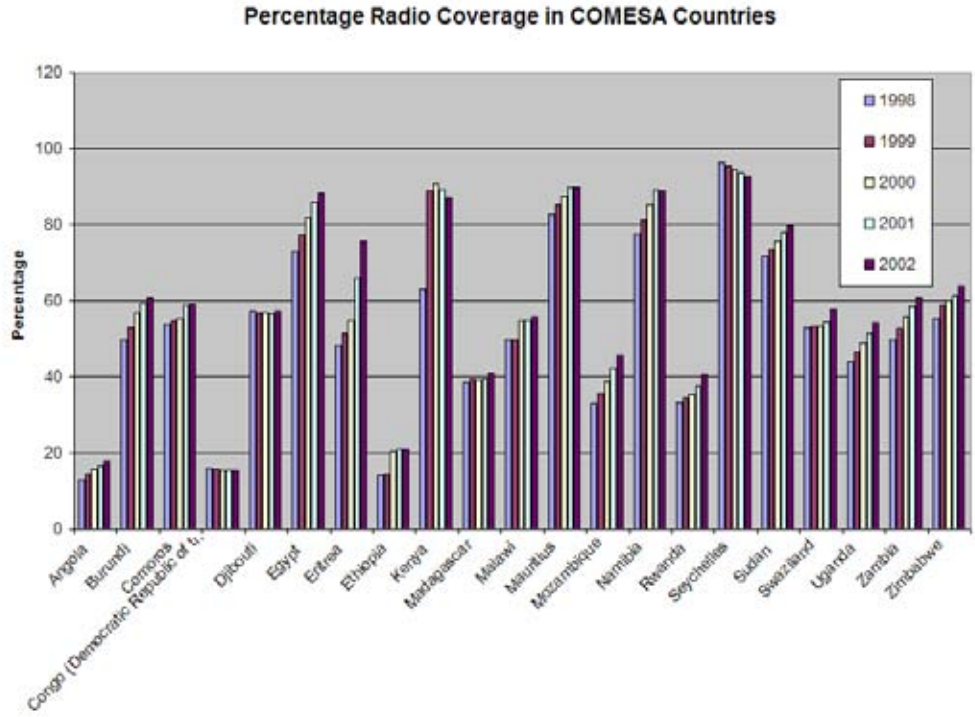
- Effective decision-making based on status of local communities with respect to capacity to sustain development.
- Preservation of local value systems.
- Empowerment of citizens at local level.

This is expected to ease the Local Government's current constrains imposed by the central government, provincial administration as well as the parliamentary constituencies (Association of Local Government Authorities of Kenya, 2004).

CDI for Local Communities and their Authorities

There is an emerging consensus that poverty reduction is what development is about, and poverty reduction is by now a major priority of the international development aid community. A recent poverty assessment shows that results are mixed, with many countries falling behind on reducing malnourishment and infant mortality and on increasing primary school enrolment and access to safe drinking water (Hjorth, 2003). The idealistic impulse to improve the standard of living of the poor is noble, but it is for the people to do what it takes to make change sustainable. But unless the actual policy solutions are well grounded in a deep understanding of the causes of poverty and how those causes have been, and can be, effectively addressed, they could end up with worse results than in the past. That is to say, in spite of best intentions, policies based on inadequate knowledge are likely to increase rather than reduce poverty. That is why there is need to develop indicators and compute an index that can help understand poverty from its roots. Effective poverty alleviation will require significant change in current structures, attitudes, and behaviour by people and their leaders alike. CDI is meant to capture this and help provide an in-depth understanding of the sources of poverty as a means of creating wealth at the local community level, and hence nationally.

Fig.1: Radio Coverage in COMESA countries Data points Source: ITU. (2005). World Telecommunications Indicators Database. 8th Edition.



CDI and the Local Authorities

Social scientists have looked at poverty from three broad definitional ways namely: absolute, relative and subjective poverty (Odhiambo, Omiti and Muthaka, 2005). Absolute poverty refers to subsistence poverty, based on assessment of minimum subsistence requirements, involving a judgment on basic human needs and measured in terms of resources required to maintain health and physical efficiency. These basic life necessities are then priced and the total figure constitutes the poverty line. Relative poverty refers to the use of relative standards in both time and place in the assessment of poverty. Therefore the notion of relative poverty is elastic and receptive to conventional and rapid changes. Lastly, subjective poverty is closely related to relative poverty in the sense that subjective poverty has to do with whether or not the individuals or groups actually feel poor. Absolute poverty is the type CDI is seeking to tackle.

Human Poverty Index for developing countries HPI-1 is based on measured effects of poverty using four indicators. These are namely: Probability at birth of not surviving to age 40; adults illiteracy rate; percentage of population without sustainable access to an improved water source; and percentage of children under weight for age (UNDP, 2004). It is important to realize that absolute poverty has continued to increase despite concerted effort to increase donor funding. This calls for a radical change in approach, particularly when it comes to information for policy makers as well as monitoring and evaluation of sustenance of community development projects by government, donors and the community itself.

Poverty alleviation strategies succeed when they emphasize on means of eliminating sources and causes of poverty. This can be done via identification of the causes and stopping them. Basically, an anti-poverty mechanism can be developed, with indicators build into a computer system for monitoring and evaluation so as to boost investment, good use of time and resources, prevent wastage, and help in wealth creation.

In the case of Kenya, funds directed at poverty alleviation such as Local Authority Transfer Fund (LATF) and the Constituency Development Fund (CDF) may be used to mobilize local human resources for productive purposes and activate new economic activities such as infrastructure building-roads, houses, bricks, plantations, transport, hotels, stone cutting, sand collection, export processing zones, irrigation projects, dams etc. Local authorities for example, may encourage the citizens to improve agricultural production: cattle, goat, sheep, poultry, fish, bee keeping, etc. Products from these projects should be sold to markets established by the local authorities. The local people may establish agro-based industries to process their farm produce, engage in candle making, skin-leather industry, production of bio-gas from cow-dung, making compost manure for organic farming, among others. Local authorities, for example, may create a wealth creation department to implement all its anti-poverty programs and activities. As a faster means, local authorities can then explore the global market through Internet and other ICT technologies with a view to enable their

communities obtain the equipment and inputs from around the world as well as sell their products regionally and globally.

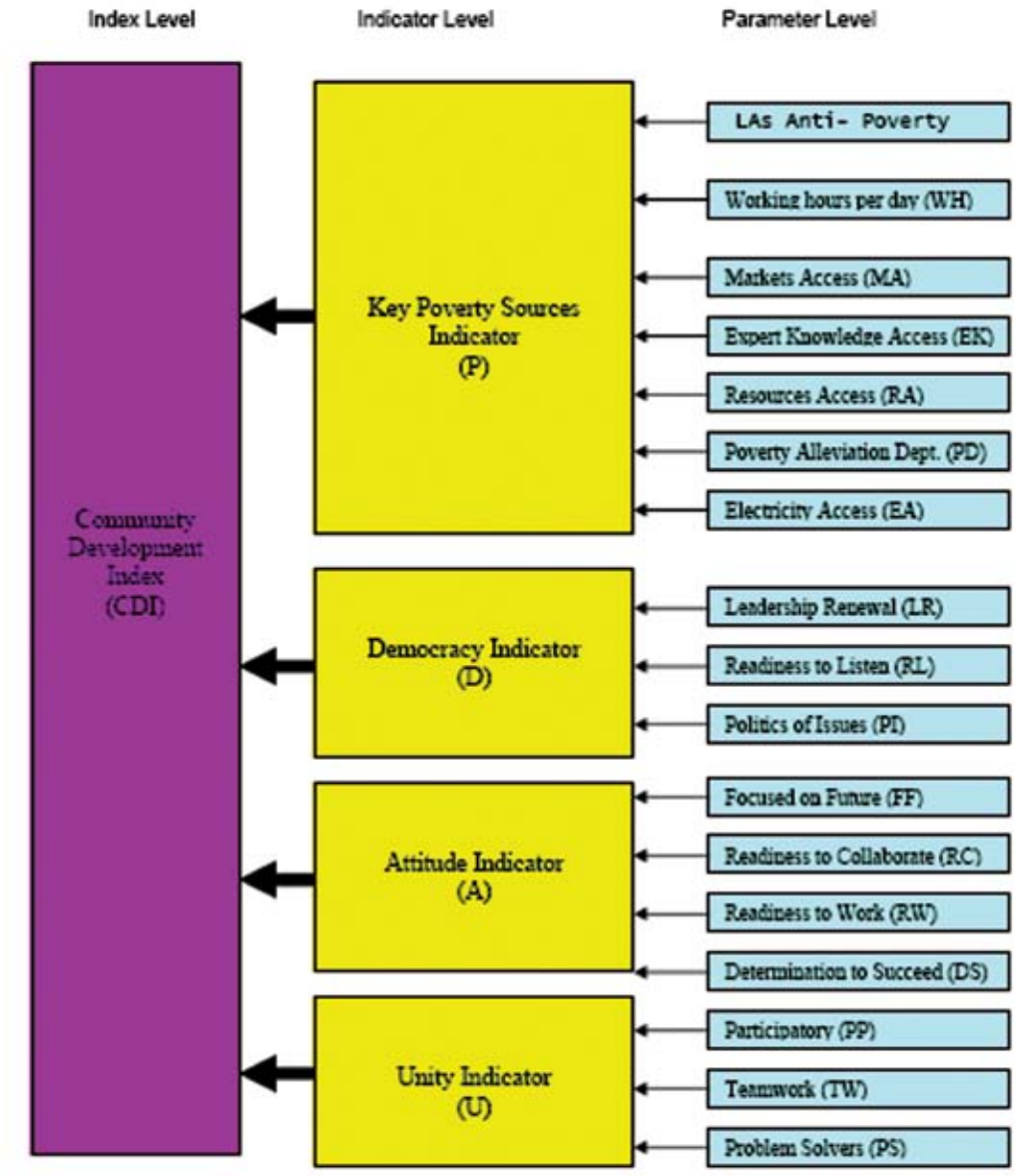
Majority of the poor and people languishing in poverty like idling and doing nothing even with the little resource available. In the case of Africa, to ensure that such people are mobilized, local authorities should pass anti-poverty By-laws requiring citizens to engage in gainful socio-economic activities. For instance, get people to register the activities they are engaged in every time they seek public service, especially if and when they are not formally employed. Once anti-poverty By-laws are in place, local authorities can enforce them with assistance of the provincial administration, the church etc. it would then be easy to identify and isolate idlers who engage in crime, robbery and other means frustrating investment, savings, and accumulation.

With such ICT-based CDI, leaders and citizens alike will be able to measure their own strengths and weaknesses, as well as recognize their opportunities. In order for communities to manage their own risks in a sustainable and responsible manner, it is imperative that they should subject themselves to technologies of self-examination and self-reflection so as to understand themselves, the risks they face, and their ability to manage those risks appropriately. Therefore the development of CDI and its availability and access, would serve the purpose of providing a mirror for communities to self-examine themselves. CDI may help create a distinction between communities that are unsustainable or 'at risk' and those that are comprised of 'active' and entrepreneurial citizens who manage their own risks in a 'healthy', 'responsible' or 'sustainable' way. Figure 2 shows the suggested different parameters and indicators that can be combined to create CDI.

Several methods are currently in use that can be applied to determine an index such as CDI. For instance, in order to produce a CDI index, four indicators shown in Fig.2 namely key Poverty Sources, Democracy, Attitude and Unit have to be summed using an algorithm. In our case, we are still at the stage of developing a meaningful theory that we shall base and justify a generalized algorithm that is applicable to any number of parameters and indicators for generating CDI.

.

Fig.2 : Fundamental Parameters and Indicators for the Community Development Index (CDI).



Conclusion

Technology works well as a powerful tool for human development and poverty alleviation when good and appropriate policies, regulators and a high degree of transparency in its deployment are exhibited.

The contemporary advanced liberal emphasis on 'bottom-up' or community-led development is indicative of the fact that the management of resources and the

attendant risks are no longer an activity of the state alone, but the responsibility of citizens and their communities wherever they may be.

ICT-based tool such as CDI can help leaders and their citizens alike to be able to measure their own strengths and weaknesses, as well as recognize their opportunities.

Therefore the development of CDI and its availability and access, would serve the purpose of providing a mirror for communities to self-examine themselves. CDI may help create a distinction between communities that are unsustainable or 'at risk' and those that are comprised of 'active' and entrepreneurial citizens who manage their own risks in a 'healthy', 'responsible' or 'sustainable' way.

An underpinning theory for a generalized algorithm that is applicable to any number of parameters and indicators for generating CDI needs to be developed. This will be achieved through case studies and the development of an appropriate frame work so as to illustrate the effectiveness of the CDI computation.

References

Association of Local Government Authorities of Kenya (ALGAK). (2004). *The Role and Place of Local Government in the East African Political Federation*. ALGAK Annual General Meeting held on December 16, 2004 at School of Monitary Studies. Nairobi.

Bratton Michael. (2003). *Support for Economic Reform? Popular Attitudes in Southern Africa.* World Development Vol.31,No.2,pp.303 –323,2003

Government of Kenya. (2003). *Economic Recovery Strategy for Wealth and Employment Creation 2003-2007*. Government of Kenya

Herbert-Cheshire Lynda and Vaughan Higgins Vaughan. (2004) From Risky to Responsible: Expert Knowledge and the Governing of Community-led Rural Development. Journal of Rural Studies 20 (2004) 289-303 Hjorth Peder. (2003). *Knowledge Development and Management for urban Poverty Alliviation.* Habitat Internation 27 (2003) 381-392.

Howard Wriggins and Gunner Adler-Karlsson. *Reducing Global Inequalities.*

Kim Pyungho. (2003). *In search of a private realm: a social perspective on Internet diffusion.* Technology in Society 25 (2003) 417–429

Kuyvenhoven Arie. (2004). Creating an enabling environment: policy conditions for less-favored areas. *Food Policy* 29 (2004) 407-429

McGinis B James. (1979). Bread and Justice: Toward a New International Economic Order

Ministry of Finance and Planning Republic of Kenya, Poverty Reduction Strategy Paper: *Report of the Sector Working Group on Information Technology*. September 2001

Odendaal Nancy. (2003). *Information and Communication Technology and Local Governance: Understanding the Difference Between Cities in Developed and Emerging Economies.* Computers, Environment and Urban Systems. 27 (2003) 585-607.

Odhiambo Walter, Omiti M John and Muthaka I David. (2005). *Qualitative and Quantitative Methods for Poverty Analysis*. KIPPRA

Olowu Dele. (2003). Challenge of Multi-level Governance in developing Countries and Possible GIS Applications. Habitat International 27 (2003) 501-522

UNDP .(2001). *Human Development Report 2001.* Making New Technologies Work for Human Development.

UNDP. (2004). *Human Development Report 2004.* Cultural Liberty in Today's Diverse World

41
Computational Resource Optimization in Ugandan Tertiary Institutions

Richard Ssekibuule, Joel Lakuma and John Ngubiri

Insufficient computational power, caused by limited budgets is a big challenge to high quality research and education in developing countries. Some institutions suffer severe shortages while others enjoy comparatively more resources. Many institutions therefore source for funds to procure more resources while paying less attention to optimally utilizing what they have or creating facilities for sharing with other institutions for mutual benefit. Current developments in Grid computing technology provide for mechanisms of consolidating isolated computing resources to provide a high power computing environment. In this project, we aim at studying ways (limited) computer resources can be optimally used in a financially constrained setting as well as ways for building an environment for providing high computing power for learning and research in a cost effective way.

1. Introduction

Information and Communication Technology (ICT) has revolutionalized ways in which businesses and day to day activities are run. This has been both at the internal unit/business operation as well as business internetworking to cater for resource sharing, communication, cooperation and other business needs. Graduates from tertiary institutions are expected to have adequate ICT knowledge related to the field of specialization and are as well expected to be adequately trained in ICT facilitated institutions. ICT is therefore both a medium and goal of tertiary training.

In financially constrained countries like Uganda (and other developing countries), there is hardly a tertiary institution with adequate ICT infrastructure to meet its demands throughout the year. Some institutions, individuals have idle resources during off peak seasons but also experience bottlenecks during peak periods. In other institutions, the shortages are experienced throught the year. This is caused by lack of money to procure adequate ICT infrastructure (and train the human resource). There is a high possibility of inadequate training of students, graduating with less skill hence not able to fully explore their potential at the work place. This leads to lower business productivity, low income, poor remuneration and low competitiveness. This vicious circle needs to be broken at the earliest opportunity to stimulate development and transform developing countries into developed countries.

To address the scarcity of ICT resources, tertiary institutions in Uganda have mainly put emphasis on sourcing for funds to procure more infrastructures to reduce the deficit. Less emphasis however has been put at optimal utilization of the existing ones, identifying specifications that can offer higher value for money in a certain situation or exploring inter-institutional resource sharing for mutual benefit.

In this study, we aim at:- (i) Investigating ways ICT facilities can be optimally utilized both with in and among tertiary institutions in Uganda; (ii) Investigating metrics institutions can use to procure ICT resources in a cost effective way in specific work and institutional environments and (iii) investigating ways different organizations can share ICT resources in a developing country environment. Overall, the research aims at improving the overall benefit per unit cost as means of increasing the relative ICT prevalence in tertiary institutions.

The rest of the paper is organized as follows; we present the motivation behind the inception of this study in Section II and work done related to it in Section III. We discuss the developing country specific challenges this research project is likely to encounter in Section IV. These challenges to a large extent make this project unique from related projects in the developing countries. We justify the project in Section VII and describe the overview of the research blocks in this project and make our conclusions in Section VIII.

2. Project Motivation

Despite the general fall in prices of ICT equipment and resources, many tertiary institutions in developing countries cannot afford them in adequate amounts. In fact, some resources like bandwidth and computer hardware are more expensive in developing countries than developed countries. While sourcing for extra funds can reduce the scarcity, we believe devising ways of using new and existing technologies that put the resources to optimal utilization can also create a substantial impact. It also generates more utility from the funds mobilized by the institutions. The following are the main factors that motivated the inception of this study/project.

A. Low utilization levels:

The individual utilization of a microcomputer is very low with resources like CPU usage hardly reaching 15% for average computer users. We need to investigate ways in which we can get optimal utilization of computer systems both inside individual tertiary institutions and across institutions and individuals in the region.

B. Cost and utility mismatch:

Application and operating system software is so dynamic (with windows changing nearly every two years and linux far less). Within a few years, the software is considered outdated. In many cases, a software upgrade is not effective since the hardware may not fit the minimum requirements of the newer software version. In most cases an operating system or application software upgrade might as well

necessitate a hardware upgrade or a total replacement. A long lasting and hence expensive computer may not be the best option in a financially constrained environment if the organization is to keep abreast with the software regime.

C. Software appropriateness:

In many cases, a computer that cannot work with specific software can work with another (especially less graphical software). This is because of the different resource requirements by different platforms. The cost of replacing the computers may be higher than the cost of acquiring a resource conserving software. Likewise, the software may have a lower rate of change in resource requirements that it can be used for a longer time on the same hardware system than another system whose resource requirement change more often. Additionally, computer hardware that could be considered 'old' for some platforms could be 'new' enough to support other software systems with lower system requirements. Old computers can therefore be reused and hence increase their utility.

D. Resource sharing:

With the existence of the Internet, Intranets and some commercial and private optical fiber networks, remote/grid computing can be a possible cost saving option. Different institutions can share or combine otherwise expensive resources like software and processors to achieve high computational power. Technological advancements in Grid computing [1] around the world provide a platform for a probable solution to large-scale resource sharing [2]. The Grid architecture can provide a means for combining low capacity inexpensive processors into a distributed mass of high capacity computational infrastructure.

E. Network for collaboration:

Institutions participating in resource sharing can easily enhance collaboration in their production or research areas that are within or outside the scope of this project for their mutual benefit.

3. Related Work

The work that we wish to carry out in this project builds on a foundation that has been created else where most especially in Europe. Several projects in Europe have been successfully implemented to facilitate optimal computation power utilization and research. In this section we present projects whose achievements and ambitions are similar to the ones in this work.

A. European Grid Computing

Through a grant from the European Union, an application test-bed for Grid computing [3] was implemented to provide a network of high performance computing within European counties. Additionally, the European Grid is used to provide support and a framework for developing Grid software among many other software and related scientific research problems.

B. Albertross

The Wide Area Cluster Computing project Albatross [4] was implemented in The Netherlands among four collaborating universities to help them investigate and understand application behavior on wide-area networks. A host of cluster computing research problems is being investigated in this project and a host of research publications have been realized as a result of the presence of this collaborative facility.

C. Particle physics data grid

The particle physics collaborative project [5] is used by physicists and grid computing scientists to investigate high energy problems and network infrastructures. This project has served needs for experimental physics and research requirements for computer scientists.

4. It Infrastructural Challenges

Unlike most of the related work carried out, this project is taking place in a technologically and financially different setting. It therefore faces some unique challenges in addition to those faced by related projects. These challenges may therefore call for different approaches. They may also render results from some related research inapplicable. While some make what would otherwise be obvious assumptions stringent constraints, they highly contribute to the originality and vitality of the study. Below, we briefly explain the nature and severity of the (potentially unique) challenges on this research.

A. Low Inter-network speeds:

Like in other developing countries, Ugandan institutions of learning and research are mostly connected to low speed Internet links. This may be a hindrance to remote processing especially when large amounts of data are to be transferred. It is also expensive for developing countries to upgrade speeds in network links.

B. Low availability of computer hardware and software:

Though the world price of computers is falling, it is still out of reach for many organizations and persons in Uganda. Considerations of the scarcity of money to procure computers and the scarcity of computers themselves in the organizations have to be put into consideration. Additionally, the total cost of ownership (TCO) remains high for Ugandan institutions. This is rarely taken into consideration by institutions when procuring computer systems.

C. Quest for high computational power in new applications:

High education demand in Uganda is increasing; this is coupled with research some of which needs high computational power (despite fewer computers). The need for high computational power in the effective day today running of the institutions will be highlighted in the research.

5. Opportunities For Research

We now discuss the items that are to make the units of this project. Just like in other research projects, new research ideas can come up leading to the deepening of the individual unit or creating another unit of research. Therefore we cannot guarantee the conclusiveness of the items just like we cannot guarantee the conclusiveness of the depth of each.

A. Utility maximization

In utility maximization, we shall seek to investigate how resources, on a single computer for example, can be put to maximum utilization. This can involve development of multi-user workstations in a local setting. It also involves investigating the optimal number of users that can run on a computer (as a function of time, application, user characteristics, etc).

B. Resource management and brokerage

This is to be in inter-unit (with in an organization) and inter-institutional sharing of resources. This is to ensure optimal resource usage, incorporation of owner policies in a shared environment, investigating ways multi agent can be used in a (micro/ mini) grid setting. This is to be done with in the constraints of a slow communication network, heterogeneous environment and relatively slower machines. We are also to investigate (multi) programming issues.

C. Security

Sharing and cooperative processing extends the computers to more threats to attacks and privacy violation. Sharing and coordinating use of computational recourses opens up new security challenges that can be grouped into three categories: (i) integration with existing systems and technologies in distribute virtual organizations [2], (ii) interoperability with different hosting environments (e.g., J2EE servers, .NET servers, Linux systems), and (iii) trust relationships among interacting hosting environments[6]. This project will give researchers in Uganda and others involved and opportunity to have hands-on research for the security challenges involved in coordination with experts in the same field.

D. Robustness

Factors like network, process or node failure cannot be ruled out. This may render the reliability of such a cooperation set up poor. More threats like exceptions and deadlocks are likely to happen. Handling them points to more research.

E. Additional system support

This is to involve development of middleware that can help in management of such cooperation network in the developing country set up.

6. Justification

Other than the increased availability of computational resources for training and research, there are other fields/organizations that can tap from this project to improve their efficiency. Other areas of research in Uganda that could tap into the computational abilities of this research work are:-

A. Support for epidemic research

Uganda is one of the key participants in Virus research and infectious diseases. Institutions that require this computation power already exist (like the Uganda Virus Research Institute, Join Clinical Research Centre) and their results and Mean Time to Deliver Results (MTDR) will be reduced by existence and utilization of a high computational infrastructure.

B. Support for other high performance applications

Current (and future) research centers in Uganda (will) need high but cheap computational power. These include weather prediction centers, support for disaster preparedness, and molecular biology research among others. Foster et al. [7] and Czajkowski et al. [8] provide examples to some of the services that can be developed to benefit from distributed resource sharing.

C. Distributed data management

In Uganda, like in many countries where there is decentralized governance, there is distributed data on a similar subject with remote data often needed for supervision and comparative purposes. This research will as well look at distributed data management which can be beneficial to such structures.

7. Project Structure

In this section, we present in chronological order the general steps that will be taken to realize successful implementation of this project.

A. Feasibility Study

In this phase of the project, we shall investigate the current state of computer availability, utilization trends, and connectivity resources and systems versioning among others. The researchers will establish the computational resources on the ground, numbers, specifications, institutional needs, gaps, bridging strategies, strengths, weaknesses and other related factors. By the end of this Phase, the researchers are expected to have information on:-

i) The current state of computational resources;
ii) The requirements of different institutions that will be expected to participate in the high power computing facility;
iii) The relationship between the ideal, reasonable and existing IT infrastructural levels;

iv) The extent to which IT has been incorporated in the management, administration and carrying out of business in tertiary institutions in Uganda;

v) The existing opportunities in IT resource optimization in Ugandan tertiary institutions and

vi) The possible threats that may be created as a result of consolidating computing resources will be investigated.

B. Main Research

The ability for this project to support and trigger a spin-off of research problems and experiences brings us to the discussion the key research components. The main research is divided into high and low level research. High level research is expected to deal with the organizational, policy and administrational issues while low level research will deal with the software development, performance and resource optimization techniques.

1) ***High level research:*** The themes to be considered in high level research include:

i) *Policy and Regulation issues;* Policies that will be used to regulate usage and distribution of computing resources will be done as high level research.

ii) *Specifications for middlewares*; This activity involves specification of the types of middleware and needs that are to be served.

iii) *Administrative & support frameworks*; This level of research in the project will define how administrative work for technological facilities will be supported.

2) ***Low level research:*** The themes to be included in low level research include:

i) *Middleware development / Deployment*; Middleware in a grid environment is supposed to provide abstraction of the underlying technical details of the hardware and operating systems from grid applications. Middleware development and deployment is key to successfully realization of grid services. InteGrade [9] is an example of middleware that leverages idle computing power for desktop machines. Grid middleware has also been used to provide transparent migration of non-grid applications to grid environments with applaudable success [10].

ii) *Resource scheduling*; Resources in a Grid system are distributed geographically in small components, which are consolidated into a single pool of large consolidated base. This pool of resources is then shared among formal or informal groups of individuals and organisations that are commonly known as virtual organisations [8] [2]. Resources in such distributed network systems need to be well managed in order to have efficient and effective support for demand and resource utilization with

economies of scale being put into consideration as illustrated in the research work done by Buyya et al. [11]. The high power computing facility will provide and opportunity for researching scheduling challenges on low speed networks. Little work has been done on low speed networks in a high computing facility, particularly because most high power computing systems are in industrialized countries where bandwidth is not a big problem like in developing countries, especially in Africa.

iii) *Transaction & deadlock management*; In a high power computing facility where jobs are being submitted by several users and organisations that may want results in the shortest time possible, deadlocks and transaction bottle necks are most likely to occur. Transaction and deadlock management will definitely be an interesting research problem for academicians who will be involved in the project. Transaction and deadlock management has been investigated by several researchers [12], [13]. We need to investigate and evaluate these solutions for wide area cluster of heterogeneous systems and probably improve on them to suite any unique challenges that we shall be facing.

iv) *N-tier service brokerage;* Most distributed applications have been built on multi-tier technology in order to support scalability in service resources and provision of simplified administrative tasks. Research work by Shan et al. [14] presents recent technologies like logical solution architecture, process choreography, business rule engine, enterprise service bus and service composition among others in service oriented solution architectures for mainly web applications. These technologies can also be investigated for application in the grid systems. Research in N-tier service brokerage will be useful for leveraging high power computing resources for grid applications.

v) *Quality of service on communication channels;* For the kind of networks that exist in Uganda and the East and Central African region, quality of service research really needs to be done whilst paying attention to the low speed networks. Consolidation of idle CPU over low speed networks and probably accessing these resources from a few high speed fiber links that exist in the country is expected to give us interesting challenges for research.

8. Conclusion

The wide attention that has been given to Grid computing by researchers around the world provides a stimulating opportunity for scientists in Uganda to actively participate and contribute to the growing technology. We hope that by bringing the high power computing facilities closer to researchers in the country at relatively affordable financial and administrative costs, will help to enhance research in the region. This project will be a brave step towards bridging the technological gap

between a developing country Uganda and the industrialized economies. The successful implementation of an inter-institutional Grid will provide computer scientist with an opportunity for hands on research. Other scientists outside computer science will have an opportunity to research in fields that were previously impossible without a high power computing resource.

References

Buyya R., Abramson D., and Giddy J. An Economy Driven Resource Management Architecture for Global Computational Power Grids. The 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000), Las Vegas, USA, June, pages 26–29, 2000.

Czajkowski K., Fitzgerald S., Foster I., and Kesselman C. Grid Information Services for Distributed Resource Sharing. 10th IEEE International Symposium on High Performance Distributed Computing, pages 181–184, 2001.

Daniel Mallmann. Application Testbed for European GRID computing. http://www.eurogrid.org/, 2004.

Foster I., Kesselman C., Nick J.M., and Tuecke S. Grid services for distributed system integration. Computer, 35(6):37–46, 2002.

Foster, C. Kesselman, and S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications, 15(3):200, 2001.

Goldchleger A. , Kon F., Goldman A., Finger M., and Bezerra G.C. InteGrade: object-oriented Grid middleware leveraging the idle computing power of desktop machines. Concurrency and Computation: Practice & Experience, 16(5):449–459, 2004.

Grid.org. Grid computing projects. http://www.grid.org/home.htm, 2004.

Nagaratnam N., Janson P., Dayka J., Nadalin A., Siebenlist F., Welch V., Foster I., and S. Tuecke. Security Architecture for Open Grid Services. GGF OGSA Security Workgroup. http://www. ggf. org/ogsa-sec-wg.

Particle Physics Data Grid. http://www.ppdg.net/, 2006.

Takemiya H., Shudo K., Tanaka Y., and Sekiguchi S. Constructing Grid Applications Using Standard Grid Middleware. Journal of Grid Computing, 1(2):117–131, 2003.

Tang F. L., Li M. L., and Cao J. A Transaction Model for Grid Computing. Proceedings of the Fifth International Workshop on Advanced Parallel Processing Technologies, pages 382–386, 2003.

Thilo Kielmann. Wide area cluster computing. http://www.cs.vu.nl/albatross/, 2002.

Turker C., Haller K., Schuler C. and Schek H.J. . How can we support Grid Transactions? Towards Peer-to-Peer Transaction Processing. Proceedings of the Second Conference on Innovative Data Systems Research, CIDR 2005, pages 174–185, 2005.

W. Bank and CTS Inc. Solution Architecture for N-Tier Applications. Proceedings of the 2006 IEEE International Conference on Services

42
Security Analysis of Remote E-Voting

Richard Ssekibuule

In this paper, we analyze security considerations for a remote Internet voting system based on the system architecture of remote Internet voting. We examine whether it is feasible to successfully carry out remote electronic voting over the existing Internet infrastructure that conforms to the requirements of a public election process of; integrity, anonymity, confidentiality, and intractability. Using the Delov-Yao threat analysis model, we perform a threat analysis on the system and highlight limitations of existing solutions to mitigate these threats. We present suggestions on how to improve some of the proposed solutions in literature and we end with our opinion about technical feasibility of remote e-voting over the Internet and the future work.

1. Introduction

The Internet has transformed the way we live, interact, and carry out transactions. Traditionally, physical contact among parties in a business transaction was used to enhance trust. But in e-enabled service, trust is built based on algorithms that define authenticity of parties and maintain their confidentiality preferences as in the natural world. The advent of online shops like EBay, Froogle, and Yahoo, online educational programs, telemedicine among others have gone a long way in enforcing the belief that the future is "< >" i.e., most of human transactions can be carried out safely at the click of a button. Although the pioneers of online services have experienced problems as regards to security, privacy, anonymity, and usability, the large amounts of transactions that are being carried online (worth about 85 billion dollars annually for e-commerce only) [22] is a good indicator of how important the Internet is to the modern societies. To continue harnessing the possibilities that the Internet can offer to human societies, researchers have proposed a number of ways to implement online remote voting over the Internet so as to enhance democracy [1]. It is through democracy that liberty and freedoms are entrenched in the human societies which are vital components for economic prosperity.

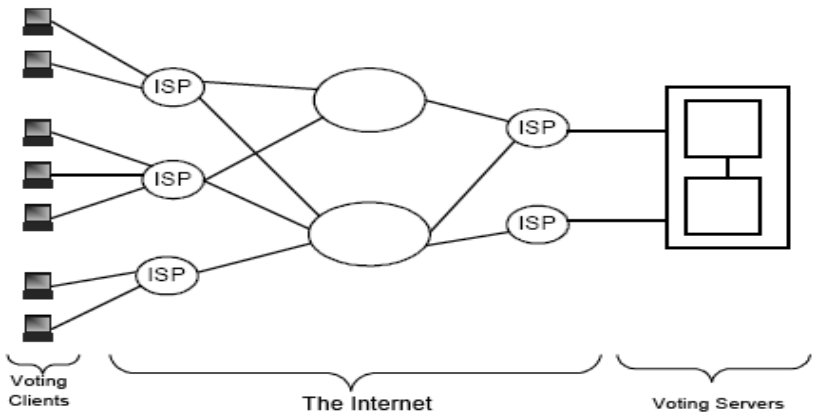
In this paper we use the phrase "Internet voting" to refer to electronic voting (e-voting) over the Internet. Unlike traditional voting systems in which voter choices and intentions are represented in form of a paper ballot or other means like a punch card, Internet voting uses electronic ballots that are used to transmit voters' choices to electoral officials over the Internet. Internet voting can be can be categorized into three forms [1, 24] that are described below.

- **Poll-site Internet voting;** in this system, voters cast their ballots from a number of designated polling stations. The controlled physical environment

at the polling site offers more possibilities of managing some security risks. Poll-site Internet voting offers more convenience and efficiency than traditional voting systems.

- Kiosk Internet voting; is similar to poll-site, but voting machines are placed away from traditional voting locations and could be set up in convenient places like schools, libraries, and malls. Like poll-site voting, kiosk voting would make it possible to manage some security risks by controlling the physical environment.
- Remote Internet voting; this scheme allows voters to cast ballots from practically any where in the world as long as they have access to the Internet link. While this offers tremendous convenience, it also introduces several potential security risks because the physical voting environment is not controlled. Issues of intimidation, voter impersonation among other due arise. Figure 1 presents a generic remote Internet Voting architectural diagram.

Fig. 1: Internet Voting Architecture



The voters can cast their ballots using *client computers* that are connected to the Internet through *Internet Service Providers (ISP)* that link the client computers to voting servers.

This paper focuses on the challenges of implementing a viable remote e-voting system. We discuss the different threats this system faces to deliver a credible election result and the current approaches to mitigate these threats. We presented the limitations to the proposed mitigation measures in literature and propose improvements on these schemes. We voice our opinion weather remote e-voting can deliver a viable election result and discuss the cost benefit analysis of such an election. We end the paper with future research directions in e-voting systems.

The rest of the paper is organized as follows; in Section II we present basic requirements of a remote e-voting system and its components. In Section III, we present an analysis of threats to the remote e-voting. In Section IV, we analyze the

proposed threat mitigation schemes in literature. In Section V, we give a summary of our analysis with suggestions of improving the proposed threat mitigation schemes in literature and Section VI presents concluding remarks and future research direction.

2. Remote e-Voting System

Just like any other system, the remote e-voting system is made up of a number of components and has stringent requirements to meet. In the proceeding Subsections, we present the requirements and the building blocks of an e-voting system.

A. Requirements

Like the traditional voting system that ensures that only registered voters participate in the voting process, and that a voter can only cast one ballot, and that the vote is cast in privacy without unlawful influences and that the voting process is transparent to all interested parities. Remote e-voting systems are also expected to provide a platform of conducting a fair and transparent election. Based on Neumann et al. [2] work, we summarize e-voting system user and functional requirements in the proceeding Sections.

1) *Ease of use*

   For e-voting systems to again acceptance, the systems should be user friendly, i.e., requiring less time to learn and operate. Users naturally desire a new system to be more user-friendly than the one being replaced [27]. A system that is functionally sound but with poor usability can be a cause of errors (on the voter's side) during and electoral process. Often times, system developers focus more on system functionalities and the expense of usability [27]. This for an e-voting system could lead to low voter turn and voters feel their time is precious to waste in learning a system that does not direct add value to their lives.

2) *Authentication*

   Authentication is very important to maintaining overall security of the system. Strong authentication mechanisms have to be implemented to grant access to authorized users and to keep out intruders in order to maintain system accountability. Additionally, mutual authentication has to be provided to protect voters from providing their security credentials to rouge servers. Most of the cases in which phishing scams are successful occur because mutual authentication is not provided because users are not enabled to authenticate servers that they are connecting to [26].

3) *Integrity*

   The integrity of votes cast and the entire voting system hardware and software should be maintained. As stated by Roy et al. [3], the counting process of votes should produce reproducibly correct results. Integrity is an important requirement that requires servers and client computers being free of trapdoors

and any other forms of internal threats that could cast doubt on the safety of the voting system.

4) *Voter anonymity*

Voters should be able to cast their votes without being traceable as is the case in traditional manual election process. The voting system should not link a cast vote to a voter [3, 18]. Failure of the voting system to provide anonymity would mean that interested parties could trace and know that someone did or did not vote for a given individual or policy. Such act would endanger the voters and compromised the fairness of the election results.

5) *System accountability*

The voting system should be transparent enough to allow accountability by interested parties in case of disputes [18]. Accountability is important for defusing disputes regarding voter complains which could involve wrongly registered votes or incorrect tallying. An audit trail that does not link a voter to a cast vote is desirable in case one wishes to know if their vote was counted.

6) *Confidentiality of the vote*

The system should maintain the confidentiality of the votes during and after the voting process. This is very important to align any fear of votes tampering.

B. System components

1) *Voters*

These are persons registered with the system, with the rights to participate in the election. They are a critical component of the remote e-voting system as most of the feasible security breach can occur at this level. Of course, these voters are expected to be humans who are registered and authorize to participate in the electoral process.

2) *System administrators*

These are persons with the authority to operate the voting system. System administrators undertake tasks of installation, upgrade and application of security patches and have privileges to access both physically and logically all components of the voting systems except client computers. .

3) *Client computers*

These are end user terminals that are remotely connected to the voting servers over the Internet from which voters cast their votes. They run generic software's and are highly vulnerable to logical attacks.

4) *Network infrastrature*

This is mainly comprised of communications media that connect the internet service providers (ISP) to the client computer, ISP gateways, interconnecting servers, layer three switches among others. The communications media consist of fiber networks, Ethernet cables, telephone lines, and wireless medium.

5) *Voting server(s)*

Voting servers are part of the Trusted Computing Base (TCB) of the voting system. A trusted computing base is that part of the system that is responsible for enforcing security policies [25] they are strategically located in the system for faster access at low risk of compromise. Normally they are physically located is a secure environment at the election organizers premises.

6) *Voting protocol*

The voting protocol is another key element in the system. The protocol governs the logic that handles security of the ballots, registration of users, authentication of participating parties, verification of votes cast and vote counting. We can as well say that the voting protocol is the heart of the voting system, without which all the designs are fruitless. The remote e-voting system requires a voting protocol that can guarantee confidentiality, integrity, and authenticity of the votes. To our knowledge, the most complete protocol so far written but with no trace of implementation was designed by Indrajit and Indrkshi [4]. This protocol is largely based on the work of prominent researchers like David Chaum [21].

7) *Voting System Software*

This is a crucial component of the voting system that has the actual implementation of the voting protocol and services that are needed in the voting process. Apart from software that exists in the network devices like routers and switches, other software components on both the client and server side have to be customary developed for the voting process. Usually, a large component of the voting system is executed from the server side and a thin client made available via network connections to clients. A secure communication between the client and server software is always expected to be provided to keep out adversaries.

3. Threat Analysis

In the following sections we present threats to the remote e-voting system. The threat can be categorized into: technical and social depending on the schemes of attacks and target components.

A. Trapdoors

This is a technical threat and previous research work by Roy et al. [3] has highlighted the danger of maintaining trapdoors. Software developers and system administrators usually create accounts that are usually not known to normal system operators (trapdoors). These accounts are used for trouble shooting purpose and at times for achieving personal goals. However, skilled hackers also obtain these accounts and even create other trapdoors which are more difficult to close or detect for their future use. Trapdoors can exist in any software that runs on a computing device. The software can be a web browser, web server, application server, word processor, a favorite screensaver among others.

B. Virus attacks

Protection against virus attacks is not a trivial in a large election in which voters use their home computers to cast their votes. It is very hard to ensure that users do not have viruses on their computers that could do something unexpected in the polling day. Most of the attacks on computer vulnerabilities are very stealth and sophisticated for an average computer user to predict or detect. The most notable user exploits are those that attack email clients like Microsoft Outlook and Outlook Express. Some of these viruses don't require the user to open an attachment or an email in order to infect his/her computer. In Outlook Express, a virus can activate even if the e-mail is only viewed through the Preview Panel [6]. Attacks during a major election are expected to be more subtle than the more famous script kiddies' attacks. Probably people who write script-kiddies maybe the ones involved for malicious intent. A nation wide election in any country is most likely to attract the attention of state enemies who may be willing to invest enough resources to employ highly skilled crackers to sabotage the voting process. This is a technical threat as well.

C. Phishing scams

Through social engineering and intimidation, eligible voters can be led into giving away their security credentials to criminals who might want to influence the outcome of the voting process. Some phishing scams deploy rogue websites that appear like genuine ones and are used by attackers to get credentials illegally from voters. This threat can be classified as either technical or social depending on the mechanism of attack used. When software is used to confuse the user into thinking that the presented interface is genuine, then a technical phishing scam is said to be used. On another hand, voters can be conned by individuals into giving away their voting credentials; in which case a social phishing scam is said to be used.

D. Compromise of voter's privacy

System designs that keep audit trails of the voting process (that can later be used to link voters to their votes) are a source of compromise to voter's privacy. If this is done, voters who are sensitive to their privacy can choose to abstain from the voting process for fear of their safety, hence influencing the electoral system.

E. Subverting System Accountability and Integrity

Though subverting system accountability is non-trivial for a well designed remote e-voting system, it still remains a threat for especially from internal organizational administrators who may take advantage of their system privileges and tamper with audit trails of the system. An attack on system accountability could be launched from the client software, where by a supposedly cast vote is either dropped or registered with changed voter intention and then tallied on the server side in accordance to the desires of the intruder. Additionally, the tallying process on the server side of the voting system can also be tampered with to favor given subject or candidate.

F. Compromise of client computers

Current research [7] has revealed that there are wider spread and reported vulnerabilities in Windows systems compared to other operating systems like UNIX, Linux, and MacOS. Almost all internet applications on Windows operating system (OS) have at one time contained security vulnerability. The continued discovery of buffer overflows in several windows systems including most internet applications is quite a big problem. This causes a big threat to remote voting over the internet, since above 90% of internet users are running windows operating systems.

Most of these flaws are known to cracker communities and can be easily exploited in a public election to interfere with the voting process in various ways (DDOS being the most likely) [8]. Since most people use windows systems with popular applications like e-mail clients, chat tools, office suites, document views like Adobe and others, a group of people from these companies can easily install a backdoor or a Trojan-horse inform of an update which can go quite unnoticed to many people as illustrated by Ken Thompson [9]. The effect of such subversion could render client computers unusable for a while during an election day, or redirecting them to dummy web server.

4 Analysis of Proposed Mitigation Schemes

In an effort to mitigate the above mentioned threats, researchers have proposed a number of mitigation controls and in the following paragraphs we summarize some.

A. Solution to mitigate authenticity

In literature [16, 18], researchers have suggested that physical and logical access to the voting systems should be based on credential and rights granted either on role based or need to know policy. Voters and administrators must gain access with nontrivial authentication mechanisms that may require use of smartcards [26] for stronger security. However, some authentications schemes which offer a strong authentication require either a user to memorize complex credentials or they are technically expensive in monetary and privacy terms. This is because users may be required to buy end user authentication devices like cryptographic calculators and biometric readers; additionally, transfer of biometric data over public networks raises privacy concerns on the side of users.

B. Virus attacks

Research indicates that sensitizing users into knowing the dangers of keeping update versions of software and being careful on the type of software they install on their computers can tremendous reduce the risks [7]. Though most antivirus software is commercial, there are also non commercial versions of software that voters could use before a voting process to ensure that their computers are free of viruses. However, these problems cannot be easily solved for all client computers participating in an election where people are voting from their homes.

C. Solution to Phishing Scams

Social phishing scams can be prevented through educating of people with detailed information of various means through which they (voters) can be exploited [13, 18]. However, this requires that the educators themselves keep updated with current methods of exploitation. Otherwise, the taught methods of attack and defense for the voters could be out dated and could still leave the voters vulnerable to social phishing scams. More importantly, technical phishing scams are more dangerous that social ones, since their effect can be easily wide spread in an election process. However, the solution equally solves the problem on a wide scale. Strong authentication is required in the voting system by means of mutual authentication. Mutual authentication schemes require the clients to be authenticated to the server software, and the server software also authenticated to the client. In that way, voters protected from technical phishing scams.

D. Solution to integrity threats

System changes must be prohibited throughout the active stages of the election process. Voting systems need to be verified by independent non partisan bodies that will look at the source code and verify that it does exactly what it was designed to do. The use of cryptography exchange of messages can guarantee integrity of information exchange. Indrajit and Indrkshi [4] developed an algorithm that protects voters' votes by use of cryptographic keys, in which it is not possible to link a voter to a vote unless the voter has cooperated. The requirements of vote secrecy and voter anonymity has not been a problem in itself, but achieving both of them (secrecy and voter anonymity) at the same time has been a problem to vote accountability and dispute resolution after voting process.

E. Subverting system accountability (voting server)

Although in some literature[2, 18], researchers have advocated for use of encryption and checksums on audit trails to help in detecting changes to file system audit trails, additional use of audited open systems code on the server environment can also minimize the risks of running source code with undesirable side effects [23].

F. Network infrastructure

Through redundancy, use of cryptograph, and the concept of honey spots, attacks on network infrastructure can be minimized. However, we note that it is fairly difficult to prevent some attacks along the communication channels like Denial of service (DDOS) [8].

G. Legal Protection

Attacks on mission critical systems in countries like the USA, UK and Brazil are being handled as criminal cases [11], [12] for which culprits have to be prosecuted. The act of hackers/crackers gaining unauthorized access to computer system can be compared to someone breaking into a house as a means of checking whether it is secure.

Microsoft is also putting a lead in this pursuit with over 100 law suits outside the USA [13] and it serves to protect electronic systems in the same way the law protects houses from bugler attacks. Without legal prosecution, then many attacks on systems will continue to be tried out and eventually some will succeed. This behavior has to be controlled legally, so that security checks can only be done by legally accepted organizations such as certified security organizations, but not any underground team of hackers who might have malicious and personal goals. Of course some sophisticated attacks can go unnoticed and other non-traceable attackers could launch successful attacks without being punished for their wrong doings. This is why security of a system cannot be left to legal protection and prosecution alone. System stake owners need to do all they can to keep the voting system technically sound.

H. Open Source Systems in Electronic voting

In literature [9, 14, 15] a concept of using open source systems for e-voting has been proposed. The debate rages on whether it is a good idea to have open source systems powering electronic voting over the Internet or not? The question of whether open source systems can be trusted more than closed source systems still stands? A Ken Thompson in his paper entitled "Reflections on Trusting Trust" indicates you can't trust code that you did not totally create yourself [9]. The paper by Ken presents and ingenious piece of code which can be used to create another program from itself in a way that is not easy to detect my non sharp-eyed programmer. Software written in a similar comportment can be used to introduce trap and back doors in an application.

The question of trust cannot certainly be left unanswered for an important democratic exercise like voting. People need be assured that there are no uncertainties regarding security for the systems that has been deployed. Experience from exposed vulnerabilities in closed source system has showed that closed systems cannot be thought of as being more secure than open systems. The most common example is of windows operating systems, where much vulnerability have been uncover by independent security experts working without access to the source code. This is not to suggest that open source systems are bullet proof, it rather shows that vulnerabilities can uncovered or even easily exploited in closed systems.

Bruce Schneider author of Practical Cryptography and one of the foremost experts on cryptography explains in his article on voting systems [14], that security is almost always in the details of the rest of the system; where by a secure system is only as strong as its weakest link. The biggest weakness of these companies (that keep closed source) is the need to keep the source code secure in order to keep the system secure. The analysis by Tadayoshi et al [15] provides an example of how vulnerabilities can be discovered in source code by someone who is not the author of that source code. In the analysis done February 2004, on AccuVote-TS electronic voting system, lots of problems, including unauthorized privilege escalation, incorrect use of cryptography, vulnerabilities to network threats, and

poor software development processes were identified. It was also discovered in the analysis done on AccuVote-TS voting system by Tadayoshi et al [15] that without any insider privileges, voters could cast an unlimited number of votes without being detected by any mechanism within the voting terminal software. In the AccuVote-TS systems, smartcards were not performing any cryptographic operations, giving way for forged smartcards to authenticate themselves. The system was found so insecure that even ballot definitions could be changed and even voting results modified by persons with forged credentials.

We note that most developers may know what is required to be done, but because of project time demands and sometimes because they do not have many people watching what type of code they are writing, many of them end up coding in undesirable styles leaving behind undocumented features. Open source developers are always aware that many people will be reviewing their code so developers do their best to have the best output.

5. Analysis Summary

Our study of remote e-voting has revealed quite a number of important critical issues that are summarized in this section. A trusted computing base (TCB) is a primary requirement for secure electronic voting over the Internet but building one is one fundamental challenge researchers are still facing.

Internet voting system cannot guarantee security to users voting from their computers operating in an insecure environment. The presence of viruses, untrusted user computer applications from various vendors and phishing scams, renders client computers vulnerable to thousands of attacks. More expensive measures can be taken by providing voters with cryptographic calculators and smart cards to provide an improved security to the client side of the TCB. However, problems concerning more subtle attacks like Distributed Denial of service (DOS) [8] attacks do not have a solid solution yet. Also, fundamental and original design flaws in Internet protocols can create an open door for quite a large number of security exploits.

DNS spoofing is a security threat that involves voters being redirected to a different server from a genuine one. This attack can have several impacts on the results of an election. Voters could be made to think that they are voting for the correct person among the candidates, yet they are voting for a dummy candidate. DNS spoofing that targets demographics that are known to vote for a particular party or candidate can negatively impact on the results of their total votes.

Buffer overflows can be exploited in poorly designed systems to alter the trend of the election. The ability for DDOS attack to be launched for a particular domain name can end the whole story of a voting process in quite a short time. Apparently the current implementation of raw sockets in windows XP has simple opened gates of possibilities for DDOS attacks. The experiences in 2003 of SCO going offline due to DDOS showed the world that more very sophisticated attacks that are not easy to filter can actually bring down a targeted network[10].

Trust is still a very big problem in electronic voting software. Apart from trusting electronic voting software, the compilers that were used for these programs/systems also need to be trusted. Presence of a Trojan-horse in widely deployed systems can alter results of an election in favor of some candidates. Open source systems and public scrutiny of source code will help in buying voters' trust in electronic systems.

Using of security independent bodies like universities and accredited security organizations to perform source code analysis for vulnerabilities will enhance the quality of source code for mission critical systems. Most of the vulnerabilities in software also arise from poor programming principles which are rather difficult to completely eliminate for programming languages like C and C++. Using a type safe language like Java helps in avoiding buffer overflows that are common in C and C++ programming languages. As indicated in the software evaluation report by Kohno et al. [16], the choice of a programming language can either lead to an increase or decrease of vulnerabilities in a system. It is easier to unknowingly introduce a bug in a C or C++ program that could be easily exploited with a buffer overflow as compare to Java or a safe dialect of C like Cyclone [17].

Possibilities of coercing voters into choosing different candidates, most especially on Election Day is a big problem to remote e-voting. Additional issues of voters' coercion, vote selling, vote solicitations that are discussed by Avi et al. [18] put remote e-voting into question, since these problems do not have solid solutions. As much as security and technological details of Internet voting systems can be perfected to an appreciatable degree, there is no clear solution as far as we know regarding vote selling if people are allowed to vote from home, or even coercion of a voter into choosing a candidate against one's choice. There must be a trade-off in any voting protocol between security and simplicity of voting as discussed in [19, 20]. In order to ensure voter trust and legitimacy of election results, all levels of Internet voting process must be observable. Because fair elections and elections perceived to be fair, are important targets in any voting system. The use of open source systems can help in buying trust of citizens; since code reviewed publicly will most likely not have unfair operations.

6. Conclusion

Our work has revealed that, public analysis of systems improves security and increases public confidence in the voting process. If the software is public, no one can insinuate that the voting system has unfairness built into the code. Proliferation of similarly programmed electronic voting systems can escalate further large scale manipulation of votes. It is very hard to guarantee security of a remote e-voting system, in an environment that cannot be explicitly controlled by the voting regulatory body. All technologies are useful only if they are used in the right way. In the AccuVote-TS voting system [15] provides a clue of how a poor usage of cryptography rendered a supposedly secure system to be flawed. Open source systems and peer reviews can help solve the problem. Independent

bodies study and evaluate systems for errors, security and design flaws. The technological advancements of e-commerce services that were never expected to be an on-line success, is a good indicator that in future we may have trusted remote voting systems. Using experimental prototypes in small election cycles will help in preparing e-voting for large scale public elections. The challenges that face Internet voting systems are not quite severe to prevent them from being used. Just like any other systems - even manual ones - that may have weakness and problems that need to be solved, Internet voting provides lots of more flexibility as compare to traditional methods of voting. The infrastructure is also relatively cheaper to maintain, considering that it is built upon existing systems that are used in everyday life of voters.

A desirable voting system should be accessible to all potential voters. In some societies like in the developing countries, not all voters have access to a computer and Internet. In fact a good number of them do not have knowledge of computer usage and the Internet. In such cases, the Internet can be used as an option to improve voter turnout. However, if the election is only facilitated by Internet voting, then the technology would end up becoming a barrier to voter participation.

References

Bruce Schneier. The problem with electronic voting machine. http://www.schneier.com/blog/archives/2004/11/the problem wit.html, accessed on Nove

Bubbleboy virus. http://www.exn.ca/nerds/19991112-04.cfm.

Caida. Code red. http://www.caida.org/analysis/security/code-red/. Accesed on August 10th, 2006

Caida. Denial of Service Attack on SCO. http://www.caida.org/analysis/security/sco-dos/. Accessed in July 2006

Chaum D. , www.chaum.com, accessed on 20 August, 2006

Comiskey, D. 2006 Online Retails Sales to Hit $100 Billion http://www.ecommerce-guide.com/news/research/article.php/3598281 , accessed on December 4th , 2006

Gibson, Steve. Distribute Denial of Service Attack. http://www.grc.com/dos/drdos.htm.

Jefferson D., A.D. Rubin, B. Simons, and D. Wagner. Analyzing internet voting security. Communications of the ACM, 47(10):59–64, 2004.

Jefferson, D., Rubin, A.D., Simons, B. and Wagner, D. A Security Analysis of the Secure Electronic Registration and Voting Experiment (SERVE), New York Times (http:// www. servesecurityreport. Org, accessed on December 19th, 2006

Jim T., G. Morrisett, D. Grossman, M. Hicks, J. Cheney, and Y. Wang. Cyclone: A safe dialect of C. USENIX Annual Technical Conference, pages 275–288, 2002.

Kohno T., A. Stubblefield, AD Rubin, DS Wallach, and UC San Diego. Analysis of an electronic voting system. Security and Privacy, 2004. Proceedings. 2004 IEEE Symposium on, pages 27–40, 2004.

Marc Friedenberg, Ben Heller, Ward McCracken, and Tim Schultz. "Evoting System Requirements: An Analysis at the legal, Ethical, Security, and Usability levels" www.marcfriedenberg.com/wp-content/evoting.pdf Accessed on Feb 16th, 2007.

MM Puigserver, JLF Gomila, and LH Rotger. A Voting System with Trusted Verifiable Services. Lecture Notes in Computer Science, pages 924–937, 2004.

National Science Foundation, USA. Internet Voting is no "Magic Ballot," Distinguished Committee Reports. http://www.nsf.gov/od/lpa/news/press/01/pr0118.htm, 2001. accessed on August 12th, 2006

Neumann P.G. Security criteria for electronic voting. 16th National Computer Security Conference, 1993.

Periklis akritidis, yiannis chatzikian, manos dramitinos, evangelos michalopoulos, dimitrios tsigos, nikolaos ventouras. Lecture Notes in Computer Science Springer-Verlag GmbH, Vol 3477/2005, pp 42

R.C. Hollinger and L. Lanza-Kaduce. The process of criminalization: The case of computer crime laws. Criminology, 26(1):101–126, 1988.

Ray I. and Narasimhamurthi N. An anonymous electronic voting protocol for voting over the internet. Proceedings of the Third International Workshop on Advanced Issues of E-Commerce and Webbased Information Systems, 2001.

Rubin A. Security Considerations for Remote Electronic Voting over the Internet. Comm. of ACM, 45:12, 2002.

Saltman Roy G. Accuracy, integrity, and security in computerized votetallying. Communications of the ACM 31, 10 October 1988.

Sinrod Eric J., Caution on Net Voting, http://www.computerworld.com/governmenttopics/government/legalissues/story/0,10801,59077,00.html , April 02, 2001 accessed on Dec 19th, 2006

Sun, H.M , An efficient remote use authentication scheme using smart cards,, IEEE Transactions on Consumer Electronic Vol 46/4, pg 858—961, 2000

Symantec security response. http://securityresponse.symantec.com/ avcenter/security/Advisories.html.

System Management Concepts: Operating System and Devices, First Edition (September 1999) http://www.unet.univie.ac.at/aix/aixbman/admnconc/tcb.htm , accessed on Feb 16th , 2007.

Tavani H.T. Defining the boundaries of computer crime: piracy, breakins, and sabotage in cyberspace. ACM SIGCAS Computers and Society, 30(3):3–9, 2000.

Thompson, . Reflections on trusting trust. Communications of the ACM 27, 8 (Aug. 1984); www.acm.org/classics/sep95.

Tom Espiner. Microsoft launches legal assault on phishers. http://news.zdnet.co.uk/0,39020330,39258528,00.htm. Accessed on November 20, 2006

43
E-government for Uganda: Challenges and Opportunities

Narcis T. Rwangoga and Asiime Patience Baryayetunga

The government of Uganda drew an e-government strategy aimed at changing the design operation and culture of the public sector to better respond to the needs of Ugandans. Some flagship programmes were included in the strategy as an opportunity for the Uganda Government to consolidate the position of ICTs in the country. These flagship programmes have been ongoing for long and according to survey reports, some of the programmes have only partially succeeded. Others are reported to have totally failed. This paper looks at what these programmes are and through interaction with personnel in the institutions concerned with the programmes and reviews of documented reports, discusses the underlying challenges faced by ICT initiatives in Uganda. Recommendations that can assist planners when designing ICT programmes are presented. These recommendations are aimed at improving the design of ICT programmes to minimize programme risks. In conclusion, we highlight one of the check tools used in ICT project planning that can be used to identify key factors that e-government project planners must address if ICT projects are to succeed.

1. Introduction

According to Bretschneider S. (1990)[1], there is insufficient evidence to suggest a direct link between ICTs and development. Danziger, J. N. &. Kraemer, K. L (2006) [2] argue that recent studies have found a positive correlation between investment in ICTs and economic growth in developed countries, but evidence for developing countries is not as extensive. However, it is emerging that ICTs in Uganda have been identified as a major tool for achieving socio-economic development by the Government of Uganda. In order for the government to implement the long term national development programmes timely, relevant information must be available at all levels of implementation. However, despite the government's will and mandate towards advancement of ICTs growth in the country, there are many limiting factors within the environment that have slowed realizations of the good intentions. For instance, lack of adequate funding to invest in ICTs, poor network infrastructure, and unaffordable ICT services for the citizens are among the limitations often cited. This study was conceived in part to look at the existing ICT initiatives in Uganda with a view of identifying areas that require special attention. The paper was prepared mainly from document analysis and the researchers' experience with ICT projects in Uganda. The Internet was also used to search for current trends of implementation of ICT projects in governments. There were meetings held with officials of the Ministry of Finance, Planning,

and Economic Development; the Office of the President; the Ministry of Local Government; and the Ministry of Works, Housing, and Communications with a view of understanding their current ICT initiatives in the country and how they are progressing.

This study was undertaken during 2006 and some of the data and the overall environment regarding the state of ICT in Uganda have undergone substantial changes. We have endeavored to explain where the ICT environment has changed since the last time of review earlier in 2006. Some changes are very positive and demonstrate how the Government of Uganda is committed to fully integrate ICT within all government processes. One such a development is the creation of the Ministry of ICT as a single face for ICT initiatives in Uganda. This is an opportunity for an environment that can be exploited to have the value of ICT enhance service delivery and policy formulation for more efficient citizen services and improved economic development in the country.

Before embarking on a broad strategy for implementing E-government, it is important to identify common practices and their trends in the existing ICT projects. Such trends will assist in identifying critical shortfalls that affect the successful implementation of E-government programmes. Further, this paper is intended to contribute towards the improvement of the processes and procedures employed in government bodies to implement ICT projects, collaborating in the effort to make an impact on service delivery to the citizens. The results of this research will help project planners and policy formulators define the areas into which they are supposed put more or less effort in a coordinated manner. This will allow them to implement a more productive work strategy whilst trekking on a steady path towards attaining the defined goals. This paper is organized as follows: Section 1 is the introduction. Section 2 presents the approach used in the study. Section 3 describes aspects related to E-government projects implementation aspects. Section 4 presents and analyses the results obtained in field research. Section 5 presents the main conclusions from this research and recommendations that can be adopted to improve project performance in government programmes.

2. Approach of the study

We reviewed literature from several documents related to E-government projects in Uganda. These include the Uganda E-Government Strategy (March 2004)[3], Uganda e-Readiness Assessment (March 2004),[4] The National ICT Policy for Uganda Implementation Framework Draft Final Report (February 2005) [5] and East African Community Regional E-Government Framework (Draft) December 2005.[6] We also reviewed literature on E-government from other countries to identify any best practices that are applicable in Uganda. From the documents reviewed, we focused on identifying evidence to confirm that the Government of Uganda is committed to a unified, integrated, and comprehensive ICT program to enable Government services to be delivered more efficiently and effectively to every segment of society. We also looked at sample projects that have already been undertaken in a sample of ministries with a focus on identifying challenges and

experiences from these flagship projects. Special attention was paid to e-government related projects based on the e-government strategic plan for the country.

3. E-government and ICT Projects

3.1. Overview

Driven by the belief that e-government is one of the key motors for development, governments are taking wide-ranging initiatives to rapidly create knowledge-based economic structures and information societies comprising networks of individuals, firms and countries interlinked electronically through webs of informational relationships. According to Datanet et al (1987) [7], the importance of expanding the access of developing countries to information and communication technologies (ICTs) has been recognized by governments and international agencies with increasing consensus that ICT-related technology should be regarded as a strategic national infrastructure. Development, in contemporary times, is characterized by various dimensions, including ICTs. A functional E-government structure is comprised of an ICT infrastructure, different computer applications, and knowledge workers who form the basis of new information societies. While the ICT infrastructure is a visible starting point, it is often very expensive to install but at the same time the easiest to see and verify. When it comes to government programmes, it is more complex to stimulate processes through which individuals, organizations, communities and countries create capacities to use information effectively in their local contexts and for their needs.

With the vision to offer better services to the citizens and business communities, the government of Uganda formulated the e-government strategy in 2004. This followed many years of government efforts to put e-government in practice through formulation of policies and structures to support its implementation. Literature available shows that Uganda received substantial support from donor agencies in the area of ICT for development. This has translated into a myriad ICT projects being implemented in various sectors of Ugandan society, most notably in rural infrastructure, education, livelihoods and health. The motivation for conducting this study has been accelerated by the need to establish the where Uganda as a country has reached on the path to full implementation of E-government, identify any challenges and where possible, focus on identifying strategies to handle the challenges.

There are indications that the Government of Uganda has recognized the critical importance of ICT in national development, and has started a policy framework to start implementing these technologies throughout the country. Several policies, statutes, and other initiatives have been undertaken toward this goal. The most recent of these include: (i) A National ICT Policy was approved in 2003 with the aim of promoting the development of ICT infrastructure in the country, with the Ministry of Works Housing, and Communications as the primary coordinated agency within the Government; (ii) A Draft broadcasting policy is in place; (iii) The Uganda Communications Commission is implementing a Rural Development

Policy; (iv) The new Communications Policy (Draft) seeks to connect all schools, sub-counties, urban centers, health centers and public libraries by 2010; (v) The Government is promoting Public-Private Partnerships to build the requisite backbone infrastructure.

Despite the above structure, individual Ministries have continued to adopt ICT initiatives based on internal factors and available opportunities for funding on an ad hoc, decentralized basis. As a result, ICT development within the Government remains more integrated at the national policy level, than it does with respect to translating that policy to a harmonized ICT implementation and operational guidelines across all Ministries.

3.2. ICT Programs in Selected Ministries

The following summaries highlight the major findings based on interview meetings and documentary reviews. The highlight are presented in Table 1, first identifying the ministry concerned, description of its responsibilities, current and planned ICT Programs and comments on ICT Program Implementation. In the commentary section, we highlight the causes of any failure so far experienced within the identified program.

Table 1: ICT Projects in selected Government Entities

Government Entity	Description of Responsibilities	Current and planned ICT Programs
Office of the President	The Office of the President is responsible for monitoring the execution and budgetary expenditures of government programs administered by the various Government Ministries. Additionally, the Office of the President is responsible for the dissemination of Public Affairs information to the citizenry and coordination of Public Affairs with Local Governments.	a. **Ministry Communications** The current planning priority is to establish an Internet/E-mail System (with services also provided to the State House and the Office of the Prime Minister). The objective is to enhance internal and external communications and data/information sharing and coordination at these key governmental offices.
		b. **Office of the President Intranet** Another focus is establishing an Intranet linking the Office of the President with the Statehouse. The objective with the Intranet is to provide a protected and secure internal data/information sharing between these physically separated elements of the staffs of the Office of the President.

Ministry of Works, Housing & Communication	The Ministry Works, Housing & Communication is responsible for all matters pertaining to Postal, Telecommunications and Infrastructure development	a. **Computerized Driving Permits (CDP)** This is an ICT project that is already in the implementation stage and is designed to provide the ministry with the following advantages or functionalities. (i) Production of machine readable driving permits; (ii) Form a CDP regional databank which can as well be used in other government sectors; (iii) Provide a vehicle tracking equipment; (iv) Improve the testing of drivers and (v) Design and enforce a driving curriculum
		b. **Vehicle Inspection** This was designed to address the following: (i) Reduction of mechanically unfit vehicles from being registered and licensed; (ii) Formation of regional data banks and (iii) Provision of vehicle tracking equipment. c. **Vehicle Registration and Licensing** This was designed to address the following: (i) Formation of regional data banks and (ii) Provision of vehicle tracking equipment.

Ministry of Education and Sports	The Ministry has the responsibility: "To provide for, support, guide, co-ordinate, regulate and promote quality education and sports to all persons in Uganda for national integration, individual and national development." The strategic vision for ICT policy in education is the mainstreaming of ICT n the Education Sector	a. **Ministry Network** At the Ministry, all the offices are connected on a network and are able to use the Internet and e-mail services as well as to access resources on the network. Ten districts have been connected on the wide area network as a pilot phase. These districts can send and receive information on that network.
		b. **Education Management Information System** (EMIS) Education Management Information System is a component of ICT in the Ministry of Education and Sports; it provides quality education statistics in a timely, cost-effective and sustainable manner. EMIS provides the education statistics and pupil details among others. After the procurement of computers, printers and accessories each district has hardware and software installed and is to carry out data gathering from schools for processing through ED*Assist application Software which is used in EMIS.
		c. **ICT Maintenance Facility** A project proposal has been completed with assistance from The International Institute for Communication and Development, The Hague, The Netherlands (IICD) for the establishment of a Support Call Centre to repair and maintain ICT equipment and to develop the capacity of users to perform preventive maintenance and basic troubleshooting of their equipment. The Ministry is now soliciting funding for the project proposal.

		d. **Workflow Management and Financial Information in Planning and Budgeting** The objective is to improve delivery of services in the Ministry by streamlining the funding cycle, providing for interfaces with external parties, and timely preparation and delivery of work plans enhancing transparency and accountability.
		e. **Connect-ED (Connectivity for Educator Development)** The Connect-ED project, initiated in May 2000, is supported by the United States Agency for International Development (USAID) in close cooperation with Uganda's Ministry of Education and Sports and within the framework of the U.S. Education for Development and Democracy Initiative (EDDI). Connect-ED is using technology to enable and enhance learning and teaching for primary educators through the creation of multifaceted approaches to integrating media and computers in the Primary Teacher Colleges (PTC) classrooms.
Ministry of Finance Planning & Economic Development	The Ministry of Finance Planning & Economic Development is the key ministry responsible for all aspects of financial data administration, planning concepts of the government as well as all economic development concepts in Uganda. This includes coordinating budget data for all the ministries within the Ugandan Government.	a. **Integrated Financial Management System (IFMS)** This project bundles all financial management functions into one suite of applications. The IFMS covers all the major Government business processes including Budgeting, Accounting and Reporting, Purchasing, Payments /Payables, Revenue management, Commitment Accounting, Cash Management, Debt Management, Fixed Assets, Fleet Management, and Inventory/ Stock Control. The Integrated Financial Management System assists the GOU entities to initiate, spend and monitor their budgets, initiate and process their payments, and manage report on their financial activities.

		b. **Information Sharing System (ISS)** This is another ICT project initiated by the Ministry of Finance Planning & Economic Development to provide a backbone infrastructure for the sharing of information within the MoFPED to ensure its fast and efficient flow together with minimizing the usage of physical paper work therein.
Ministry of Local Government	The Ministry of Local Government (MoLG) is charged with the responsibility of supporting and ensuring the efficient and effective operations of Local Governments (LGs) through proper management and coordination of the Decentralization process.	a. **Local Government Information Communication System ((LOGICS)** LoGICS was developed under the Local Government Development Program (LGDP I) and is comprised of three integrated parts, namely: (i) Monitoring and Evaluation Sub-system; (ii) Compliance Inspection Sub-system; (iii) Computerized Software Sub-system, which enables the data generated from the M&E sub-system and CI sub-system to be entered, verified, analysed, stored and disseminated to the various stakeholders. LoGICS is a multi-sectoral information system covering all sectors in a Local Government including: Education, Health, Water, Roads, Prisons, Police, Production, Planning, Finance and Administration, Council, Social Services etc.

		b. **Local Government Financial Information Analysis System (LGFIAS)** This system captures all relevant financial data on revenues and expenditure for all levels of Local Governments. The system has been designed with facilities to analyse and generate in-depth reports on revenue performance, expenditure, donor funds and Central Government transfers to the Local Governments. The reports generated are used by the Local Authorities, Central Government, Development Partners, NGOs and other stakeholders for decentralized fiscal planning, policy formulation and decision making functions.
		c. **Performance Monitoring Management Information System (PMMIS)** The Ministry hired a Consultant to develop a Performance Monitoring Management Information System and a Client Feedback System. This reporting framework, when complete, will provide end-users with means of accessing data from both systems, as if it were stored in a single system. It will also provide users with means of drilling down/up and to dynamically generate reports of interest. This reporting framework is currently undergoing testing at the ministry.

4. Key findings

In general, the development and integration of ICT within the Government is uneven, with the lack of adequate resources to dedicate to ICT programs. Therefore, programs that enlist international donor organizations have been the primary catalyst for ICT penetration into the Government sector. Some ministries i.e. Finance, have substantial electronic records processing, databases and information retrieval systems, internal LANs and external networking to other Ministries. The World Bank provided much of the funding for these initiatives. Many other Ministries are still working to establish internal networking, information exchange, and preliminary planning for database management systems. Efforts are very limited to date in reviewing business processes and realignment of staffs to promote efficient application of electronic Government processes and applications. Specifically, the survey reveals several key factors that must be considered as part of an effective ICT strategy development:

(i) ICT programs must build on the unanimous agreement throughout the Government, and the political will, demonstrated through the expressed commitment of the President, to enable the Government of Uganda through comprehensive implementation of ICT technologies.

(ii) The lessons learned from initiatives sponsored by International Donor Organizations should serve as templates in building a comprehensive implementation strategy

(iii) ICT spending on equipment, as well as training, is not coordinated across the government.

(iv) ICT investment remains an "ad hoc" affair, with each individual Ministry seeking ICT funding primarily from defined donor project resources to offset the minimal funding available through the governmental budgetary channels.

(v) A majority of the installed systems are underutilized owing to a series of factors, including a lack of operator training, lack of connectivity, and absence of automated processes.

(vi) No series of commonly accepted standards is in place, or even informally agreed upon, for equipment, applications, or connectivity, for the current initiatives.

(vii) A lack of coordinated ICT investment strategy and tracking of ICT investments and performance monitoring within the Government.

In summary, a comprehensive ICT implementation strategy, cross-cutting all Ministries, requires a centralized and coordinated organizational structure to ensure the most rational and cost-effective utilization of scarce resources and fully standardized and inter-operable systems throughout the Government. With the creation of a Ministry of ICT, there is anticipation that this challenge will be addressed.

Using the dimension model that has been used in other work by Vitalari N.

(1988) [8], Table 2 below shows components that can be used to realistically address challenges bound to face ICT projects in government before projects fail. The components represent dimensions and the issues of concern are described beside each dimension.

Table 2: Dimension Model Components (according to Vitalari N (1988)

Dimension/ Component	Description
(i) Information	All planning and framework documents reviewed show that there is an assumption that creation of formal strategic information would be of value to government functioning. This is a system design issue. In reality, informal information and gut feelings are what decision makers' value and use.
(ii) Technology	The design often assumes the use of a broad range of software and hardware, with two PCs per department in all the offices of the Service. The initial reality was manual operations: paper supported by typewriters, phone, fax and post. This created a large design-reality gap on this dimension.
(iii) Process	In all project documents reviewed, the system design assumes that a rational model of structured decision-making hold sway within the government. This mismatches the dominant reality of personalized, even politicized, unstructured decision-making. Designs often assume that automation of pre-existing processes with some amendments made to the way in which data was gathered, processed, stored and output – with many previously human processes (including checking and retyping of figures) being altered to computerized processes. Almost all of the key public decisions remain as they were prior to computerization (only assumed to be made more efficiently and effectively). This creates a medium design-reality gap on this dimension.
(iv) Objectives and values	The findings indicate that the systems are usually designed within, and reflecting, a scientific environment which has a 'role culture' that values rules and logic. In reality, systems are to be used in a political environment which has a 'power culture' that values self-interest and hidden agendas. Some of the projects do not progress because funds are diverted or equipments are procured wrongly. The design often assumes that the objectives of the project (automation of processes, better decision-making) are shared by all stakeholders. In reality, prior to computerization, most senior officers supported these objectives, since it was they who had initiated the project. However, most staff within the departments initially opposed the system; they fear changes in their working patterns and they fear job losses.

(v) Management systems and structures	From what was observed, systems are designed for institutions that have both structures and systems to support strategic decision making. In reality, such structures and systems do not exist within the government establishments. With the creation of the Ministry of ICT and according to its mandate, there is a promise that the management systems and structures will be modified to ensure successful project implementation.
(vi) Staffing and skills	The designs make two significant assumptions. First, it assumed the ongoing presence during and after implementation of a cadre of staff with strong IT and information systems skills. The initial reality is that no such staffs exist in most government departments. Second, it assumed a reduction by 50% in the numbers of staff within the departments since human intervention in many data-handling processes would no longer be required. Clearly, this is significantly different from the initial reality before computerization. The design also assumes that the addition of some new skills, but no other changes in large numbers of jobs within the Service.

5. Recommendations and conclusion

5.1 Recommendations

From the identified challenges in each dimension in Table 2, project planners within Minsitries need to determine the course of action to address the challenges identified. Using the methodology recommended by Norris, D. F. & Moon, M. J. (2005) [9], the real gap on each dimension can be addressed. For illustration purposes, Table 3 presents the methodology of addressing gaps identified in each dimension as recommended by Garson, G.D (2004) [10]**.** This methodology proposes that if a project has a significant overall design-reality gap, the planners should take action since the project may be heading for failure. Similarly, if there is a significant design-reality gap on a particular dimension, the planners should take action since this may cause problems on the overall project. Assessing whether a gap is 'significant' is a matter for discussion, debate and judgment. Taking action means either changing the design of the e-government project to make it more like reality, and/or changing current reality to make it more like the assumptions/requirements within project design. Selected techniques will, of course, depend on which dimension(s) the gap occurs. In selecting a technique for a particular project, planners need to ensure that the technique is not only desirable but also feasible. There is no point considering techniques that could reduce risks in theory, but not be implemented in practice. Table 3 as mentioned earlier illustrates how this methodology can work in government projects.

Table 3: Dimension-Specific Gap Reduction Techniques

Dimension	Applicable Techniques
Information dimension.	(i) Undertake a professional requirements analysis in order to draw out true information needs of stakeholders. (ii) Use prototyping – getting users to use a test version of the e-government application – in order to help them explain what information they really need.
Technology dimension.	(i) Investigate ways in which government reforms could be delivered without ICTs. (ii) Investigate ways in which government reforms could be delivered using the existing ICT infrastructure. (iii) Avoid leading-edge technologies in your design. (iv) Investigate opportunities for use of donated or recycled equipment.
Process dimension.	(i) Keep doing things the same way, only with the addition of some new technology (see generic point above about automation). (ii) Avoid business process reengineering; instead, at most, look at optimization or minor modification of existing processes within the E-government application design. (iii) Consider a two-stage approach: in the first stage, processes are optimised without any change to ICTs; in the second and later stage, new ICTs are brought in.
Objectives and Values dimension.	(i) Use rewards to alter stakeholder objectives and values (e.g. messages of management support, better pay, better working conditions, career advancement, etc.). (ii) Use punishments to alter stakeholder objectives and values (e.g. threats, reprimands, transfers, worsened pay and conditions, etc.). (iii) Communicate with stakeholders about the system: sell the true benefits and address the true negative aspects. (iv) Get key stakeholders (those regarded as key opinion formers or those vociferous in their resistance to the e-government application) to participate in the analysis and/or design of the e-government application. (v) Base e-government application design on a consensus view of all main stakeholders. (vi) Use prototyping: this helps incorporate stakeholder objectives in the design, and also helps to make actual stakeholder objectives more realistic. (vii) If feasible in skill, time and motivational terms, get users to help develop and build the e-government application.

Staffing and Skills dimension.	(i) Outsource contracts in order to improve the current reality of available competencies (though this may increase other gaps). (ii) Train staff to improve current reality of competencies. (iii) Improve recruitment and retention techniques to reduce competency (staff) turnover. (iv) Make use of external consultants (though this may increase other gaps). (v) Hire new staff to expand the volume of current competencies.
Management Systems and Structures dimension.	(i) Make an explicit commitment to retain the existing management systems and structures within e-government application design.
Other Resources dimension	(i) Prioritise e-government applications that maximise revenue generation for Government (e.g. those dealing with tax, fees, fines, etc). (ii) Seek additional financing from donor or central government agencies. (iii) Get private firms to develop, own and operate the e-government application. (iv) Charge business or wealthier users of the e-government system. (v) Scale-down ambitions of the e-government project. (vi) Negotiate central/shared agency IT agreements to reduce hardware and software costs.

5.1 Conclusion

Our review of current programs revealed, in general, that the development and integration of ICT within the Government is uneven, with the lack of adequate resources to dedicate to ICT programs. Therefore, programs that enlist international donor organizations have been the primary catalyst for ICT penetration into the Government sector. Recent Government initiatives have recognized the need to provide more centralized planning and implementation of ICT initiatives. The Uganda E-Government Strategic Framework (Final Draft) January 2006 proposes a fully fledged Ministry of ICT to be established by Government to provide both political and technical leadership in the overall coordination and harmonization of policy development and implementation for Information Technology (IT), Information Management (IM) services, Communications (Tel & Postal) services and Information and Broadcasting services. With the recent authorization and activation of the ICT Ministry provides unified coordination and harmonization of these initiatives within one political and technical leadership will probably avoid the current duplication existing in both the Central and Local Government

projects. The structure now in place, given the necessary authority and mandate should be fully capable of coordinating all of the appropriate governmental entities under a unified organizational structure that can effectively address cross cutting issues.

References

Bretschneider, S. (1990). Management Information Systems in Public and Private Organizations: An Empirical Test. *Public Administration Review*, 50(5), 536-545.

Danziger, J. N. &. Kraemer, K. L. (2006). *People and Computers.* New York: Columbia University Press.

The Uganda E-Government Strategy by Techno - Brain (March 2004),

Uganda e-Readiness Assessment (March 2004), The National ICT Policy for Uganda Implementation Framework Draft Final Report (February 2005)

East African Community Regional E-Government Framework (Draft) December 2005 Datanet K. N; Blumler, J. G., & Kraemer, K. L. (1987). *Wired Cities: Shaping the Future of Communications.* Boston, MA: G.K. Hall.

Vitalari, N. (1988). The Impact of Information Technology on Organizational Design and the Emergence of the Distributed Organization. Irvine, CA, Public Policy Research Organization, University of California, Irvine.

Norris, D. F. & Moon, M. J. (2005). Advancing E-Government at the Grass Roots: Tortoise or Hare? *Public Administration Review*, 65(1), 64-75.

Garson, G.D. (2004). The Promise of Digital Government. In Garson (Eds.), *Digital Government Principles and Best Practices* (pp. 2-15). Hershey, PA: Idea Group Publishing.