Proceedings of the
IADIS International Conferences
Informatics 2011
Wireless Applications and Computing 2011
Telecommunications, Networks and Systems 2011

EDITED BY
Hans Weghorn,
Jörg Roth and
Pedro Isaías
PROCEEDINGS OF THE IADIS INTERNATIONAL CONFERENCES
INFORMATICS 2011
WIRELESS APPLICATIONS AND COMPUTING 2011
TELECOMMUNICATIONS, NETWORKS AND SYSTEMS 2011

part of the
IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND
INFORMATION SYSTEMS 2011

Rome, Italy
JULY 20 - 22, 2011

Organised by
IADIS
International Association for Development of the Information Society
# TABLE OF CONTENTS

**FOREWORD**  
ix  
**PROGRAM COMMITTEE**  
xi  
**KEYNOTE LECTURE**  
xxix  

## FULL PAPERS

**CHALLENGES IN ADAPTIVE SYSTEMS DESIGN:**  
THE CASE OF E-ENABLED PERFORMANCE MANAGEMENT  
*Polyxeni Vassilakopoulou and Theodore Darmaros*  

**ALGORITHM TO OPTIMIZE THE POWER CONSUMPTION OF SCALABLE CLUSTERS GUARANTEEING RESPONSE TIMES**  
*Daniel F. García, Ramón Medrano and Joaquín Entrialgo*  

**THE IMPACT OF ENVIRONMENTAL DYNAMISM ON THE EFFECTIVENESS OF SISP: A COMPARATIVE STUDY OF MANUFACTURING AND NON-MANUFACTURING MEDIUM ENTERPRISES**  
*Luka M. Musangu and Ray M. Kekwaletswe*  

**AN O(N) APPROXIMATION FOR THE DOUBLE BOUNDING BOX PROBLEM**  
*Jörg Roth*  

**HYBRID FUZZY CLASSIFIER DESIGN WITH COEVOLUTIONARY GENETIC ALGORITHM**  
*Roman B. Sergienko, Eugene S. Semenkin and Vladimir V. Bukhtoyarov*  

**A QOS-AWARE APPROACH FOR WEB SERVICE SELECTION BASED ON PROBABILITY EVALUATION**  
*Dessislava Petrova-Antonova*  

**AGILE DATA WAREHOUSING: THE SUITABILITY OF SCRUM AS DEVELOPMENT METHODOLOGY**  
*Roelien Goede*  

**INTEROPERABLE PROCESS MONITORING USING CLINICAL IT-STANDARDS AND HEALTHCARE INFORMATION SYSTEMS**  
*Eva Gattinar, Okan Ekinci and Vesselin Detschew*  

**ALTERNATIVE QUERY DISCOVERY FROM THE WEB FOR DAILY MOBILE DECISION SUPPORT**  
*Shun Hattori*
AN EMPOWERED ENERGY EFFICIENT APPROACH FOR WSN BASED TRACKING APPLICATIONS
Tareq Alhmiedat and Shuang-Hua Yang

A COST-AWARE MULTI-PATH ROUTING PROTOCOL FOR MULTI-INTERFACE MULTI-CHANNEL MANETS
Sudan Li, Dan He and Zhili Sun

A COMPLETE AND EFFICIENT STRATEGY FOR COOPERATIVE CACHE IN MOBILE INFORMATION SYSTEM
Thu Tran Minh Nguyen and Thuy Thi Bich Dong

QOS PROVISIONING IN CONTENTION AWARE MANETS USING FLOW-AWARE ADMISSION CONTROL PROTOCOL
Muhammad Asif, Zhili Sun, Haitham Cruickshank and Naveed Ahmad

THE SHAPE OF THINGS TO COME: SCENARIOS AND VISUAL STORIES FOR TELECOMMUNICATION IN 2020
Henning Breuer, Hartwig Grabowski and Heinrich Arnold

SHORT PAPERS

REVIEW OF QUERY PROCESSING IN DISTRIBUTED SYSTEMS
Dinesh K. Saini, Nebras N. Hasoon, Feras N. Hasoon and Mustafa Hasan

GRAPH-BASED AUTHENTICATION ALGORITHMS FOR VEHICULAR AD-HOC NETWORKS
P. Caballero-Gil, C. Caballero-Gil and J. Molina-Gil

ONLINE VERSUS OFFLINE CRITICAL GEOGRAPHIC INFORMATION SYSTEM
Jan Hodicky and Petr Frantis

GRANULAR COMPUTING ALGORITHM TO DECREASE THE UNCERTAINTY OF KNOWLEDGE EXTRACTION
H. S. Alinia, M. R. Delavar, M. Zare and A. Mohseni

MOBILE DETECTION OF SOCIAL SITUATIONS WITH TURN-TAKING PATTERNS
Georg Groh, Alexander Lehmann and Manuel de Souza

A SERVICE DISCOVERY PROTOCOL FOR WIRELESS MESH NETWORKS
Omonvo Oluwagba and Chris Guy

ENHANCED SIMPLE FAIR EXECUTION TIME ESTIMATION (ESFETE) SCHEDULING ALGORITHM FOR GRID SCHEDULER
Rohaya Latip and Adam Malik Sahar

A PROPOSAL OF ACCESS-BASED CONTENTS GROUPING ON P2P NETWORK
Takuya Sasaki, Jun Sawamoto and Yuji Wada
REFLECTION PAPER

THE ROLE OF INFORMATION TECHNOLOGY ON SUCCESSFUL KNOWLEDGE SHARING IN KNOWLEDGE MANAGEMENT PROJECTS: CASE STUDY
Lila Rajabion, Azadeh Kheirabadi, Alireza Khanjari and Kamran Rezaei

POSTERS

SOME KNOWLEDGE NORMALIZATION METHODS
Philippe Martin

A DOMAIN SPECIFIC LANGUAGE TO REPRESENT STRUCTURED DATA FOR MULTIPLE PROGRAMMING LANGUAGES
Kazuaki Maeda

AUTHOR INDEX
FOREWORD

These proceedings contain the papers of the IADIS International Conferences Informatics 2011, Wireless Applications and Computing 2011 and Telecommunications, Networks and Systems 2011, which was organised by the International Association for Development of the Information Society in Rome, Italy, 20 – 22 July, 2011. These conferences are part of the Multi Conference on Computer Science and Information Systems, 20 - 26 July 2011, which had a total of 1402 submissions.

The IADIS Informatics 2011 conference shall host fundamental topics on Informatics. In addition, its scope is not just limited to fundamental theory, but it should also cover the impact of Informatics on society and human life, and it shall furthermore complement theory foundations with technical considerations and practice.

The IADIS Wireless Applications and Computing 2011 addresses several themes related to theory and practice within wireless networks, computing and application related areas.

Enormous developments in wireless technologies made in recent years enabled both researchers and industry to create new, innovative wireless and mobile application and services, protocols, middleware platforms and application frameworks. Corresponding research is related to all communication layers, but also includes application-related topics, theoretical results and non-technical issues. Main topics have been identified. However, innovative contributions that don’t fit into these areas will also be considered since they might be of benefit to conference attendees.

The IADIS Telecommunications, Networks and Systems 2011 covers theory, design and application of computer and telecommunication networks and systems. During recent years there has been an impressive increase in the use of networked applications and networks are now key resources in any information system configuration. Wireless and fixed-line networks complemented by a growing range of mobile devices are having a significant impact on the way we run our lives and our businesses.

These events received 99 submissions from more than 26 countries. Each submission has been anonymously reviewed by an average of five independent reviewers, to ensure that accepted submissions were of a high standard. Consequently only 14 full papers were published. The overall acceptance rate corresponds to about 14 %. A few more papers were accepted as short papers, reflection papers and posters. An extended version of the best papers will be published in the IADIS International Journal on Computer Science and Information Systems (ISSN: 1646-3692) and/or in the IADIS International Journal on WWW/Internet (ISSN: 1645-7641) and also in other selected journals, including journals from Inderscience.
Besides the presentation of full papers, short papers, reflection papers and posters, the conferences also included a keynote presentation from an internationally distinguished researcher. We would therefore like to express our gratitude to Prof. Saman K. Halgamuge, University of Melbourne, Australia, for accepting our invitation as keynote speaker.

As we all know, organising these conferences requires the effort of many individuals. We would like to thank all members of the Program Committees, for their hard work in reviewing and selecting the papers that appear in the proceedings.

This volume has taken shape as a result of the contributions from a number of individuals. We are grateful to all authors who have submitted their papers to enrich the conferences’ proceedings. We wish to thank all members of the organizing committee, delegates, invitees and guests whose contribution and involvement are crucial for the success of the conferences.

Last but not the least, we hope that everybody will have a good time in Rome, and we invite all participants for the newly created Theory and Practice in Modern Computing 2012 conference that will be held in Lisbon, Portugal.

Hans Weghorn
BW Cooperative State University Stuttgart
Germany,
Informatics 2011 Conference Program Chair

Jörg Roth
University of Applied Sciences Nuremberg
Germany
Wireless Applications and Computing 2011 Program Chair

Pedro Isaías
Universidade Aberta (Portuguese Open University)
Portugal
Telecommunications, Networks and Systems 2011 Program Chair

Piet Kommers, University of Twente, The Netherlands
Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal
MCCSIS 2011 General Conference Co-Chairs

Rome, Italy
July 2011
PROGRAM COMMITTEE

INFORMATICS CONFERENCE PROGRAM CHAIR
Hans Weghorn, University of Cooperative Education, Germany

WIRELESS APPLICATIONS AND COMPUTING CONFERENCE PROGRAM CHAIR
Jörg Roth, University of Applied Sciences Nuremberg, Germany

TELECOMMUNICATIONS, NETWORKS AND SYSTEMS CONFERENCE PROGRAM CHAIR
Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal

MCCSIS GENERAL CONFERENCE CO-CHAIRS
Piet Kommers, University of Twente, The Netherlands
Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal

INFORMATICS COMMITTEE MEMBERS
Abdel-badeeh M. Salem, Ain Shams University, Egypt
Ake Sivertun, Swedish National Defence College, Sweden
Amnart Pohthong, Prince Of Songkla University, Thailand
Ana Paula Appel, Federal University Of Espirito Santo, Brazil
Anca-Juliana Stoica, IT University Uppsala, Sweden
Andrea Perego, Universita Degli Studi Dell insubria, Italy
Andrzej Goscinski, Deakin University, Australia
Antonio Cau, De Montfort University, United Kingdom
Antonio Martins, Universidade Estadual Paulista Júlio De Mesquita F, Brazil
Arnab Dhua, A9.com Inc., USA
Ateet Bhalla, NRI Institute of Information Science and Technolog, India
Chandana Unnithan, Deakin University, Australia
Chantana Phongpensri, Silpakorn University, Thailand
Chara Mitropoulou, National Technical University Of Athens, Greece
Chih-chieh Hung, National Chiao Tung University, Taiwan
Chris Scogings, Massey University, New Zealand
Colette Johnen, Université Bordeaux, France
Cyril Onwubiko, Kingston University, United Kingdom
Dave Parry, Auckland University Of Technology, New Zealand
Debasis Giri, Haldia Institute Of Technology, India
Deepak Narasimha, University Of Malaya, Malaysia
Des Watson, University Of Sussex, United Kingdom
Dumitru Burdescu, University Of Craiova, Romania
Elad Michael Schiller, Chalmers University Of Technology, Sweden
Eleni Vlahogianni, National Technical University Of Athens, Greece
Enzo Seraphim, Federal University Of Itajuba, Brazil
Eric Pardede, La Trobe University, Australia
Erich Marth, University Of Vienna, Austria
Ezendu Ariwa, London Metropolitan University, United Kingdom
Fausto Fasano, Università Degli Studi Del Molise, Italy
Francesca Scozzari, Dipartimento Di Scienze, Italy
Georg Duftschiindl, Medical University Of Vienna, Austria
George Gravanis, Democritus University Of Thrace, Greece
Graça Bressan, Escola Politécnica Da Universidade De São Paulo, Brazil
Helge Janicke, De Montfort University, United Kingdom
Huei Lee, Universidade Estadual Do Oeste Do Paraná, Brazil
Humberto Razente, University Of São Paulo, Brazil
Imad Abugessaisa, Karolinska Institutet, Sweden
Indrajit Ray, Colorado State University, USA
Itana Gimenes, Universidade Estadual De Maringá, Brazil
Jean-louis Boulanger, Laboratoire Heudiasyc, France
Jerzy Dabkowski, Cracow University Of Technology, Poland
Jing Yuan, Microsoft Research Asia, China
José Guimarães, Universidade Federal De São Carlos (ufscar), Brazil
Justin Rough, Deakin University, Australia
Kalpdrum Passi, Laurentian University, Canada
Karl Froschln, University Of Vienna, Austria
Katti Faceli, University Of São Paulo, Brazil
Ken Hawick, Massey University - Albany, New Zealand
Kevin Daimi, University Of Detroit Mercy, USA
Kok-leong Ong, Deakin University, Australia
Krassie Petrova, Auckland University Of Technology, New Zealand
Laura Carnevali, DSI, Italy
Leonor Barroca, Open University, United Kingdom
Lu-an Tang, University Of Illinois At Urbana-champaign, USA
Lucia Rapanotti, The Open University, United Kingdom
Luciana Zaina, Ufscar, Brazil
Luis Scott, Federal University Of Abc, Brazil
Marc Conrad, University Of Bedfordshire, United Kingdom
Marcelo Do Nascimento, Federal University Of ABC, Brazil
Marcio Oikawa, Federal University of Sao Carlos, Brazil
Marek Stanuszek, Cracow University Of Technology, Poland
Maria Camila Barioni, Ufabc, Brazil
Martin Sandrieser, University Of Vienna, Austria
Michael Le Duc, Mälardalen University, Sweden
Mustafa Atay, Winston-Salem State University, USA
Nikos Lagaros, National Technical University Of Athens, Greece
Nithya Vijayakumar, Cisco Systems, Inc., USA
Pablo Bringas, University Of Deusto, Spain
Paddy Krishnan, Bond University, Australia
Pierre Leone, Université De Genève, Switzerland
Piotr Andrzej Kowalski, Cracow University Of Technology, Poland
Reagan Moore, Renaissance Computing Institute, USA
Rebeca Cortazar, Universidad de Deusto, Spain
Richard Colbaugh, Sandia National Laboratories, USA
Robert Dew, Deakin University, Australia
Ronald Prati, Universidade Federal do ABC, Brazil
Ryosuke Saga, Kanagawa Institute Of Technology, Japan
Sarka Kvetonova, Brno University Of Technology, Czech Republic
Saulius Gudas, Kaunas University Of Technology, Lithuania
Shaojie Qiao, Southwest Jiaotong University, China
Shoba Tegginmath, Auckland University Of Technology, New Zealand
Siegfried Benkner, University Of Vienna, Austria
Sung Ho Ha, Kyungpook National University, Korea, Republic Of
Tee Sim Hui, Multimedia University, Malaysia
Thilo Bauer, bluewater multimedia concepts, Germany
Tomas Skersys, Kaunas University Of Technology, Lithuania
Vladimir Oleshchuk, University of Agder, Norway
Wichian Sittiprapaporn, Mahasarakham University, Thailand
Witold Pedrycz, University Of Alberta, Canada
Yeda Regina Venturini, Universidade Federal De Sao Carlos, Brazil
Yoseba Penya, University Of Deusto, Spain
Zbigniew Kokosinski, Cracow University Of Technology, Poland
Zhong-hui Duan, University of Akron, USA

WIRELESS APPLICATIONS AND COMPUTING COMMITTEE MEMBERS

Alexander Klein, Technische Universität München, Germany
Alexander Lazovik, Inria, France
Andre Aquino, IC-UFAL, Brazil
Antonio De La Oliva, Universidad Carlos III De Madrid, Spain
TELECOMMUNICATIONS, NETWORKS AND SYSTEMS COMMITTEE
MEMBERS

Abderrahmane Lakas, United Arab Emirates University, United Arab Emirates
Abdul Hamid, Hankuk University Of Foreign Studies, South Korea
Afrodite Sevasti, GRNET, Greece
Alan Ruddle, University Of St Andrews, United Kingdom
Albert Levi, Sabanci University, Turkey
Alex Galis, University College London, United Kingdom
Alfonso Rodriguez Rios, Universidad Del Bio-bio, Chile
Al-sakib Khan Pathan, International Islamic University Malaysia, Malaysia
Andres Muñoz, Universidad De Murcia, Spain
Andreu Sancho Asensio, Universitat La Salle, Spain
Andrzej Goscinski, Deakin University, Australia
Angel Moreno, Universidad De Alcala, Spain
Antonio De Domenico, Minatec, France
Antonio Izquierdo Manzanares, Nist, USA
Antonio Ruiz Martinez, University Of Murcia, Spain
Arup Acharya, Ibm Tj Watson Research Center, USA
Atilla Elci, Toros University, Turkey
Bob Askwith, Liverpool John Moores University, United Kingdom
Brigitte Kervella, Upmc - Paris Universitas, France
Bulent Tavli, Tobb Economy And Technology University, Turkey
Byoung Uk Kim, Ridgetop Group, Inc., USA
Carlos Becker Westphall, Federal University Of Santa Catarina, Brazil
Carlos Juiz, Universidad De Las Islas Baleares, Spain
Carlos Serrao, ISCTE/DCTI, Portugal
Chris Guy, University Of Reading, United Kingdom
Claudio Palazzi, University of Padova, Italy
Cristina Alcaraz, University Of Malaga, Spain
Dai Davies, Dante Ltd, United Kingdom
Dan He, University Of Surrey, United Kingdom
Daniel Caragata, University Of Nantes, France
Dario Bottazzi, Laboratori Guglielmo Marconi, Italy
Djamel Belaïd, Institut Telecom - Telcom SudParis, France
Dr. Md. Abdur Razzaque, University Of Dhaka, Bangladesh
Emilio Serrano, Universidad De Murcia, Spain
Enrique De La Hoz, University Of Alcalá De Henares, Spain
Erica Cecilia Ruiz Ibarra, Instituto Tecnológico De Sonora, Mexico
Farouk Kamoun, University Of Manouba Ensri, Tunisia
Faycal Bouhafs, Liverpool John Moores University, United Kingdom
Federico Montesino Pouzols, Helsinki University of Technology, Finland
Felix Strohmeier, Salzburg Research, Austria
Felix J. Garcia, University Of Murcia, Spain
Fernando Pereñiguez Garcia, University Of Murcia, Spain
ABSTRACT

Propagation of a natural disaster across a landscape is stochastic in nature. However, the pattern may have an underlying structure in a given geographical primitive (alongside the river banks in the case of a flood, driven by the wind patterns in the case of a forest fire, etc). We discuss about models that combines prior knowledge with in-coming stochastic data to predict expected times of failure at various locations to optimally prepare sensor nodes to capture vital environmental events with limited on-board power, a hopping mechanism to enable sensor nodes to be mobile to achieve maximum resolution of sensing, and a new method of orphan node management enabling the sensor network to be robust in the presence of random malfunctioning of sensor nodes. We will also present some of the recent sensor network deployments.
Full Papers
CHALLENGES IN ADAPTIVE SYSTEMS DESIGN: 
THE CASE OF E-ENABLED PERFORMANCE MANAGEMENT

Polyxeni Vassilakopoulou\textsuperscript{1} and Theodore Darmaros\textsuperscript{2}
\textsuperscript{1}National Technical University of Athens, School of Mechanical Engineering, 
\textsuperscript{2}PlanetSA

ABSTRACT
Managing enterprise performance is intrinsically complex due to the need to cater for diverse objectives in organizational settings that operate within evolving environments. This paper aims to investigate design approaches for performance management tools focusing on the notion of “design for adaptability”. The empirical basis is a longitudinal case study that follows the evolution of a performance management system in a mid-sized bank. The main contributions of the paper are (i) drawing parallels among different design and deployment approaches that foster system adaptability (ii) proposing a combination of approaches to address challenges in adaptable systems design and deployment.

KEYWORDS
Information Systems, Performance Management, Design Principles, Adaptability, Sociotechnical Design

1. INTRODUCTION

In this paper we observe the adaptability challenges related to the design and deployment of information systems for performance management using a case study in a mid sized bank. Interpreting the case study as an instance of “sociotechnical design” that facilitates system adaptability by establishing change mechanisms, we propose that this sociotechnical approach can precede customization/parametrisation activities required for the introduction of the relevant components of modern Enterprise Resource Management Systems (ERPs). ERPs cater for adaptability via advanced functionality and control mechanisms. We thus see complementarities between two different approaches that follow different design paradigms. Concluding the paper we propose that systems implementation can combine different approaches as organizations reach different levels of maturity in their understanding of needs.

2. A SHIFT TO ADAPTABILITY

Recent literature proposes complexity as a lens through which to conceptualize information systems and theorise on their design (Kovács, and Ueno, 2004; Merali, 2006). This, signifies a shift towards adaptability for the information system discipline: “We argue first that managers should view Information System Development (ISD) projects as complex adaptive systems so as to more effectively cope with the challenges of evolutionary complexity in changing environments. We further suggest an adaptation perspective of ISD that rests primarily on co-evolutionary theory, which we believe will be useful for managing the emergent nature of most information systems” (Benbya and McKelvey, 2006). Going beyond automation or support of well defined simple tasks most systems designed today are aimed to support complex and dynamic activities. A significant proportion of the systems being built have to cater for fluid relationships and goals that evolve as the environmental conditions change. For such cases, traditional design focusing on the attributes of components with fixed interactions has limited use, which explains the recent trend towards adaptability. The key challenge for this shift towards adaptive systems is how to facilitate adaptation for technical components, user activities and organizational structures while retaining control: “research on the interplay between
dimensions of change and control will be vital in developing an understanding of future IT uses in organizational contexts” (Tilson et al, 2010).

In the sections that follow we discuss a case study on a system introduced to support Performance Management within a large-scale organization (a bank). Performance management requires data assemblage, analysis, interpretation, and communication (Bose, 2006). It is an exemplary complex and adaptive work activity as measurement, feedback, and orientation for improved performance are continuously adjusted responding to changes of environmental conditions. This adjustment can be achieved when systems are stable enough to be used as a basis for further development and flexible enough to allow quick alterations. There is a wealth of literature spanning more than two decades on performance management systems (Johnson and Kaplan, 1987; Dixon et al., 1990; Eccles and Pyburn, 1992; Kaplan and Norton, 1992, 1996; Neely and Adams, 2001; Marr and Schiuma, 2003; Busci and Bititci, 2006; Ferreiraa and Otley, 2009). Most of these well-referenced writings focus on systems’ characteristics and not on implementation approaches to be followed in order to achieve the sought-after adaptability. In this paper we present two different approaches for adaptability: namely, the customization/parametrisation approach put forward by major ERP system providers and the “sociotechnical” approach from information systems literature. Both approaches aim to deliver adaptable and efficient systems catering to evolving needs but are stemming from different traditions. ERPs have been developed within the engineering domain and naturally have a strong focus to control, while, the sociotechnical perspectives for information systems came out of the social science domain that sought to balance the rigidity of techno-driven approaches with contingent, pragmatic provisions for change. We postulate that the two approaches are not contradictory but complementary resolving the tension between change and control in different ways. Depending on organization’s preparedness and functional requirements’ maturity one of the two approaches or a combination of both can be suitable for application.

2.1 Modern Integrated Solution Implementation

Modern integrated solutions (exemplified by systems for Enterprise Resource Planning – ERPs) provide ways to deal with the complex co-evolution of processes, structures and technologies catering for the relationships among technical, functional, and institutional components of organizations. The key idea is to offer a robust basis where the nonsingular characteristics of each organization are handled with ready-made solutions while the more idiosyncratic parts can be handled via customisation or parametrisation activities. Such systems are usually designed as sets of modules built around functions transcending organizational structures that are interlinked exchanging data and sharing common repositories. The different components can evolve at different paces as long as the interrelationships are catered for. The adaptability of the systems relies to customisation-parametrisation activities that can range from simple end-user “preferences” setting to highly technical programming using vendor-specific languages. Big ERP vendors are enhancing their products widening their scope and building further on the “adaptability” ideal. The following excerpts are taken from the promotional brochures of two key vendors (SAP and Oracle):

• “To address changes in the business environment, you need to investigate disparate disciplines into a single solution that can handle the demand for unified information, collaborative decisions, and business network optimization” (SAP, 2009); “SAP provides the comprehensive, forward-looking support you need for strategic alignment, predictable performance, and confident decisions” (SAP, 2010); stressing the need for adaptability SAP foresees “a seismic shift from today’s technology architecture, involving completely new paradigms for underlying software and hardware infrastructure, in the way software is developed and deployed, and in usability, adaptability, flexibility, and integration” (SAP, 2006).

• “Management excellence requires an integrated set of performance metrics and management processes so you can link strategic scenarios to your financial and operational management...Oracle’s EPM system includes a suite of performance management applications ... easy to use, easy to manage, and the lowest in cost to deploy” (Oracle, 2010); Oracle’s offering enhances the “ability to respond to demand signals in real time, configure operations in response to those signals, and produce product in a timely fashion to meet customer expectations. Flexibility in and adaptability of operations requires a higher degree of information availability and collaboration than ever before” (Oracle, 2004).

Both companies propose solutions for performance management based on two key points: orchestrating different functions via a common robust and mature platform, ensuring adaptability to future needs.
2.2 Sociotechnical Systems Design for Adaptability

In the sociotechnical approach the interrelatedness of social and technical aspects is acknowledged and a shared emphasis is given to the organizational and technical elements. Cherns (1976, 1987) developed a set of sociotechnical design principles aiming to the “joint optimization of the technical and the social aspects, thus exploiting the adaptability and innovativeness of people in attaining goals instead of over-determining technically the manner in which these goals should be attained”. The ten principles are briefly recited here: “Compatibility”, “Minimal Critical Specification”, “Variance Control”, “Boundary Location”, “Information Flow”, “Power and Authority”, “The Multifunctional Principle”, “Support Congruence”, “Transitional Organization”, “Incompletion” (as they were listed in his 1987 paper).

Another set of design principles of relevance to performance management systems is the one proposed for supporting “emergent knowledge processes” (Markus et al. 2002) that posit the challenge of “information requirements that are complex and include general and situated knowledge distributed across experts and non-experts” (ibid). The six design principles proposed by Markus et al. are stemming also from the sociotechnical tradition and are the following: “Design for Customer Engagement by Seeking Out Naive Users”, “Design for Knowledge Translation Through Radical Iteration with Functional Prototypes”, “Design for Offline Action”, “Integrate Expert Knowledge with Local Knowledge Sharing”, “Design for Implicit Guidance Through a Dialectical Development Process”, “Componentize Everything Including the Knowledge-base”.

Both sets of principles put special emphasis to preparedness for change: Cherns caters for this with two principles (“Transitional Organization” and “Incompletion”) while Markus et al. address change “Through Radical Iteration with Functional Prototypes”. None of the two elaborate on the control and stabilization requirements that are essential for mature, efficient solutions.

3. AN EMPIRICAL CASE STUDY

3.1 Case Background and Method

In what follows, we present a reconstruction of a bank’s experience with performance management. This reconstruction is based on the accounts of key participants that are internal or external to the organization (bank: middle management, IT support, branch personnel, externals: management consultants and developers) and on the review of documentation (internal project documents, reporting to the top management, guidelines for bank personnel). We follow the interpretive case study research approach (Klein and Myers, 1999; Walsham, 2006). Our aim is to gain insight through tracing assumptions and problems of involved actors as the case unfolds, echoing inevitably our own “reading” of the story.

The commercial mechanism of the bank is based on a widely distributed branch network of more than four hundred branches coached by “area managers” that manage groups of 25 to 35 branches. Overall planning and direction originates from central divisions. A strong management reporting and accounting division undertakes at the tactical level most of the coordination of central divisions’ directions. Target setting and monitoring is the main performance management mechanism and up to the year 2000, target setting was tightly linked to the annual budgeting process: targets were directly derived from the annual budget and they were monitored with an IT application that collected information directly from the General Ledger. Branch Managers, Area Managers and the Financial Department of the bank had access to this application. The bank used only accounting measures as targets since this was the only easily retrievable data set available. Data not related to accounting (such as: number of new customers, number of complaints filed) would have to be retrieved from various databases and then handled with the local use of spreadsheets. In order to improve control over branch performance, align branch activities to the bank strategy and enhance their ability to contribute to the business, the management decided to introduce new performance management processes and establish continuous monitoring of a broader set of performance. There is a wealth of literature on the limitations of traditional approaches (Hope and Fraser, 2000; Jensen, 2003, Neely et al., 2003) that explain the rationale for going for a new system. Common directions found in the literature...
include: devolution of authority more effectively to the front line, creation of a single and widely accessible view of metrics, acting upon environment shifts before financial outcomes occur.

3.2 Case Analysis

In brief, the key objectives set at the outset of the endeavour, where: a) broaden the set of information used, b) expand the circle of stakeholders involved in the process accommodating diverse perspectives and c) facilitate two-way communication with the front-line. The approach followed can be described as “evolutionary”: early stage definitions of requirements were used to design and develop early system capabilities that were put to use. The early system capabilities supported the development of more detailed requirements and consequently, more advanced functionalities and processes. This led to cyclic refinements and enhancements of both the organizational and technological components of the system. There was a very pragmatic reason for adopting this approach. The whole effort started at the end of the third quarter of 2001 and there was very little time available before the launching of the traditional budgeting process. Missing the opportunity to apply it to the forthcoming business circle would result to postponing implementation for a whole year. So, the idea was to put some core capabilities in place as soon as possible and allow room for augmenting and refining within the year to come.

The functionality that was gradually developed for the new system was pretty sophisticated. The substantial increase in the number of actors to be involved in the implementation of the new system created the need for presenting diverse data subsets, relevant to each party’s role in the process. For example, the financial division would not accept any broadening of the targets if that meant that its ability to monitor the branch network’s financial performance would be impaired. Since not all financial measures represented actionable sales targets, there was a need to offer special data views for this division averting branch overload with information. Therefore, the new system would have to accommodate, in a flexible manner, different views of the performance information, depending on the viewer. The requirement for accommodating different views was mandated also by the dominant bank culture, according to which the access to performance information is clearly delineated by formal organizational structure and reporting lines. For example, each branch was only allowed to view its own targets and performance results, area managers were restricted to information for their own area, product divisions to their own products etc. Thus, the need for information representations that allowed comparison among peers without disclosing actual results had to be satisfied. Also, specific tools for the facilitation of two-way communication were designed. The idea was building mutual understanding and creating a learning environment within the bank sensitising headquarters’ employees to branch local conditions. In that respect, it was important for central divisions not only to access the outcomes of efforts made locally but also to have some descriptive information of the local conditions. Figure 1 that follows illustrates the performance management system that was put in place.
As the bank was at the same time embarking in a large scale MIS project to reap the fruits of an ongoing ERP implementation the idea was that performance management would eventually be supported by the wider platform while the prototype would be in place for the transition period. The prototype was prepared as a web-based application that would be accessed via the bank intranet. In order to minimise technical hitches, a mature and simple infrastructure was selected; the application would use a relational database and the views required would be dynamic web pages. The first version of the new set of metrics, processes and tool were launched four months from the project outset. During the years to come, a series of adjustments were made to both the processes and the tool. Some of them are the well-anticipated changes like: accommodating organizational restructuring, and introducing new approaches such as rolling target-setting fashioned by industry leaders. Another set of adjustments was made towards the achievement of greater flexibility and control of the tool without involving IT personnel (e.g. being able to define compound algorithms for aggregating data over time or at different organizational levels, being able to assign attributes to metrics that enable them to appear and disappear from users’ screens triggered by centrally defined events, etc.). At the same time, the ERP kept expanding and the use of Business Intelligence tools grew significantly within the bank. After a long period of “perpetual piloting” the bank specifications for “performance management” became mature enough for transferring to the integrated ERP solution. In Table 1 we summarise the different provisions for change and control of the two implementation approaches presented and we comment on them using the case study experience.

Table 1. Different provisions for change and control in two different approaches: drawing parallels with the case study

<table>
<thead>
<tr>
<th>Socio-technical approaches</th>
<th>Tailorable fully integrated system approaches</th>
<th>Drawing parallels with this case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cherns’ Principles, Markus et al. Design Theory for ERPs)</td>
<td>(ERP approaches)</td>
<td></td>
</tr>
</tbody>
</table>

**Provisions for Change:**


**Provisions for Control:**

Cherns: “Variance Control”, “Power and Authority”, “Support Congruence”

Markus et.al: “Integrate Expert Knowledge with Local Knowledge Sharing”

**Provisions for Change:**

ERPs can be configured to support the majority of business processes. Depending on the provider some of them have a basic core and rely on the ease of customizing to develop sophisticated functionality while others are more comprehensive and in simple cases can be put to use without writing any code (so for them in simple cases no adaptation is required). All solutions provide flexibility and configurability - some are more efficient than others.

**Provisions for Control:**

The discipline required with complex large scale systems is well support by ERPs. System configuration changes are controlled to ensure stability and reliability. Control mechanisms cover verification, validation, audit and security management.

Most socio-technical design “guidelines” seem to have been inadvertently followed. E.g.: building for immediate usefulness, implementing the minimal critical specifications and iterating fully functional prototypes. Modularization was achieved in a spiral way designing from the core to the periphery of the functionality. Simplicity and minimal design for immediate use are incompatible with the ERP approach. The benefits of the mature and robust integrated solutions can be reaped when such systems are introduced after clear views on the key needs of the organization have been established. In our case, the prolonged experimentation period signifies the initial lack of maturity that led to a series of explorations. These explorations would have been too expensive or cumbersome in a fully fledged solution but here were possible without a lot of hassle and led to the gradual crystallization of functionality needs. Still, as the “fully functional prototype” fulfilled its role, the organisation can have a series of benefits migrating to the well governed performance management component of the ERP suite which is already in place.
4. DISCUSSION AND CONCLUSION

The key lesson of this case study is that in an era of fluidity, blurring boundaries and ambiguity in everyday work, organizations might have to combine more than one approach in order to build systems that suit their needs. An initial period of iterative piloting can be planned to lead to some mature specifications to be inscribed in a robust and durable system. Adaptability is a key feature of all systems implemented today but this adaptability does not mean that organizations that have no mature articulated needs can successfully adopt fully fledged systems hoping to adjust them along the way. Although the functionality that allows system adaptation is built-in in such systems, it is frequently reported that adaptation can be lengthy, expensive and full of complications (Davenport, 1998; Robey et al. 2002; Kallinikos, 2004; Dechow and Mourtisen, 2005). These problems can be reduced if a preparation and experimentation phase using simple technologies precedes the introduction of the fully fledged solution. Even though modern systems allow full customization, exploring needs and requirements while large scale complex systems are deployed and put in use can result to lengthy and costly reworking. Indeed this seems to be the root in many reported troubled implementations: organisations get either trapped to naïve specifications developed in early stages to avoid expenses and risks of system destabilisation or get into tedious and perplexed reworks of large scale systems that are not designed to be continuously changed. It seems evident that only after a certain level of maturity has been reached an effective mix of organizational change and software configuration can take place to reap the advantages of the mature ERP solutions (Holland and Light, 1999; Milford and Stewart, 2000; O’Leary, 2000). Furthermore, it seems that a not fully institutionalised system launch may be an enabling factor for users expressing their needs in an “open culture”, which encourages admitting mistakes, self-learning and adaptation.

In this paper we highlighted the adaptability aspects of information systems design. We posit that “design for adaptability” is of particular relevance in today’s fluid environments and we illustrate the adaptability needs that emerged during the first years of implementation of an electronic solution for performance management. Clearly, organizations have to manage the tension between change and control when aiming for systems that can support complex and evolving work activities. A pragmatic approach for addressing the challenges can be the combination of “sociotechnical design” that facilitates system adaptability by establishing change mechanisms in the early phases with the customization/parametrization of modern Enterprise Resource Management Systems (ERPs) as requirements mature.

REFERENCES


SAP. (2006), Top ten reasons customers choose SAP for Business Transformation, *Brochure*: (1/06)

SAP. (2009), Financial Excellence Solutions, *Brochure*: 50 092 610 (09/02)


ALGORITHM TO OPTIMIZE THE POWER CONSUMPTION OF SCALABLE CLUSTERS GUARANTEEING RESPONSE TIMES

Daniel F. García, Ramón Medrano and Joaquín Entrialgo
University of Oviedo - Department of Informatics - 33204 Gijón, Spain

ABSTRACT
The power consumption of large computer clusters greatly affects the exploitation costs of datacenters. Therefore, the utilization of power minimization algorithms is essential for a profitable exploitation of datacenters. Few of the available algorithms guarantee a minimum quality of the service (QoS) provided by the cluster, and many require complex models of the cluster and of the workload evolution. Rather, they solve optimization problems off-line and tabulate results to be used on-line. In contrast, our algorithm uses only two input-output models for each computer type of the cluster. These models characterize the response time (QoS) and the power consumption as a function of the workload, and can be easily obtained from a load experiment. Based on these models, the algorithm works on-line maintaining only the most power-efficient computers which can provide the service with the desired quality turned on. The deployment simplicity of this new algorithm will contribute favorably to the integration of Green IT policies in datacenters.

KEYWORDS
Power consumption optimization, Response time guarantee, Scalable cluster management, Datacenter management, Green IT.

1. INTRODUCTION
The power consumption of large clusters is a significant part of the exploitation costs of datacenters. Power saving policies are necessary, but any viable policy must also guarantee the quality of the services (QoS) provided by the cluster.

In this paper we present a simple and effective algorithm to optimize the power consumption of any scalable cluster. The algorithm turns on nodes of the cluster as the load increases and turns them off as the load decreases. The algorithm maintains only the nodes needed to process the load with the required quality of service turned on. The stability and the behavior of the algorithm have been checked in a real cluster executing a synthetic transactional application.

2. THE SCALABLE CLUSTER
The algorithm presented in this paper optimizes the operation of scalable clusters. In this section, we present the kind of cluster for which the optimization algorithm has been designed. It is composed of multiple computing nodes which can either be identical (homogeneous cluster), or varied (heterogeneous cluster).

All the nodes provide the same set of web services to the clients of the cluster. To work with the cluster a client opens a session in the cluster, and then invokes services to carry out transactions.

The load supported by the cluster at any instant can be measured in sessions, that is, by the number of clients connected to the cluster at that instant. In order to balance the total load among the nodes of the cluster, a special node, called the distribution node, runs an algorithm to assign each new session to a specific computing node of the cluster. The new session is executed in the assigned node until it finishes. The migration of sessions between computing nodes is not considered in this work.
The distribution node operates as a communication tunnel between two TCP connections. One connection is maintained between the client and the distributor, and another between the distributor and a node of the cluster. The client sends requests through its connection to the distributor, which are then sent by the distributor to the assigned computing node. After processing the transaction, the node sends the response to the distributor, which sends it to the client. Figure 1 shows a scheme of this kind of cluster.

The optimization of the operation of a cluster like the one described above must fulfill the following basic requirements:

- The response time of transactions must be under a specified limit, commonly called SLA (Service Level Agreement). Often, the SLA is the maximum value allowed for the 90 percentile of the response time of transactions.
- The power consumption of the cluster must be as low as possible, while allowing the fulfillment of the SLA.
- The availability of the services provided by the cluster must be maximized.

3. PREVIOUS RELATED WORK

Before showing the design of the new algorithm, we present a review of the current available proposals for power saving in clusters. Algorithms based on dynamic provisioning of cluster nodes was initially studied by Chase and Doyle (2001). They proposed a research agenda to deploy a power efficiency update on Internet servers. In this way, Pinheiro et al (2001) developed an algorithm to concentrate the load of the cluster by turning nodes on and off dynamically.

Later, several algorithms integrated the dynamic provisioning of nodes with the dynamic voltage scaling (DVS) for the processors, operating on homogeneous (Elnozahy et al, 2002) and heterogeneous (Bertini et al, 2010) clusters. We find a major problem with these algorithms: they do not consider the SLA explicitly in order to manage and minimize power consumption without violating the SLA. Nor do they take into account the energy efficiency of the computing nodes being used to run the workload (Ranganathan, 2010).

There are also proposals that consider the SLA explicitly. The proposal of Chen et al (2005) is suitable for homogeneous clusters and uses a prediction model in order to fulfill the SLA. Another approach that considers the SLA explicitly is proposed by Heat et al (2005), but it does not use appropriate metrics for measuring it. Kusic et al (2010) proposed a system that maintains the desired QoS using dynamic virtual and physical machine provisioning. This system fulfills all the requirements of the SLA, but its major drawback is its complexity. Additionally, optimality analyses of different turn on/off strategies have been done (Gandhi et al, 2010) and models have been integrated with virtualized environments (Petrucci et al, 2010).

Summarizing, the algorithms analyzed do not consider the SLA in order to guarantee the QoS, or are very complex to implement in widespread systems. For example, some of the algorithms analyzed require a previous phase to model the cluster response, or need the use of specialized frameworks to implement the
power management policy. Metrics used to measure the QoS provided to users are often inexistent or inadequate, and the use of static prediction methods does not allow the system to evolve during its operation.

We propose a simple and effective algorithm to optimize power consumption of large clusters by modeling the system as a black box that is unaware of the internal structure of the cluster, taking decisions based strictly on the QoS provided to the users and the power efficiency of the nodes.

4. THE OPTIMIZATION ALGORITHM

The optimization algorithm we propose runs in the distribution node to take the proper scheduling decisions about the sessions. In order to take appropriate optimization decisions, the algorithm must have knowledge about the response time of each node ($R_i$) as a function of the load (number of concurrent sessions, $N_i$) that it supports and also about the power efficiency of each node.

The response time of each node is measured off-line with increasing loads and adjusted by a third order polynomial. Figure 2 shows the results for three different nodes of a heterogeneous cluster.

The power consumption of each node is also measured off-line with increasing loads and recorded as a table. Figure 3 shows examples of these curves for different nodes of a cluster. The low initial values of the curves are due to the DVS, which reduces the power consumption of processors when they are operating with low loads.

The left curve of Figure 3 clearly corresponds to an inefficient node. The power consumption with very few sessions is only slightly lower than with many sessions. The central curve corresponds to a more efficient node; the power consumption with no load ($\approx 100$W) is half of the consumption with full load ($\approx 200$W), and the consumption increases with the load until saturation. Finally, the right curve corresponds to a highly efficient node. The power consumption with no load is very low and grows progressively with the increment of the load.

From these curves, other curves representing the sessions per watt supported by a node as a function of the total number of sessions under processing can be easily derived, and later used on-line to estimate the power efficiency of the node for any load regime.

The actions that the algorithm must take are analyzed below. In order to limit the response time of transactions, the algorithm must distribute the incoming sessions optimally between the available nodes of the cluster. Of course, if the number of incoming sessions exceeds the total capacity of the cluster, the algorithm will reject sessions, thereby working as an admission controller.
In order to optimize the power consumption of the cluster, the algorithm must turn on nodes of the cluster when the load (number of sessions) increases and turn off nodes when the load decreases. When the cluster is heterogeneous, the algorithm must turn on the most energy efficient node of those available, and turn off the least energy efficient node.

Finally, in order to maximize availability, the algorithm is informed when the connections between the distributor and a node are broken. This indicates that the node has just failed. In this situation, the distributor re-opens the sessions in other available nodes of the cluster, and then turns on a new node.

In this architecture, the distribution node is a “single point of failure”. To increase its reliability and the availability, a fault tolerant cluster of two nodes with a shared cabinet of disks can be used. In this fault-tolerant architecture one node operates as the primary working node and the other as the secondary hot-spare node.

The question now is to define when the optimization algorithm must be executed to develop new optimization actions. Under normal operating conditions, the algorithm must be executed when the load increases and decreases. Under failure conditions, the algorithm must be executed just after the detection of the failure.

5. OPTIMIZATION ACTIONS WHEN LOAD INCREASES

Immediately after the distributor receives a request from a client to open a new session in the cluster, the optimization algorithm starts its operation. Figure 4 shows the steps of the algorithm. First, the algorithm calculates the capacity of each node, which is defined as the maximum number of concurrent sessions ($M_i$) that it can support without violating the SLA. The value of $M_i$ is obtained from the intersection of the polynomial curve that represents the response time with the horizontal line that represents the SLA. This analytic procedure is graphically represented in Figure 2.

Next, the algorithm calculates the utilization of each node ($U_i$), which is a fraction of the total computational capacity of the node that is being used. It is computed dividing the number of sessions in the node by the capacity of the node. The algorithm then applies admission control if the utilization of all the nodes is equal to or greater than 1.0 and there are no available turned off nodes.

After admitting a session, the algorithm must assign it to one of the available working nodes. In order to reach the best possible balance of the load among the nodes, the algorithm always selects the node with the lowest utilization. The algorithm also updates the utilization of the node that receives the session.

These first steps of the algorithm guarantee the fulfillment of the SLA, and the next steps minimize the power consumption of the cluster. The algorithm checks if the utilization of all the working nodes of the cluster is equal to or greater than a maximum utilization threshold ($U_{\max}$). If so, the algorithm will turn on the most power efficient node of the available (inactive) nodes. The $U_{\max}$ threshold must be slightly under 1.0.

6. OPTIMIZATION ACTIONS WHEN LOAD DECREASES

Immediately after the distributor detects that a client has finished his session and thereby closed his connection, the optimization algorithm is activated. In this case some of the steps differ from the case of load increments. Figure 5 shows a flow diagram of the steps of the algorithm when the load decreases.

Firstly, the algorithm updates the utilization of the node that has just lost a session. The next step is to minimize the energy consumption of the cluster. The algorithm checks if all the working nodes of the cluster have a utilization equal to or less than a minimum utilization threshold ($U_{\min}$). If so, the algorithm checks if there are two or more nodes working. The algorithm always leaves at least one node working to serve the requests of the clients. If there are more than one node working, the algorithm estimates their power efficiency and puts the less power efficient node in offloading state. The distributor never assigns new sessions to a node in offloading state. As the remaining sessions of the node finish, the node will gradually lose its load.
7. TRACKING THE STATES OF THE CLUSTER

During operation, the algorithm must track the states of all the nodes of the cluster. Figure 6 shows the states of any node and the transitions between them. Each state is represented by a name and its associated number. The nodes usually remain in the “on” and “off” states for long periods, because these are the stationary states. During the transitions between these two stationary states a node can be in any of the transitory states.

When the algorithm takes the decision to turn on a node, it executes the corresponding action immediately by sending a Wake-on-LAN message to the selected node and changing the state of that node to “turning on”. After a few minutes, the node will be fully operational and ready to process user sessions. Then, the node sends a message to the algorithm to confirm its readiness. When the algorithm receives this message, it changes the state of the node to “on”. If the message is not received before the predefined period of five minutes, the algorithm will assign the state “broken” to the node.
When the algorithm takes the decision to turn off a node, it changes the state of that node to “offloading”. In this state the node will lose its load progressively, because the sessions processed in the node will finish and the distributor will not assign new sessions to the node. When the node loses its last session, the algorithm executes the action to turn off the node, and after a predefined period of time, usually four minutes, the algorithm will assign the state “off” to the node. In addition, during the offloading period, if the load supported by the cluster grows quickly, the algorithm can assign the state “on” to any node that is offloading.

If a node fails while operating in any of the states, the algorithm should change its state to “broken”. But the algorithm is not always capable of tracking the state of the nodes properly. A failure during the “turning on” state can be detected as a time-out. During the “on” and “offloading” states a break of the connections of the sessions renders the node unavailable. The algorithm can detect this situation easily and assign the state “broken” to the node. However, detecting a failure while the node is turning off is impossible. The algorithm will detect the failure later, when it tries to turn on the node without success. Furthermore, accidental faults can appear. As an example, the power or network cables of a node could be disconnected during cleaning operations. Later, this will provoke a failure when the algorithm tries to turn the node on. After the reparation of the fault, the operator must notify the algorithm that the node can be used again, changing its state to “off” manually.

Correct tracking of the state of the nodes by the algorithm is essential because the states define the set of nodes that are used to take turn on/off decisions based on their utilization and the type of actions that the algorithm can carry out with a node.

![State transition diagram of the nodes](image)

**Figure 6. State transition diagram of the nodes**

8. **EXPERIMENTAL RESULTS**

The experimental environment consists of a cluster of four computing nodes and a distribution node. The nodes are HP ProLiant DL380 G6 servers, which make up a homogeneous cluster. Each node of the cluster has the capacity for 102 sessions with a 90-percentile response time of 2 seconds. In order to demonstrate the correct operation of our optimization algorithm, Figure 7 shows the results of an experiment in which the cluster supported first an increase and then a decrease in the number of concurrent sessions under processing. The next paragraphs describe the evolution of the curves of Figure 7.

During the first 1690 seconds, the workload consisted of 90 sessions. Only Node 1 was active and it had a utilization of 0.88. In order to show how the algorithm turns on new nodes, at 1690 seconds the workload was increased to 110 sessions. As the utilization of the only node active surpassed the $U_{\text{max}}$ threshold (0.95), the algorithm turned on a second node, increasing the capacity of the cluster to 204 sessions, as shown in the second graph from the top in Figure 7. The utilization of both nodes stabilized at 0.52. Next, at 2420 seconds the workload was increased to 190 sessions; the utilization of the nodes grew to 0.85. As can be seen in the figure, the distributor balanced the workload between both nodes perfectly.
The next change in workload, at 3490 seconds, brought the number of sessions injected to 210, over the maximum capacity of the cluster with two nodes (204 sessions). The utilization would have surpassed the $U_{\text{max}}$ threshold if no node were turned on; therefore, the algorithm turned on another node and the workload was balanced between the three active nodes.

![Graph of operational variables](image)

Figure 7. Evolution of the operational variables of the cluster during a load experiment

The workload was increased again and the fourth node was turned on. At 7630 seconds, the workload grew to 400 sessions and the utilization of all the nodes was slightly lower than 1.0, but the admission control did not reject any sessions. In order to assess the functionality of the admission control feature, the workload injected to the cluster was increased to 500 sessions at 8390 seconds. As can be seen in the top graph of Figure 7, the algorithm limited the workload processed by the cluster to 408 sessions, which is the maximum capacity of the cluster in order to guarantee an SLA of 2 seconds.

In order to show how the algorithm works with a decreasing workload, the workload was decreased in several steps from 400 sessions at 9390 seconds, to 290 sessions at 12410 seconds. The utilization of all the nodes decreased accordingly, but no node was turned off because the utilization of the remaining nodes
would have been higher than the $U_{\text{max}}$ threshold (0.7). As can be seen in Figure 7, the workload was balanced perfectly.

At 13450 seconds, the workload was decreased again, leaving 210 sessions. As the utilization of three operational nodes would have been less than the $U_{\text{max}}$ threshold (0.7), the algorithm turned off Node 2, selected at random because the four nodes are identical. The utilization of Node 2 dropped progressively until it reached zero, while the utilization of the other nodes grew gradually when they received new sessions no longer handled by Node 2. In the last part of the experiment, the workload was decreased and the turning off of another two nodes can be seen.

The evolution of the 90 percentile of response time is shown in the bottom graph of Figure 7. The percentile is under the SLA (2 seconds) except in the period between 6450 and 11450 seconds, in which the cluster is operating very near and over its total capacity. In this period the percentile is around 2 seconds, so the algorithm guarantees the maximum percentile allowed while minimizing the power consumption.

9. CONCLUSIONS

In this work, we have presented a simple and effective algorithm to optimize the power consumption of clusters while guaranteeing the SLA. The main contribution of our algorithm is that it is based on input-output models of the response time and the power consumption of the nodes of the cluster, which can be easily obtained from a load experiment.

Whereas most of the previous algorithms for power optimization require the development and adjustment of complex models of the cluster, our algorithm requires only two configuration parameters: the maximum and minimum utilization thresholds of the nodes, which are very easy to establish.

ACKNOWLEDGEMENT

This research work has been partially supported by the project TIN2008-06045-C02-01 of the Spanish National Plan of Research, Development and Innovation.

REFERENCES

THE IMPACT OF ENVIRONMENTAL DYNAMISM ON THE EFFECTIVENESS OF SISP: A COMPARATIVE STUDY OF MANUFACTURING AND NON-MANUFACTURING MEDIUM ENTERPRISES

Luka M. Musangu and Ray M. Kekwaletswe
Faculty of ICT, Tshwane University of Technology, Pretoria, South Africa

ABSTRACT
The effectiveness of SISP is an important research aspect in the management information systems. This study investigates the influence of environmental dynamism in the relationship between strategic information systems planning process and strategic information systems planning success in manufacturing and non-manufacturing medium enterprises. Based on responses from 118 South Africa CIOs or head of IS/IT, our empirical investigation provided the following results: of all five hypotheses tested, four were supported and one was not. The findings of study demonstrate that dynamic environment strongly influences the relationship between SISPP and SISPS in manufacturing enterprises than in non-manufacturing. The implications and limitations of this study are also discussed.

KEYWORDS
Environmental uncertainty; Environmental dynamism; Strategic information systems planning process; Strategic information systems planning success

1. INTRODUCTION
External environment and strategic information systems planning (SISP) have received significant scholarly attention in the past two decades. The previous researches identify the competitive pressure of the external environment as an important facilitator for SISP use (Teo and King, 1997). The importance of SISP has increased in relevance among information systems researchers and practitioner and it remains as one of the top issues in information systems area. In today’s dynamic market environment, business success relies heavily on the effectiveness of SISP, which is the first stage of IS planning model. According to Bechor, Seev, Moshe and Glezer (2010) SISP refers to the process of strategic thinking that identifies the most desirable IS on which the firm can implement and enforce its long-term IT activities and policies. It is a mechanism for assuring that IS activities are aligned with the organisation’s evolving needs and strategies (Bechor et al., 2010; Sabherwal and Chau, 2001).

There are many different ways of defining small, micro and medium enterprises (SMMEs), varying from size to revenue. The present study has been conducted within the South African context. The appropriate strategy here would be to adopt the definition provided by the National Small Business Act of 1996, as amended by Act 26 of 2003, which defines SMME as: “Any entity, whether or not incorporated or registered under any law, consisting mainly of persons carrying on small enterprise concerns in any economic sector and established for the purpose of promoting the interests of or representing small enterprise concerns, and includes any federation consisting wholly or partly of such association and organisation”. From that definition the South Africa SMME sector is classified in four respective categories: survivalist enterprise, micro enterprise, small enterprise and medium enterprise.

According to Le Roux (2006), South Africa’s 2 million SMMEs represent 98% of the country’s total number of firms, employ 55% of the labour force and contribute 42% of the country’s wage bill. From these statistics manufacturing SMMEs remains an important sector contributing 18.6% to Gross Domestic Product (GDP) and employing more than one million people (South African, 2008).
The organisation target of the present study was the medium enterprises; it has been described as the largest of the SMMEs. It is owner controlled, the ownership and management structure is more complex and it employs a maximum of between 100 and 200 people (South Africa, 2008).

The purposes of this study is to empirically examine whether the presence of dynamic environment can strongly moderate the relationship between strategic information systems planning process (SISPP) phases and strategic information systems planning success (SISPS) in manufacturing and non-manufacturing medium enterprises.

To describe the outcomes of this research, this paper is organised in the following way: The next sections describes the theoretical background of research variables and research hypotheses. Following that, there is a section describing the research methods. The analysis of the data together with a discussion of the results, implications and conclusions of this research are discussed in the last section.

2. THEORETICAL BACKGROUND, RESEARCH VARIABLES AND RESEARCH HYPOTHESES

2.1 Theoretical Background

Previous IS researchers have found out that external environments defy today’s business managers. It forces them to be more careful strategic planners (Newkirk and Lederer, 2007). Many authors contend that it is under pressure of external environment that strategic planning could come to the fore. Supported by the information uncertainty perspective, they argue that external environment increases the need for information gathering and therefore SISP (Cohen, 2001). SISP can be used to reduce uncertainty, and is increasingly recognised as a necessity for organisations to survive and be able to tackle uncertainty quickly and robustly in order to sustain and enhance business competitiveness (Brown, 2008).

Sabherwal and King (1992) suggested that more extensive SISP would be more successful because it would help planners understand the impact of the external environment and better respond to it. Teo and King (1997) identified the competitive pressure of the uncertain environment as an important facilitator for SISP use. Hopkins and Hopkins (1997) found that environmental factors emerged as a more influential determinant of SISP. Choe et al. (1998) empirically showed a positive relationship between perceived environmental uncertainty and SISP. Grover and Lederer (1999) studied the influence of environmental uncertainty on the strategic use of information systems and found out that environmental uncertainty is an important predictor of SISP practices.

The present research contained one independent variable: Strategic Information Systems Planning Process (SISPP) with five phases namely: strategic awareness, situation analysis, strategy conception, strategy formulation and strategy implementation. One dependent variable: Strategic Information Systems Planning Success (SISPS) with four dimensions, namely: alignment, analysis, cooperation and improvement in capabilities. And one moderated variable: external environment characterised on the purpose of the present study by the environmental dynamism.

2.2 Research Variables

2.2.1 Strategic Information Systems Planning Process (SISPP)

SISPP has been the core of much interest in the previous SISP literature but there is no agreement about what the formal methodology is. Mirchandani and Lederer (2008) stated that there was no industry standard for SISP approaches, as organisations, possibly with the help of external consultants, developed their own in-house approaches. Whatever approach or combination of approaches is chosen, it will then have to be adapted to suit the environment, culture, experience and skills existing within the organisation.

SISPP, has been defined in the context of the current research as a process of investigating external and internal organisation’s environments with the aim of positioning information systems in relation to the business. Issues and opportunities identified in this assessment should result in generating strategic directions of future information systems investments (Zijad, 2007).
Previous studies have developed and used SISPP model which aim to increase the practical usefulness of information systems (IS) planning within organisations. This SISPP model contained three process elements, namely: phases, stages, and modules. The phases are generic strategy formulation steps that can be applied to any corporate strategy development process. There are five phases in SISPP model which are: strategic awareness, situation analysis, strategy conception, strategy formulation and strategy implementation. Each phase is divided into stages. Stages are considered to be semi-autonomous components of work which can be planned relatively independently. Stages are further divided into modules. Modules can either be units of work or collections of activities (Musangu and Kekwaletswe, 2011; Newkirk and Lederer, 2006; Mentzas, 1997).

The first phase of SISPP model called strategic awareness carries purpose to formulate major questions and provide answers relatively to the organisation and its competition. The situation analysis is the second phase. Its goal is to examine the existing business and IS situation in the organisation and identifies the specific strengths and weaknesses. The third phase is strategy conception. In this phase the IT purposes, opportunities for improvement and the important IT strategies are identified. The fourth phase, strategy formulation clarifies new business processes, new IT architectures, specific new projects and the priorities for new projects in terms of the functions, hierarchies and responsibilities. Final phase is called strategy implementation, in this phase the management approach, action plan, follow-up and control procedure are described in order to monitor and control the cost and to manage the implementation process (Newkirk and Lederer, 2007; Mentzas, 1997).

The observation of the extent to which an organisation carries out each phase and task may be used to assess the state of the SISPP.

2.2.2 Strategic Information Systems Planning Success (SISPS)

The success of SISP has been operationalised in terms of the achievement of organisation planning objectives as either a single or multi-factor construct (Mirchandani and Lederer, 2008; Segars and Grover, 1999; Raghunathan and Raghunathan, 1994).

According to Segars and Grover (1998) the assessment of SISPS cannot be reduced to such simple financial measures as return on investment, payback or internal rate of return. In this context, they built an instrument and empirically verified a second-order model based on these two perspectives for SISP success. Their instrument used four dimensions to measure SISPS, namely: alignment, analysis, cooperation and improvement in capabilities.

Alignment dimension was described as the linkage between IS strategy and business strategy (Grover and Segars, 2005). Analysis dimension has been described as the understanding of internal operations of the organisation in terms of processes, procedures and technologies (Newkirk, Lederer and Srinivasan, 2003). Cooperation dimension has been identified as the agreement between managers and users concerning development priorities, implementation schedules and managerial responsibilities. The purpose of cooperation is to avoid conflict which may destroy the implementation of strategic IS plan (Newkirk and Lederer, 2007; Grover and Segars, 2005). The last dimension was capabilities. It has been described as the improvement in the potential of the planning system. An effective SISP should have a high perceived level of contribution to various aspects of organisational effectiveness (Newkirk and Lederer, 2007; Grover and Segars, 2005; Newkirk, Lederer and Srinivasan, 2003).

The dimensions developed by Segars and Grover (1998) have been used to assess the SISPS in this research.

2.2.3 External Environment

External environment has been defined as physical and social determinants outside the boundaries of the enterprise that are taken directly into consideration (Chi, Jones, Lederer, Li, Newkirk and Sethi, 2005). External environment is an important concept in the SISP research but only little study has been conducted to investigate the influence of external environment on SISP (Brown, 2008). Nearly all previous external environment and SISP studies have integrated their stability aspect or environment uncertainty.

Environmental uncertainty is defined as the frequent and unpredictable changes in customer preference, technological development, and competitive behaviour perceived by managers (Lin and Ho, 2010). It has been regarded as the most important factor influencing managers’ decision making. Teo and King (1997)
characterised environmental uncertainty by the presence of environmental dynamism, environmental heterogeneity and environmental hostility.

For the purpose of this paper, external environment is characterised by the presence of environmental dynamism. All organisations operate in dynamic market and most of the time market’s realities are quickly become obsolete (Cohen, 2008) - taking into account that the dynamic aspect of environmental dynamism has been characterised by the rate of change and innovation in the industry, as well as the unpredictability of the actions of competitors and customers (Lederer and Sethi, 1998). Teo and King (1997) found out that environmental dynamism was associated with greater levels of integration between business planning and SISP. Greater dynamism requires higher levels of cooperation and coordination between business strategies and IS (Brown, 2008; Grover and Segars, 2005).

Sabherwal and King (1992) view dynamism as the unpredictability of the external business environment, as evidenced by change and obsolescence. Dynamism affects the importance of information processing because as environmental dynamism increases, successful firms must make strategic decisions more quickly. Dynamic environments increase the firm’s need for information resources which can facilitate inter-organisational linkage by improving information processing capabilities such as the building of a knowledge base, the seeking out of new business opportunities, and the development and management of competitive initiatives (Grover, 1997).

The questionnaire developed and used previously by Teo and King (1997) has been used in this study to measure environmental dynamism.

2.3 Hypotheses

It is well known that the unpredictability of environmental change creates uncertainty, taking into account - the fact that environmental uncertainty challenges managers in decision making. Analysis in environmental dynamism might be expected to help manager in getting more knowledge about environmental changes. This would make it possible to understand and develop plans that are less vulnerable to consequences of that change (Newkirk and Lederer, 2006).

2.3.1 SISPP and SISPS under Environmental Dynamism

In dynamism environment context, managers can use SISP to reduce the rate and unpredictability of environment change. High level of environmental dynamism can increase the likelihood that manufacturing enterprises will identify opportunities for using SISP resources (Grover and Lederer, 1999).

According to Newkirk and Lederer (2006) the achievement of SISP in environmental dynamism is critical to planners. Analysis in dynamism environment might be expected to produce greater knowledge about competitors, resources, customers and regulators. Thus, the presence of environmental dynamism can strongly moderate the relationship between SISPP phases to success in manufacturing enterprises than in non-manufacturing enterprises. These arguments support the set of hypotheses:

H1: Environmental dynamism will strongly moderate the association between strategy awareness and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

H2: Environmental dynamism will strongly moderate the association between situation analysis and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

H3: Environmental dynamism will strongly moderate the association between strategic conception and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

H4: Environmental dynamism will strongly moderate the association between strategic formulation and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

H5: Environmental dynamism will strongly moderate the association between strategic implementation and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

H5: Environmental dynamism will strongly moderate the association between strategic implementation
3. RESEARCH METHODOLOGY

3.1 The Instrument

Demographic characteristics were measured by means of demographic information (DI), which included individual factors such as gender, age, educational level, IS experience, SISP experience, duration of employment with the current SMMEs, scope of SISP, planning horizons, IS employees and SMMEs primary industry.

The measuring instruments used in this study have been used in the previous studies (e.g., Mirchandani and Lederer, 2008; Newkirk and Lederer, 2007; 2006; Segars and Grover, 1998, Teo and King, 1997 and Mentzas, 1997).

The above researchers used the questionnaire instrument, which consists of five-point Likert-scales, in order to operationalise the following three constructs: SISPP, SISPS and environmental dynamism

3.2 Sampling Frame and Sampling Procedure

The sampling frame adopted was the 2009 edition of “Who Owns Whom in South Africa”, published by McGregor. This directory contains the names, titles, addresses of top computer executives in South Africa. The entities within the directory include small enterprises, micro enterprises, medium enterprises, large firms, educational institutions, hospitals and governmental agencies.

In developing a desirable sub frame, (i) all large firms, hospitals, educational institutions and governmental agencies were eliminated from consideration; (ii) the job titles of key informants remaining in the frame were examined as a means of determining the level of planning activity; (iii) medium enterprises with a senior executive carrying the job title of chief information officer, vice president, director of strategic planning, director of MIS or head of IS/IT established in Gauteng province were retained. This resultant sub frame contained 518 SMMEs. From this frame 350 SMMEs were chosen at random.

3.3 Data Collection

A questionnaire was mailed out to 350 CIOs, information systems managers or head of information systems in medium enterprises established in South Africa’s Gauteng province. Of the 350, 92 keys informant returned the survey. Out of the total returned (92), 73 were usable for analysis and 19 were not. In addition, a total of 48 surveys returned during the follow-up, 45 were usable for analysis and 3 were not. On the whole, a total of 140 responses have been received from both the initial and follow-up mailing, 118 were usable for analysis but 22 were not. Of the total, 22 responses were found unusable for analysis, 13 were discarded from analysis because some questions were not completed, 2 respondents returned with notes that their organisations do not participate in surveys and 7 returned with notes that their organisations were not using SISP. Thus, the gross responses rate of the research survey was 40 %, of which, 118 returns i.e., 34% were suitable for analysis. The gross response rate and usable response rate received for the present study is quite high compared with previous SISP studies conducted previously (Mirchandani and Lederer, 2008; Newkirk and Lederer, 2007; 2006; Chi et al., 2005)

4. DATA ANALYSIS

4.1 Common Method Variance, Response Bias, Reliability and Validity of Research Variables

The returned surveys were examined for non-response bias to confirm that the decision to respond was uninfluenced by non-random events or motives. If the decision to respond is random, then the timing of the response should not significantly influence the value of survey measures (Kerlinger and Lee, 2000). Multivariate analysis of variance was used to evaluate whether differences among early and late responders
were associated with different responses. The analysis indicated no significant differences in several key variables tested for the surveys. This is consistent with the absence of non response bias in the surveys.

Since dependent and independent variable data was collected from a single key informant (head of IS/IT). The head of IS/IT is typically seen as the most knowledgeable person to assess SISP, its context, and its outcomes (Premkumar and King, 1991). However, multiple subjects per organisation are preferred in order to reduce common method variance, which arises from using one individual and can account for a relationship between similar measures (Newkirk and Lederer, 2006).

Harman’s one factor test was used to test for the presence of common method variance bias. The results of this analysis on our data revealed 9 factors with an Eigen value greater than one and no single factor explained most of the variance. Such results are consistent with the absence of a significant variance common to the measures.

In order to assess the reliability of all research variables measures, internal consistency was calculated using Cronbach’s alpha which indicated the degree of internal consistency among the measurement items and is inversely related to the degree to which a measure is contaminated by random errors (Wang and Tai, 2003). As provided in table 1, the results in the present research indicate that all research variables constructs have an acceptable Cronbach’s alpha level.

### Table 1. Reliability coefficient of research variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of items</th>
<th>Cronbach's alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISPP</td>
<td>26</td>
<td>.76</td>
</tr>
<tr>
<td>Strategic awareness</td>
<td>5</td>
<td>.70</td>
</tr>
<tr>
<td>Situation analysis</td>
<td>6</td>
<td>.79</td>
</tr>
<tr>
<td>Strategy conception</td>
<td>4</td>
<td>.75</td>
</tr>
<tr>
<td>Strategy formulation</td>
<td>5</td>
<td>.82</td>
</tr>
<tr>
<td>Strategy implementation</td>
<td>5</td>
<td>.82</td>
</tr>
<tr>
<td>SISPS</td>
<td>30</td>
<td>.87</td>
</tr>
<tr>
<td>Alignment</td>
<td>8</td>
<td>.80</td>
</tr>
<tr>
<td>Analysis</td>
<td>8</td>
<td>.87</td>
</tr>
<tr>
<td>Cooperation</td>
<td>7</td>
<td>.82</td>
</tr>
<tr>
<td>Capabilities</td>
<td>7</td>
<td>.86</td>
</tr>
<tr>
<td>ENVIRONMENT DYNAMISM</td>
<td>4</td>
<td>.78</td>
</tr>
</tbody>
</table>

#### 4.2 Profile of Respondents

Summarise of responses for demographics data are provided for both manufacturing and non-manufacturing medium enterprises.

The percentages of respondent gender were 14% female and 86% were male in the manufacturing enterprises whereas 29% were female and 71% were male in the non-manufacturing enterprises. The mode of respondent age was between 36 and 45 years for the both categories. Most of the respondents had a postgraduate degree in both enterprises followed by some postgraduate school in manufacturing and 2 years college in non-manufacturing enterprises. The information systems experience was higher for non-manufacturing enterprises IS executives than for their counterparts in manufacturing enterprises. The SISP experience was not differing for manufacturing and non-manufacturing enterprises. More than 80% of respondents had a SISP experience ranged between 1 to 10 years. Only 3% had SISP experience over 11. The percentage of respondent working in current enterprise is higher in manufacturing enterprises than in non-manufacturing. 43% of manufacturing respondents were employed in the current enterprise for a period ranged between 11 and 15 years whereas 39% were employed in non-manufacturing for a period between 6 and 10 years. Enterprise was chosen as SISP scope for 96% of manufacturing and 70% of non-manufacturing. The planning horizon was normally distributed for the two samples with 71% of manufacturing using 3 years planning horizon and 52% of non-manufacturing in the same level. And finally, the results indicated that more than 70% of enterprise employed less than 10 staff in the IS department in both enterprises.

#### 4.3 Multiple Moderated Regressions (MMR)

To investigating the moderate influence of environmental dynamism in the relationship between SISPP phases and the success of SISP, the multiple moderated regressions was performed.
Two simultaneous moderated regression equations were carried to examine the moderated role of environmental dynamism in the relationship between SISPP phases and SISP success. For each equation, subjects were separated in manufacturing enterprises and non-manufacturing enterprises.

Both equations were statistically significant at \( p < .0001 \). The values of \( R^2 \) for non-manufacturing enterprises are higher than for manufacturing. Four of the five beta coefficients are significant for manufacturing enterprises and only two of the five beta coefficients are significant for non-manufacturing enterprises. The results provide in table 2 suggest that environmental dynamism moderate more the relationship between SISPP and SISPS in manufacturing enterprises than in non-manufacturing enterprises.

### Table 2. Results of multiple moderated regressions

<table>
<thead>
<tr>
<th>Independents variables: SISPP Phases</th>
<th>Dependent variables: SISPS</th>
<th>Moderated variable: (SISPP Phases * Environmental Dynamism)</th>
<th>Standardized coefficients</th>
<th>Standardized coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANUFACTURING ENTERPRISES (n = 28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategic awareness</td>
<td>.158</td>
<td>.143</td>
<td>.143</td>
<td>.136</td>
</tr>
<tr>
<td>Situation analysis</td>
<td>.267</td>
<td>.167</td>
<td>.167</td>
<td>.167</td>
</tr>
<tr>
<td>Strategy conception</td>
<td>.247</td>
<td>.205</td>
<td>.205</td>
<td>.205</td>
</tr>
<tr>
<td>Strategy formulation</td>
<td>.281</td>
<td>.198</td>
<td>.198</td>
<td>.198</td>
</tr>
<tr>
<td>Strategy Implementation</td>
<td>.744</td>
<td>.104</td>
<td>.104</td>
<td>.104</td>
</tr>
<tr>
<td>NON-MANUFACTURING ENTERPRISES (n = 90)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategic awareness</td>
<td>.126</td>
<td>.110</td>
<td>.110</td>
<td>.110</td>
</tr>
<tr>
<td>Situation analysis</td>
<td>.170</td>
<td>.103</td>
<td>.103</td>
<td>.103</td>
</tr>
<tr>
<td>Strategy formulation</td>
<td>.110</td>
<td>.057</td>
<td>.057</td>
<td>.057</td>
</tr>
<tr>
<td>Strategy Implementation</td>
<td>.599</td>
<td>.065</td>
<td>.065</td>
<td>.065</td>
</tr>
</tbody>
</table>

Significance at * \( p < .05 \), ** \( p < .01 \), *** \( p < .001 \), **** \( p < .0001 \)

### 5. DISCUSSION

The study objective was to empirically examine if the presence of dynamic environment can strongly moderate the relationship between strategic information systems planning process (SISPP) phases and strategic information systems planning success (SISPS). The results of hypotheses testing are summarised in table 3. Of the five hypotheses four hypotheses were statistically supported (H1, H2, H3 and H4) but one was not supported (H5).

### Table 3. Summary of results

<table>
<thead>
<tr>
<th>HYPOTHESES</th>
<th>MULTIPLE MODERATED REGRESSION</th>
<th>DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>*</td>
<td>Supported</td>
</tr>
<tr>
<td>H2</td>
<td>*</td>
<td>Supported</td>
</tr>
<tr>
<td>H3</td>
<td>**</td>
<td>Supported</td>
</tr>
<tr>
<td>H4</td>
<td>**</td>
<td>Supported</td>
</tr>
<tr>
<td>H5</td>
<td>ns</td>
<td>No supported</td>
</tr>
</tbody>
</table>

Significance at * \( p < .05 \), ** \( p < .01 \), *** \( p < .001 \), **** \( p < .0001 \)

Table 2 presents the empirical results of examining moderating role of dynamic environment in the relationship between strategic information systems planning process (SISPP) and strategic information systems planning success (SISPS). Two moderated regression equations were performed to test the study hypotheses. Of the five manufacturing enterprises hypotheses four were statistically significant and one was not. On the other hand, two were statistically significant in the non-manufacturing enterprises and three were not. These results make supported hypotheses H1, H2, H3 and H4 while H5 was not supported. It is well known that environmental dynamism and unpredictability of environmental change can challenge today’s managers in the decision making process and the use of planning process can reduce dynamic environment impact and leads to success (Newkirk and Lederer, 2006). The results suggest that environmental dynamism
strongly influence the relationship between SISPP and SISPS in the following phases: strategic awareness, situation analysis, strategy conception and strategy formulation in the case of manufacturing enterprises. Planners in manufacturing medium enterprises have to be careful in the following aspects: (i) the process of formulating major questions and provide answers relative to the organisation and its competition. (ii) The examination of existing business and IS situation in the enterprises and identifies the specific strengths and weaknesses. (iii) The identification of IT purposes, opportunities for improvement and the important IT strategies. And finally, the clarification of new business processes, new IT architectures, specific new projects and the priorities for the new projects in terms of the functions, hierarchies and responsibilities.

6. IMPLICATIONS FOR FUTURE RESEARCH

The study findings are especially relevant and supported previous research that environmental uncertainty was found as an important predictor of SISP (Nerkirk and Lederer, 2006). In fact, the present study demonstrates that planning process can leads to the successful of SISP and environmental dynamism impact the relationship between SISPP and SISPS in South Africa manufacturing enterprises.

Four of the five hypotheses were supported (H1, H2, H3 and H4) and one hypothesis was not supported. Future researchers should attempt to determine why the data did not support this hypothesis.

Future researchers could use more contextual factors to investigate the impact of external environment on the effectivenss of SISP.

This comparative study is important for CEOs, CIOs and planners in manufacturing medium enterprises. The increase of freedom in the world market, the removing of trade border in the regional community oblige today’s managers to know exactly what is the characteristics in is business environment. Most of the time these realities can change in relationship of each sector and context. The challenge for CIO in manufacturing sector established in South Africa will not be the same comparing to the CIO in the same sector established in Swaziland.

The findings of this study can raise manufacturing practitioners’ awareness of the phases of SISP. Practitioners should be knowledgeable about the four involved phases, but the individual item in each might suggest overlooked activities. These might impede the manufacturing from achieving its planning objectives and thus from realising greater value.

7. LIMITATIONS OF THE STUDY

There were some limitations associated to the study. Firstly, the present study used a cross-sectional research study, where the unit of analysis was observed at one point in time. While it provided a useful “snapshot” in collecting data over a period of weeks to help in understanding the phenomenon under investigation, it could not explain possible changes in the respondents’ attitudes over time. Generally, it is recognised that longitudinal studies (or at least a series of cross-sectional studies) can detect attitudes changes over time and allow stronger inferences to be drawn about the dynamic elements of behaviour (Ikart, 2005).

Secondly, the small number of manufacturing enterprises in the sample and data collected from a sample of medium South Africa enterprises may limit the results’ generalisation.

Despite these limitations, the present study provides valuable insights into the study of SISP and external environmental in South Africa. The limitations acknowledged above therefore provide some suggestions for further research.

REFERENCES


AN O(N) APPROXIMATION FOR THE DOUBLE BOUNDING BOX PROBLEM

Jörg Roth
Univ. of Applied Sciences Nuremberg - Kesslerplatz 12, 90489 Nuremberg, Germany

ABSTRACT
Minimal bounding rectangles are a simple and efficient tool for approximating geometries, particularly for accelerating spatial queries. If a spatial object fills a rectangular shape to a certain extent, then the minimal bounding rectangle is a reasonable approximation. Unfortunately some geo objects, such as streets or rivers have a small area but large bounding rectangles. In this paper we suggest an approximation with two bounding rectangles instead of a single one. Since the corresponding shape provides a better approximation, we get a greater average benefit. However, the computation of two bounding boxes with minimal area requires O(n log n) steps for n geometry points. That may be a crucial point for geometries with large amounts of geometry points. In this paper, we introduce an approximation that requires O(n) steps but only produces approx. 11% more false hits compared to the theoretical optimum.

KEYWORDS
Spatial data, geometric approximation, bounding box

1. INTRODUCTION
Bounding boxes are a simple and efficient tool to speed up geometric or spatial operations. For two dimensions, a bounding box (often called Minimal Bounding Rectangle) is the smallest rectangle that encloses a geometry such as a polygon or line string (fig. 1a). Although bounding boxes may have more dimensions, two dimensions are sufficient for typical spatial queries for geo objects on the Earth’s surface used by reverse geocoding, map painting or queries in current location-based services.

Bounding boxes provide only a rough approximation, but there exist efficient a priori checks for spatial properties. For example, if two geometries overlap, their bounding boxes overlap as well. Obviously, the other direction is not always true. Thus, we can use bounding boxes to check quickly whether or not some properties could potentially be true. If the bounding boxes pass the check, the corresponding exact geometry checks may fail, so bounding boxes can only reduce the set of candidates. Since an exact geometric test (e.g. if polygons overlap) may be time consuming, one goal is to have a small candidate set. Unfortunately for some spatial objects the bounding box area is very large compared to the original geometry. Line objects such as rivers or roads, in particular, may generate very large bounding boxes.

One way to reduce the candidate set is to reduce the surface area of the approximating construction. The better the approximation of a geometry, the lower is the number of false positive hits by a priori checks. To both retain the idea of bounding boxes and to simultaneously reduce the candidate set, we suggest Double Bounding Boxes (DBBs, fig. 1b). They consist of two simple bounding boxes, which may overlap. The area of the double bounding boxes fully encloses the geometry. To distinguish the single bounding box from the double bounding boxes described in this paper, we call the traditional box Single Bounding Box (SBB).

The greatest benefit have DBBs with a minimal total area. For the construction of such DBBs, there exists an algorithm which requires O(n log n) steps for n geometry points (Roth, 2011). DBBs with minimal areas reduce the amount of false positive hits to 47.8% of the single bounding box false hits for typical geo data queries. In this paper we introduce an approximation called Quick DBB, that only requires O(n) steps without a significant drawback concerning false positive hits for realistic data.
2. RELATED WORK

Bounding boxes are used mainly for three purposes: 1. they can be used to speed up spatial comparisons; 2. they are used to create metadata of spatial objects; and 3. they may represent objects inside spatial indexes.

In addition to bounding boxes we have approximating shapes such as circles, ellipses or convex hulls (Gartner & Schonherr, 1997, Barequet & Har-Peled, 1999, Hill, 2006, Yang et al., 2008, Welzl, 1991). The main goal is to provide a quick a priori test for a spatial condition. If this a priori test is true then it makes sense to perform an additional check which is more exact but also more costly. If the a priori test was not true, then no further checks are needed.

A bounding box does not necessarily have to have aligned axes. If a box is rotated, it may be smaller and still enclose a geometry. This leads to the construction of Oriented Bounding Boxes (Yuan et al., 2006), often used in the area of game engines for collision detection or Optical Character Recognition (OCR). Such boxes can be computed using so-called rotating calipers (Toussaint, 1983). We did not follow that approach in this paper because it does not suit our intended usage.

To give an impression of how an aligned bounding box may speed up a test, consider a case where we have stored two polygonal geometries A and B. In addition to the polygon’s vertices $p_i$, the bounding box is also stored, defined by two corners $c_1, c_2$ where $c_{1x} = \min(p_{ix})$, $c_{1y} = \min(p_{iy})$, $c_{2x} = \max(p_{ix})$, $c_{2y} = \max(p_{iy})$. We now could query whether or not A and B overlap, whether A is completely inside B (or vice versa), whether A and B only touch (only share a line or point), or whether A and B are disjoint. For each of these geometric properties there exists a bounding box test. To test whether the geometries overlap, for example, we check
\[
\text{overlap}(\text{SBB}_A, \text{SBB}_B) = \begin{array}{c}
\text{true} & \text{if } c_{Bx} \leq c_{Ax} \text{ and } c_{By} \leq c_{Ay} \text{ and } c_{Ax} \leq c_{Bx} \text{ and } c_{Ay} \leq c_{By}
\end{array}.
\]

The drawback of this approach is the number of successful box checks where the exact test fails (false positives). This happens particularly when the actual geometry’s area is small compared to the box area.

It is important to note that the a priori check for aligned bounding boxes (in contrast to oriented bounding boxes) can even be processed inside a standard database query. Thus, an approach to use non-spatial databases for spatial objects is to store the box parameters in table columns and the exact geometry as a binary large object. Only if the database passes the a priori check, and only then, the geometry is deserialised from the table and the exact check is performed (Roth, 2009).

Instead of a bounding box we could also use a bounding circle, defined by centre $c_x, c_y$ and radius $r$. To compute an optimal centre (that leads to a minimal $r$) is not trivial (Elzinga, 1972) and requires a longer computation than a bounding box, but there exist approximations (Welzl, 1991, Hearn & Vijay, 1982). Once the bounding circle is computed, the overlap a priori test is
\[
(c_{Ax}-c_{Bx})^2 + (c_{Ay}-c_{By})^2 \leq (r_A + r_B)^2
\]
which can also be processed inside a standard database query.

A third approximation shape is the convex hull, which is the minimal convex polygonal area containing the original geometry. A convex hull provides a better approximation than a rectangle or a circle. There exist algorithms with $O(n \log n)$ complexity to create convex hulls for a given geometry with $n$ vertices, but there are also some drawbacks. First, the time to check geometric properties increases with the number of convex hull vertices whereas the time for rectangles and circles is always constant. Second, the number of convex hull vertices is limited only by the number of vertices of the given geometry and thus may by very high. Hastings presents an approach to split convex hulls into multiple convex hulls (called multihulls) to provide a

![Figure 1. Single and Double Bounding Boxes and construction ideas](image-url)
better approximation (Hastings, 2009). His main idea is to identify the vertex of the original geometry with the largest distance to the convex hull, which is then used to cut the convex hull into two parts. As a result, the union of hulls provides an increasingly accurate approximation of the original geometry, with the drawback that the time to check geometric properties tends to the time for exact checks.

The second purpose of bounding boxes is to have a short representation of a spatial object used, for example, in metadata records. There often exists a simple string representation of this box that can be used, for example, in web pages for URLs or sent by email to identify spatial objects. Some examples: The Dublin Core Metadata Initiative (DCMI) (Kunze & Baker, 2007) is an open organization that defines metadata standards. DCMI provides a box encoding scheme for spatial objects (Cox et al., 2006). The vertical and horizontal spatial extents (which are our bounding boxes) are part of the ISO 19115 Metadata Standard for geographic information (ISO 19115, 2003). The bounding box (called envelope) is part of the Geography Markup Language (GML) (Portele, 2007).

The last significant role of bounding boxes is to represent a geometry inside spatial indexes. A common approach to quickly access geometries is to first approximate the shape of an object by a bounding box that is inserted into a tree structure. Common spatial indexes are Quadtrees (Finkel & Bentley, 1974) or variations of R-Trees (Guttman, 1984). They mainly differ in how a tree is efficiently built and maintained when objects are inserted, changed or removed. A query goes down through the tree until an appropriate tree node is found. The corresponding bounding box then can be used to identify a (hopefully small) set of candidates that must then undergo further geometric checks.

Very often the shapes of spatial objects do not fill axis-aligned rectangles, thus there is a significant loss of shape information. In the following we suggest an approach that approximates arbitrary shapes with the help of two bounding boxes.

3. THE QUICK DBB IDEA

Using two bounding boxes first seems to be a disadvantage. E.g., for the a priori test if two geometries A, B overlap, a single test overlap(SBB_A, SBB_B) is required. Using DBBs we have four checks: overlap(DBB_A1, DBB_B1) OR overlap(DBB_A2, DBB_B1) OR overlap(DBB_A1, DBB_B2) OR overlap(DBB_A2, DBB_B2)

Compared to the exact checks the time for box checks can be neglected as they only require few simple number comparisons. We later show that they significantly reduce the number of candidates for exact checks. Before we present the quick algorithm to create DBBs we briefly discuss the theoretical optimum.

3.1 The Exact Algorithm

Fig. 1a shows a traditional SBB. In the following, north (N), south (S), east (E) and west (W) denote the four main directions. Our goal is to find two axis-aligned bounding boxes that enclose the geometry (fig. 1b). To compute a DBB, it is reasonable to look at the ‘inverse’ areas. So, we do not look at the covered area of a geometry, but rather at the empty area in the corners (e.g. the areas (1) and (2) in fig. 1b). For this, we introduce so-called void rectangles (fig. 1c): a void rectangle is aligned to the DBB’s axes and one corner is an SBB’s corner. In addition, it touches the geometry, but it does not overlap. DBB and void rectangles have aligned borders. For example, the north/west rectangle’s east border is aligned to the north/east void rectangle’s west border. If we respect some additional conditions, we thus can infer DBBs from the void rectangles.

Fig. 1d demonstrates the idea of generating a void rectangle. The north/east corners of all south/west void rectangles reside on a special line string we call the corner profile. To find appropriate void rectangles, we have to check corners that touch their corner profile. Note that even if corner profiles have a finite number of vertices, we get an infinite number of possible void rectangles, as they may touch the profile at an arbitrary position inside a line segment.

To get a complete list of DBB types, we must answer the following question: Which layout of void rectangles allows the completion of the remaining area by exactly two more rectangles (our DBB)? The systematic way to list all cases is to iterate through the possible numbers of void rectangles (fig. 2):

- 1 void rectangle: For any single void rectangle, two more rectangles can always be added to fill the entire space (case A), we have 4 sub-cases: the single void corner can be at NW, NE, SW, SE.
• 2 void rectangles: There exist three cases B, C, and D. Case B has 2 sub-cases: the void rectangles can be at NW/SE or NE/SW. Case C has 4 sub-cases: the common border can be at N, S, E, or W. Case D has 4 sub-cases: the void rectangles can be at NW/SE or NE/SW and they can share the border in direction NS or EW.

• 3 void rectangles: There is no way to generate DBBs from three given void rectangles.

• 4 void rectangles: The only chance to generate DBBs from 4 given void rectangles requires every pair of neighbor void rectangles to have the same height (east-west neighbors) or the same width (north-south neighbors). This is case E.

In summary, there exist 15 cases. The exact algorithm described in (Roth, 2011) iterates through these cases and computes the respective largest void rectangles. The overall maximum leads to the minimal DBB. Note that for cases C, D, and E, the void rectangles are aligned (e.g. have same heights), thus the algorithm must compute the maximum sum area for two or four aligned void rectangles.

The algorithm uses a plane sweep algorithm that sorts all line segments by their x-coordinate of the left vertex. Efficient sorting requires $O(n \cdot \log n)$ steps. It requires additional $O(n)$ steps to ensure monotony since the sorted line list can be iterated element by element again. In summary, it requires $O(n \cdot \log n)$ steps to generate the four corner profiles. To compute the respective maximum void rectangles then can be performed in $O(n)$ steps, thus the overall runtime complexity is $O(n \cdot \log n)$.

### 3.2 The O(n) Approximation

We can significantly reduce the amount of time to compute void rectangles if we introduce an additional constraint: for an SBB with a width to height ratio of $s/t$ we only consider void rectangles with the same ratio of $w/h = s/t$. This means, the void rectangles inner corners reside on the SBB’s diagonals (fig. 3a). We call a double bounding box based on these void rectangles the Quick Double Bounding Box (QDBB).

![Figure 3. Quick approximation idea](image)

Obviously, the void rectangles usually are not maximal. E.g. two of the three void rectangles in fig. 3a) can be extended. We show later that for real geo data this difference to the optimum is acceptable.

The benefit is that we can compute void rectangles very quickly (fig. 3b). E.g., for the SW void rectangle we use the following approach: for each point $(p_x, p_y)$ of a geometry do the following:

- if $(p_x, p_y)$ resides in area (1) and $p_x < w$ assign $w = p_x$, $h = w/t/s$;
- if $(p_x, p_y)$ resides in area (2) and $p_y < h$ assign $h = p_y$, $w = h/s/t$.

For each intersection $X = (X_x, X_y)$ of a geometry’s line segment with the diagonal:
• if $X_s w$, assign $w = X_s$, $h = X_r$.

For the other three void rectangles we perform respective steps. As a great benefit, these steps do not require any sorting of the original geometry points, thus can be performed in $O(n)$ steps.

Fig. 3c) shows the case E of 4 aligned void rectangles. Even though three corners have their own maximum void rectangle, the smallest rectangle (here SW) determines all corners for case E. In this example, the optimal area is not achieved as the two east rectangles are too small. Again, for real data this usually is acceptable (see section 4).

### 3.3 Justification for the Void Rectangle Ratio, Worst Case Considerations

At first view, the width to height ratio of $w/h = s/t$ seems to be an arbitrary choice, as other ratios e.g. $w/h = 1$ could also be considered. The justification for the used ratio is illustrated in fig. 4.

**Figure 4. Justification for the specific void rectangle ratio**

The crucial cases are monotonic line geometries that often represent streets or rivers. For $w/h = s/t$ the two void rectangles share a corner (fig. 4, left). For this scenario the QDBB that is constructed according to case D is very close to the theoretical optimum. If we deviate from this ratio (fig. 4 middle or right), the void rectangles do not construct a DBB as the remaining area cannot be covered by two rectangles, but would require three. As a result, ratios of $w/h \neq s/t$ often produce void rectangles that cannot be represented by any DBB.

Before the quick algorithm is empirically compared to the exact algorithm we discuss the theoretical worst case. We consider case A to illustrate the worst result of the quick DBB algorithm (fig. 5).

**Figure 5. Worst case scenario**

The worst case occurs, if the largest void rectangle nearly covers the SBB’s width, whereas the QDBB’s void rectangle is limited by the diagonal corner. The QDBB’s area then is

$$A_Q = s \cdot t - A_{voidQ} = s \cdot t - h^2 \frac{s}{t} = s \cdot \left( t - \frac{h^2}{t} \right)$$

whereas the exact DBB area is only

$$A_E = s \cdot (t - h).$$

The ratio between the QBB area and the optimum is thus

$$\frac{A_Q}{A_E} = \frac{s \cdot \left( t - \frac{h^2}{t} \right)}{s \cdot (t - h)} = \frac{h + t}{t}.$$

For $h \rightarrow t$ the ratio tends to 2. This means, in worst case the QDBB’s area has double the size of the optimum. As such, this result is disappointing, but
• the QDBB’s size is smaller compared to the SBB’s size $A_S = s \cdot t$; actually $A_S/A_Q$ tends to $\infty$ for $h \to t$;
• such a scenario virtually never occurs in real data as it requires a very specific pattern.
In the following we provide an evaluation of how good the quick approximation is for real data.

4. EVALUATION

We evaluated our approach with the help of the OpenStreetMap database (Ramm & Topf, 2010). We used
the file ‘Germany’ which contains approx. 4 mio. areas and line objects (file of Sept. 2010). The first step was
to measure the creation time for quick DBBs compared to exact DBBs. The DBB algorithms run in Java (SE
1.6) on a XEON L5506 with 2.13 GHz (Windows Server 2008). Fig. 6 shows the results.

![Figure 6. Processing time to create DBBs](image)

As expected, the measurements approve the runtime behavior of $O(n)$ for the quick approximation and
$O(n \cdot \log n)$ for the exact algorithm. Even though the exact algorithm runs in acceptable time, for geometries
with large amounts of vertices (e.g. of city or district boundaries), the runtime benefit of the quick algorithm
is significant. As a result, QDBBs are efficient even for large and fine-grained geometries.

We now show that the approximation does not lead to significant loss of information. For this, we
checked the approach for typical spatial queries. Our test queries checked whether stored geo objects overlap
with randomly generated geometries. We generated five different types of query geometries: lines with four
points, lines with two points, triangles, rotated rectangles and circles. The circle diameters and line, triangle
and rectangles sizes randomly ranged from 0 to 5000m. The center, size and orientation of the geometries
were randomly generated using a uniformly distributed random number. Fig. 7 shows the results.

For each of the five geometry types we produced 100,000 random geometries. Each of them undergoes an
overlap check against all geo objects stored in the database. For each geometry type we plot the average
number of hits against the query geometry size

- using SBBs for query and geo object geometries,
- using QDBBs for query and geo object geometries,
- using exact DBBs for query and geo object geometries, and
- using an exact overlap check that compares the original geometries.

Line segments produce the lowest numbers of exact hits with a considerable amount of SBB and DBB
hits. As expected, line segments usually have a poor box approximation, however (Q)DBBs show
significantly better values than SBBs. For circles the difference between SBB, (Q)DBB and exact hits is
small. This is because an SBB area is only factor $4/\pi=1.27$ and a DBB is only $(4\sqrt{2}-2)/\pi=1.16$ (case E shape)
larger than the corresponding circle area. Thus boxes already provide a good approximation for circles.
To explore the overall benefit of QDBBs compared to SBBs we introduce the false hit ratio

\[
\text{false \_hit \_ratio} = \frac{\text{SBB \_hits} - \text{exact \_hits}}{\text{QDBB \_hits} - \text{exact \_hits}}
\]

that describes the amount of false SBB candidates compared the false QDBB candidates. A false hit ratio of, e.g., 2 means: we get twice as much false SBB candidates compared to QDBBs.

Fig. 7a) presents the false hit ratios for the five geometry types. The improvement ranges from approx. 1.5 to 2.9 (average 2.03). This is a great benefit, as QDBBs significantly reduced the number of false candidates compared to SBBs. We roughly get half as much false candidates.

As an important observation, exact and quick DBBs have very similar hit ratios. In the cases 2-point-lines, rectangles and circles, the plots are virtually identical. Only for 4-point-lines and triangles we can detect a small loss. In our experiments, QDBBs only produce 10.9% more false hits on average than the exact algorithm. As a major result, QDBBs can replace the exact DBBs in real scenarios without significant drawbacks.
5. CONCLUSION AND FUTURE WORK

In this paper we presented an approach to approximate two-dimensional geometries with the help of two bounding rectangles. In addition to the optimal algorithm that runs with complexity of $O(n \log n)$, we presented a quick approximation that requires $O(n)$ steps for $n$ geometry points. For typical geo data the quick algorithm only produces 10.9% more false hits. Compared to traditional single bounding boxes, quick DBB produce only less than half the amount of false hits, thus the QDBB is a good candidate to replace the traditional bounding box as the major tool to approximate geometries for a priori geometric checks.

In the future we want to investigate the benefits of using more than two boxes for approximation with so-called Multiple Bounding Boxes (MBBs). A first approach reuses QDBBs to recursively segment a geometry. A second approach avoids recursion, but requires a quick segmentation algorithm that ideally also runs with $O(n)$ steps.

REFERENCES

HYBRID FUZZY CLASSIFIER DESIGN WITH COEVOLUTIONARY GENETIC ALGORITHM

Roman B. Sergienko, Eugene S. Semenkin and Vladimir V. Bukhtoyarov
Department of System Analysis and Operations Research, Siberian State Aerospace University after academician M.F. Reshetnev - Krasnoyarsk, Russian Federation

ABSTRACT
A new method combination Michigan and Pittsburgh approaches for fuzzy classifier design with evolutionary algorithms is presented. Fuzzy classifier design consists of two stages. At the first stage Michigan method is used for fuzzy rule search with a high grade of certainty. At the second stage Pittsburgh method is applied for search of a rule subset with good performance and a predefined number of the rules. Besides, coevolutionary algorithm for strategy adaptation is used on both stages of fuzzy classifier design. This algorithm automatically solves the problem of genetic algorithm parameters setting. Classification results for machine learning problems from UCI repository and comparison with different alternative classifiers are presented.

KEYWORDS
Fuzzy classifier, Michigan method, Pittsburgh method, coevolutionary algorithm, strategy adaptation.

1. INTRODUCTION
Fuzzy classifier [1] is a classification algorithm based on fuzzy rules extraction from numerical data. Superiority of this method upon other classification algorithms (for example, neural networks) is provided by fuzzy rules which are linguistic expressions and available for human understanding. Thus fuzzy classifier is one of the data mining methods for knowledge discovery.

As a fuzzy classifier is a data mining tool it is more preferable to have minimal number of rules as a base and to have minimal number of conditions in the antecedent part of rules. In [2] fuzzy rule extraction has been formulated as a combinatorial optimization problem with three objectives: to maximize the performance of an extracted rule set, to minimize a number of extracted rules, and to minimize the total length of extracted rules.


A new method of Michigan and Pittsburgh approaches combination for fuzzy classifier rule base design with evolutionary algorithms is presented in this paper. Fuzzy classifier rule base design consists of two stages. At the first stage Michigan method is used for fuzzy rules search with high grade of certainty. At the second stage Pittsburgh method is applied for subset of the rules searching with good performance and given number of rules. The constraint for number of rules is used at the second stage of fuzzy classifier generating. This method requires less computational power than multiobjective optimization for fuzzy rules extraction.

Besides coevolutionary genetic algorithm for strategy adaptation [7] is used for both stages of fuzzy classifier rule base design. This algorithm solves complicated problem of genetic algorithm parameters setting. Coevolutionary algorithm for unconstrained optimization is applied at the first stage (Michigan
approach) and coevolutionary algorithm for constrained optimization is used at the second stage (Pittsburgh approach).

This method for fuzzy classifier rule base design has been applied for some machine learning problems from UCI repository [8]. Results of classification (performance values and fuzzy rule bases) are demonstrated. Classification performance values were compared with results of alternative classifiers.

The development of the new Michigan and Pittsburgh methods combination for fuzzy classifier rule base design is introduced in Section 2. The details of the coevolutionary genetic algorithm for strategy adaptation are described in Section 3. The machine learning problems from UCI repository, the results of numerical experiments and results of comparison with alternative classifiers are presented in Section 4. Summary is given in Section 5.

2. MICHIGAN AND PITTSBURGH METHOD COMBINATION FOR FUZZY CLASSIFIER RULE BASE DESIGN

The Michigan and Pittsburgh methods combination for fuzzy classifier generating implicates sequential using of the first and second approaches. But a priori to this there are two important preparatory steps of fuzzy classifier generating: attribute fuzzification (fuzzy number semantic setting) and initial population of fuzzy rules forming for Michigan approach with using a priori information from a learning sample. So let’s consider every preparatory step and the main stage of fuzzy classifier generating in more details.

Attribute fuzzification. In this work for each attribute of a machine learning problem five triangular fuzzy numbers and a term “ignore” are determined:

1. “very small”
2. “small”
3. “average”
4. “large”
5. “very large”
6. “ignore”.

If an attribute has both negative and positive values it is better to use words “negative”, “positive” and “null” instead “small”, “average” and “large” accordingly.

A triangular fuzzy number is characterized by three parameters: left boundary \( a \), center \( b \), and right boundary \( c \). For uniform filling of attribute variability interval \([A; B]\) by fuzzy numbers these parameters are defined by the following equations:

\[
\begin{align*}
    a_i &= A + \frac{1}{6} \cdot (B - A) \cdot (i - 1), i = 2..5, \\
    b_i &= A + \frac{1}{6} + \frac{1}{6} \cdot (B - A) \cdot (i - 1), i = 1..5, \\
    c_i &= A + \frac{1}{3} + \frac{1}{6} \cdot (B - A) \cdot (i - 1), i = 1..4.
\end{align*}
\]

The maximum \( B \) and minimum \( A \) values of the attribute are determined with a learning sample.

Initial population of fuzzy rules forming. This step is very important because random generating of fuzzy rules is unacceptable. Classification problems can have a lot of attributes. In this case probability of random fuzzy rule generating that would have at least one corresponding element from a learning sample is very low. It is considerable even for classification problem dimension and it is equal to four or more. So using a priori information from a learning sample is necessary for initial population of fuzzy rules forming. This procedure is considered in the following.

Let \( n \) be a number of rules at Michigan-style stage of fuzzy classifier generating, \( k \) be a number of the class, \( l \) be a number of attributes. Steps of initial population of fuzzy rules forming:

1) Put \( m = \text{int}(nk) \).
2) Sort a learning sample by number of the class and determine boundaries for each class in the sorted array.
3) For \( i = 1 \) to \( k \)
   For \( j = 1 \) to \( m \)
      Perform random selection of the element for the class \( i \) from a sorted learning sample.
For $t := 1$ to $l$

Determine the nearest center of a fuzzy number for attribute $t$.
Put the corresponding fuzzy number as element of the generated fuzzy rule.
Exchange the fuzzy number to a term “ignore” with probability equals to 0.33.

4) Fill population for Michigan-style stage by generated fuzzy rules.

**Michigan-style stage.** The chromosomes are the fuzzy rules. Chromosome length is equal to the number of attributes, each gene is a figure for the corresponding fuzzy number (1..6). Fitness function is grade certainty of the fuzzy rule calculated by a learning sample [1]. Genetic algorithm for unconstrained optimization is applied. New population forming method is modified. After genetic generation performing parents and child are combined to common array, $m$ different fuzzy rules with the best values of fitness function for each class are selected to the next generation. This new population forming method provides diversity of rules for each class and diversity of classes in population. For each generation classification performance is calculated for population at whole. Population with the best value of classification performance is used for the next stage of fuzzy classifier generating.

**Pittsburgh-style stage.** The chromosomes are the fuzzy rule sets. Chromosome length is equal to the population size for Michigan-style stage. Chromosome genes are binary. Value “1” means using of the corresponding rule in the set, value “0” means not using the corresponding rule. Fitness function is classification performance. Constraint for number of rules is used. This value is specified by a researcher. The constraint is used because it is better to have small number of rules in the final rule base. Genetic algorithm for constrained optimization is applied. New generation forming method is standard.

Thus this method of Michigan and Pittsburgh approach combination for fuzzy classifier rule base generating provides simultaneous advantages of both methods. At Michigan-style stage we get different rules with high values of grade certainty for different classes and at Pittsburgh-style stage we get the rule set with the maximum value of classification performance and necessary number of the rules. As the used rules are fixed and they cannot be changed Pittsburgh-style stage doesn’t require a lot of computational resources. Special method of initial population generating for the Michigan-style stage provides using of a priori information from learning sample.

3. COEVOLUTIONARY GENETIC ALGORITHM FOR STRATEGY ADAPTATION


One of the most complicated problems for application of genetic algorithm is the algorithm parameters setting. Conventional genetic algorithm has at least three methods of selection (proportional, tournament, and rank), three methods of recombination (one-point, two-point, and uniform). Mutation probability requires tuning too. For constrained optimization problems it is necessary to choose a constraint handling method. Some of various combinations can be estimated at tens. Exhaustive search of combinations requires a lot of time and computational power. Parameters combination selection by chance is also a bad idea as algorithm efficiency the same problem can differ very much for various parameters setting.

In [9] strategy adaptation by competing subpopulations has been suggested. Each subpopulation has its own search strategy (algorithm’s parameters combination). Resource redistribution provides domination of the subpopulation with the best for problem-in-hand search strategy. This method can be considered as an example of coevolutionary genetic algorithm.

We develop slightly different approach [7] that uses both competition and cooperation of individual genetic algorithms with its own parameters setting. Cooperation of individual conventional genetic algorithms is provided by migration of the best solutions to all of the individual genetic algorithms. So, coevolutionary algorithm efficiency can increase because of positive effect of subpopulations interaction.

This coevolutionary genetic algorithm needs no tuning of special parameters.

Let’s give some definitions.

*Common (shared) resource* is a number of individuals used by coevolutionary algorithm in current generation (population size). It’s constant.

*Algorithm’s resource* is a number of individuals used by individual genetic algorithm in current
generation (subpopulation size). It’s variable. Algorithm’s resource value is changed by resource redistribution operator. Initial resources of all individual algorithms are equal.

Adaptation interval is a number of generations for independent processing individual genetic algorithm. After each adaptation interval, resource redistribution and migration operators are performed. Adaptation interval is constant.

Algorithm’s fitness is individual genetic algorithm efficiency criterion. It’s necessary to calculate algorithm’s fitness for resource redistribution operator performing. Algorithms’ fitness values are calculated after each adaptation interval. Let \( T \) be adaptation interval value, \( k \) be counter of generations in adaptation interval (\( k=0 \) for last generation, \( k=1 \) for previous one, \( \ldots \), \( k=T-1 \) for the first generation in an adaptation interval), \( i \) be counter of individual genetic algorithms (subpopulations), \( b_i(k) \) be binary variable. \( b_i(k)=1 \) if the subpopulation \( i \) contains the individual with the global best fitness value, \( b_i(k)=0 \) otherwise. Algorithm’s fitness \( q_i \) is calculated by the following equation:

\[
q_i = \sum_{k=0}^{T-1} \frac{T-k}{k+1} b_i(k)
\]

Penalty size determines population size part of unsuccessful individual genetic algorithms (subpopulations) which is passed to the winner algorithms by resource redistribution operator after each adaptation interval. Penalty size is constant.

“Social card” size is minimal value of individual algorithm’s population size. Individual algorithm’s population size cannot be less than this value. “Social card” is constant.

Resource redistribution is an operator that provides competition in the coevolutionary genetic algorithm and strategy adaptation. Algorithm’s fitness is calculated for each subpopulation for each adaptation interval. In standard resource redistribution, one individual algorithm with the best fitness value (or a few algorithms with the identical best fitness value) gets the resource (population size part) from other algorithms according to penalty size. Applying tournament method resource redistribution is performed for each couple of algorithms. Tournament resource redistribution efficiency unessentially differs from standard operator efficiency.

Migration is an operator that provides cooperation among individual conventional genetic algorithms. After each adaptation interval all subpopulations are consolidated into the array. This array is sorted using individual’s fitness function. Then each individual algorithm (subpopulation) gets necessary amount of the best individuals from this array. We always start from the best chromosome to build up the size for each subpopulation, i.e. there is an overlap between subpopulations.

The experimental investigation [8] has determined reasonable values of adaptation interval, penalty size and “social card” size, the same for all the test problems. Adaptation interval should be equal to 5 generations. Penalty size and “social card” size are equal to approximately 10 percents. Significant deviations from these values have negative effect for all the test problems according to [8]. So these parameters don’t need tuning.

The pseudo code of the coevolutionary genetic algorithm for unconstrained problems is presented below.

1) Choose \( N \) individual genetic algorithm with different search strategies (type of selection, recombination, and mutation).
2) Define common resource \( M \), number of generations \( G \), adaptation interval \( T \), penalty size \( P \), “social card” size \( S \).
3) Initialize populations of individual genetic algorithms.
4) Put \( j=0 \).
5) For \( k=T-1 \) to 0 perform all individual genetic algorithms.
6) \( j:=j+T \).
7) For \( i=1 \) to \( N \) calculate algorithm’s fitness \( F_i \)
8) Perform resource redistribution operator.
9) Perform migration operator.
10) If \( (j\neq G) \) go to 5.
11) Determine final solution of optimization problem.

Coevolutionary genetic algorithm for unconstrained optimization has been tested on typical set of GA-community unconstrained optimization test problems in [10]. The reliability of optimum point catching has been used as the efficiency criterion. The common result of those investigations was the observation that coevolutionary algorithm competing with some dozens individual algorithms was mostly the second or third
best one among them and always more effective than individual genetic algorithm with average effectiveness. It seems to be close to success as coevolutionary algorithm has demonstrated its ability to provide effective problem-in-hand solving procedure without extra efforts for setting GA parameters.

The main idea of coevolutionary algorithm adaptation for constrained optimization uses different methods of constraint handling (“death” penalty, dynamic or adaptive penalty function [11]) in search strategies of individual genetic algorithm. Migration method was modified for algorithm adaption for constrained optimization [12].

Coevolutionary genetic algorithm has equal or better reliability than the best conventional genetic algorithm. Also coevolutionary algorithm has better convergence rate than the best conventional genetic algorithm. These effects are provided by competitive cooperation between subpopulations in coevolutionary algorithm. Besides coevolutionary genetic algorithm is a very appropriate tool for parallel computing with multiprocessors.

Coevolutionary genetic algorithm for unconstrained optimization is applied at the Michigan-style stage of fuzzy classifier generating and coevolutionary genetic algorithm for constrained optimization is used at the Pittsburgh-style stage. So this provides the solution of genetic algorithm parameters setting problem.

4. TEST PROBLEMS AND NUMERICAL EXPERIMENTS

The developed method of fuzzy classifier rule base design has been applied for a number of classification machine learning problems from UCI repository [8]:
- Credit (Australia-1);
- Credit (Germany);
- Liver Disorder;
- Iris;
- Yeast;
- Glass Identification;
- Landsat Images.

There are 36 attributes for Landsat Images problem: four spectral values (red, green, and two infra-red) for each pixel in 3*3 groups. In this work mean spectral values for a pixel group are used. Acceptability of this operation was proved by method of principal components. So we have only four attributes for this problem.

For each problem classification performance values and other parameters are presented. For the first three problems comparison with alternative classification methods has been performed. These algorithms are Bayesian approach, multilayer perceptron [13], boosting [14], bagging [15], random subspace method (RSM) [16], and cooperative coevolution ensemble learning (CCEL) [17]. Performance value is the part of test sample that is classified correctly by an algorithm.

The classification performance comparison with alternative algorithms for Credit (Australia-1), Credit (Germany), and Liver Disorder problems is presented in Table 1. The results of the suggested method performance are demonstrated in Table 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Credit (Australia-1)</th>
<th>Credit (Germany)</th>
<th>Liver Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy classifier</td>
<td>0.891</td>
<td>0.794</td>
<td>0.725</td>
</tr>
<tr>
<td>Bayesian approach</td>
<td>0.847</td>
<td>0.679</td>
<td>0.629</td>
</tr>
<tr>
<td>Multilayer perception</td>
<td>0.833</td>
<td>0.716</td>
<td>0.693</td>
</tr>
<tr>
<td>Boosting</td>
<td>0.760</td>
<td>0.700</td>
<td>0.656</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.847</td>
<td>0.684</td>
<td>0.630</td>
</tr>
<tr>
<td>RSM</td>
<td>0.852</td>
<td>0.677</td>
<td>0.632</td>
</tr>
<tr>
<td>CCEL</td>
<td>0.866</td>
<td>0.746</td>
<td>0.644</td>
</tr>
</tbody>
</table>
Table 2. Michigan and Pittsburgh method combination for fuzzy classifier design with coevolutionary algorithm

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Problem 1</th>
<th>Problem 2</th>
<th>Problem 3</th>
<th>Problem 4</th>
<th>Problem 5</th>
<th>Problem 6</th>
<th>Problem 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial number of rules</td>
<td>200</td>
<td>200</td>
<td>100</td>
<td>50</td>
<td>200</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Average classification efficiency with initial rules</td>
<td>0.854</td>
<td>0.782</td>
<td>0.595</td>
<td>0.945</td>
<td>0.417</td>
<td>0.640</td>
<td>0.793</td>
</tr>
<tr>
<td>Average classification efficiency after Michigan stage</td>
<td>0.870</td>
<td>0.791</td>
<td>0.653</td>
<td>0.979</td>
<td>0.495</td>
<td>0.687</td>
<td>0.812</td>
</tr>
<tr>
<td>Average classification efficiency after Pittsburgh stage (feasible number of rules in brackets)</td>
<td>0.827(10)</td>
<td>0.762(50)</td>
<td>0.666(10)</td>
<td>0.908(3)</td>
<td>0.573(20)</td>
<td>0.737(20)</td>
<td>0.838(10)</td>
</tr>
<tr>
<td></td>
<td>0.861(20)</td>
<td>0.790(80)</td>
<td>0.682(15)</td>
<td>0.951(4)</td>
<td>0.586(30)</td>
<td>0.781(30)</td>
<td>0.847(15)</td>
</tr>
<tr>
<td></td>
<td>0.873(30)</td>
<td>0.692(30)</td>
<td>0.971(5)</td>
<td>0.975(6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum classification efficiency after Pittsburgh stage (feasible number of rules in brackets)</td>
<td>0.870(10)</td>
<td>0.767(50)</td>
<td>0.687(10)</td>
<td>0.947(3)</td>
<td>0.598(20)</td>
<td>0.757(20)</td>
<td>0.849(10)</td>
</tr>
<tr>
<td></td>
<td>0.890(20)</td>
<td>0.710(15)</td>
<td>0.973(4)</td>
<td>0.606(30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.891(30)</td>
<td>0.725(20)</td>
<td>0.987(5)</td>
<td>0.626(60)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We can see in Table 1 that Michigan and Pittsburgh method combination for fuzzy classifier design with coevolutionary algorithm for strategy adaptation provides higher classification efficiency than other classification algorithms.

Using the Table 2 we can see features of method processing. Finally we get a compact fuzzy rule set with high efficiency.

In the Table 3 we can see fuzzy rule base for LandSat Images problem. The database consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. This database was generated from Landsat Multi-Spectral Scanner image data. One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the near infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The aim is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number. The number is a code for the following classes:

1 - red soil;
2 - cotton crop;
3 - grey soil;
4 - damp grey soil;
5 - soil with vegetation stubble;
6 - very damp grey soil.

In our problem we use five fuzzy terms for each attribute:
1 – “very weak”;  
2 – “average weak”;  
3 – “average”;  
4 – “average intensive”  
5 – “very intensive”.

Table 3. The fuzzy rule set for landsat images problem

<table>
<thead>
<tr>
<th>Green Spectrum</th>
<th>Red Spectrum</th>
<th>1st Infra-red Spectrum</th>
<th>2nd Infra-red Spectrum</th>
<th>Class</th>
<th>Grade of Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0.60</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0.96</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0.76</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>0.68</td>
</tr>
</tbody>
</table>
The fuzzy rule set consists of ten rules. Classification accuracy for train sample is 0.8490. Classification accuracy for test sample is 0.8435. So we can conclude that fuzzy classifier has good generalizing ability.

5. CONCLUSION

Thus new method of Michigan and Pittsburgh approach combination for fuzzy classifier rule base design has developed and investigated on some classification problem from UCI repository. This method has high operation speed and efficiency because advantages of both approaches are used. Michigan-style stage provides fast search of fuzzy rules with the best grade of certainty values for different classes. Pittsburgh-style stage provides thes rules subset search with good performance and predefined number of the rules and doesn’t require a lot of computational power because the rules are not changed and binary individuals are used. Besides constraint optimization is not as complicated as multiobjective optimization for computation. We can see that huge populations are not used for both stages of fuzzy classifier generating.

Very important preparatory step is initial population forming for Michigan-style stage. In Table 2 classification performance values for initial populations don’t differ very much from performance values after Michigan-style stage performing. Thereby the introduced method of initial population forming allows to use a priori information from a learning sample at most. Michigan-style stage is necessary for light increment of fuzzy rule set efficiency and smoothing of randomness at initial population forming.

In Table 2 we can see that classification performance values after Pittsburgh-style stage performing are better than performance values after Michigan-style stage although number of rules are reduced. This means that a large rule set is not always better by performance. “Bad” rules have damaged the effect.

Coevolutionary genetic algorithm for strategy adaptation using at both stages of fuzzy classifier design allows to refuse the genetic algorithm parameters setting without negative effect for algorithm efficiency. Coevolutionary genetic algorithm for unconstrained optimization is applied at Michigan-style stage of fuzzy classifier design and coevolutionary genetic algorithm for constrained optimization is used at the Pittsburgh-style stage of fuzzy classifier design.

Fuzzy classifier comparison with alternative classification methods by performance value (see Table 1) demonstrates that fuzzy classifier has either the same efficiency as present-day classification algorithms or even more. However a fuzzy classifier is not only classification method but data mining algorithm for knowledge discovery as well. It is the main advantage of a fuzzy classifier. The introduced fuzzy rule extraction method provides a fuzzy rule base with a small number of the rules. Besides the rules have often the term “ignore” for some attributes that makes them short. These rules are available for human understanding and they can be very useful for experts and specialists from corresponding problem domain (medicine for Liver Disorder problem, finances for credit problems, biology for Iris and Yeast problems, pedology and earth science for Landsat Images problem, and material science for Glass Identification problem).

REFERENCES


A QoS-AWARE APPROACH FOR WEB SERVICE SELECTION BASED ON PROBABILITY EVALUATION

Dessislava Petrova-Antonova
Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski” - 125, Tzarigradsko shosse Blvd., 1113 Sofia, Bulgaria

ABSTRACT
The rising number of available web services with the same functionality hampers the web service discovery process and complicates clients while searching for a web service that best meets their requirements. That is why the web service description and discovery need to concern not only functional but also non-functional (QoS) properties of web services. This paper contributes to that challenge by proposing a QoS-aware approach for web service selection. The QoS data collected during web service executions is used for prediction of future values of QoS properties, based on probability evaluation. As a result, the web service with the largest probability to satisfy the requirements of the client is identified and recommended to him/her as the best one.

KEYWORDS
Probability, QoS, selection, web service.

1. INTRODUCTION
Service Oriented Architecture (SOA) provides a new generation of software architectures that attracts attention as a promising way for smooth integration of loose coupled software applications. The web services provide a popular technology that overcomes the limitations of traditional middleware and can be used to implement the SOA model. The current web service architecture is based on Web Service Description Language (WSDL) and Universal Description Discovery and Integration (UDDI) standards that support functional web service description, publication and discovery. Due to the rising number of web services that share the same functionality, often clients are complicated while searching for a web service that best meets their requirements. That is why many research efforts focus on non-functional, also called Quality of Service (QoS), properties of web services.

The QoS can be applied over the whole process of web service description, publication and discovery. In this paper, it is utilized to differentiate a single web service, which best meets client requirements among multiple web services with the same functional properties. The QoS data collected during web service executions is used for prediction of the future values of QoS properties, based on probability evaluation. Thus, the client will be able to select a web service with the largest probability of having QoS properties that are as close as possible to the preliminary defined requirements.

The remainder of this paper is structured as follows. In Section 2, some background information related to QoS properties and discrete random variables is presented. In Section 3, a QoS-aware approach for web service selection is described in details. In Section 4, an illustrative example of the proposed approach is given. Section 5 introduces related work. Finally, Section 6 concludes the paper.

2. BACKGROUND
This section presents a preliminary knowledge, including QoS specification and some aspects of the theory of discrete random variables. Section 2.1 provides several classifications of QoS properties, while in Section 2.2, the probability functions applied in the proposed approach for web service selection are shown.
2.1 QoS Properties

QoS properties present the non-functional characteristics of web services. They have significant role in web service discovery process, helping clients to select a web service that meets best their requirements.

Various classifications of QoS properties exist in literature. The QoS properties can be distinguished on the basis of their values. If the client requires the values of QoS properties to be as higher as possible, then these QoS properties can be classified as increasing or positive. Otherwise, they are classified as decreasing or negative. Examples for positive QoS properties are Availability and Throughput. In contrast, Response time and Cost are negative QoS properties. Such classifications are given by Tran et al (2009) and D’Mello et al (2008).

As not all of the QoS properties can be quantified, Tran et al (2009) classify the QoS properties as unit and unitless. For example, QoS properties, such as Supported standards, Authentication method and Data encryption are unitless, because they do not have numeric values. Similarly, Kritikos and Plexousakis (2009) distinguish between measurable and unmeasurable QoS properties. They propose a categorization of QoS properties dividing them in two sets, namely domain-independent and domain-dependant. The domain-independent QoS properties refer to the technical aspects of web services without respect to application domain. They are organized in seven categories, as follows: Performance, Dependability, Integrity, Security, Network and infrastructure related, Cost, and Other. The last category is created to include the QoS properties that are not covered by any other category. The domain-dependant QoS properties rely on the application domain of web services. For example, the domain-dependant QoS properties for the application domain of Vehicle Traffic Monitoring are covered area, routes set, detail level, accuracy, completeness, validity, timeliness, and coverage. They characterized the data processed and produced by the monitoring system.

2.2 Discrete Random Variables

The probability theory finds applications in every area of scholarly activities as well as in daily experience like weather and stock market prediction. In any experiment there are a finite number of possible outcomes. For example, the values of QoS properties recorded in some history log at a given time can be seen as such possible outcomes. Each outcome can be associated with a number according to the specific rule of association called random variable (Devore 2008). Normally, random variables are denoted with uppercase letters and their values are denoted with lowercase letters. Each random variable can be discrete or continuous. The possible values of discrete random variable form a finite set or infinite sequence having a first element, a second element, and so on. The possible values of continuous random variables consist of an entire interval on the number line (Devore 2008). In our sample a given QoS property such as Response time can be represented with a discrete random variable \( X \) with several possible values \( x \) that correspond to the recorded values in the history log.

The random variable can be described by a probability mass function (PMF), which captures the probabilities of the values that the random variable can take. If \( x \) is any possible value of discrete random variable \( X \), the PMF of \( x \), denoted \( p_x(x) \) is the probability of the event \( \{X = x\} \) consisting of all outcomes that give rise to a value of \( X \) equal to \( x \) (Bertsekas and Tsitsiklis 2000):

\[
p_x = P(X = x)
\]

For example, let the experiment consist of six independent measurements of the Response time of a particular web service and let \( X \) be the value measured. Let the event \( \{X=1.8\} \) is happened one time, the event \( \{X=2\} \) is happened four times, and the event \( \{X=3\} \) is happened one time. Then the PMF of \( x \) is:

\[
p_x(x) = \begin{cases} 
1 & \text{if } x = 1.8 \\
2 & \text{if } x = 2 \\
3 & \text{if } x = 3 \\
6 & \text{otherwise}
\end{cases}
\]
Sometimes the information that the PMF gives for a given random variable $X$ is needed to be summarized in a single representative number. This is accomplished by an expected value of $X$, denoted $E(X)$ or $\mu_X$:

$$E(X) = \mu_X = \sum x p_X(x)$$

(2)

For the example above, the expected value of the Response time is calculated as follows:

$$E(X) = \mu_X = 1.8 \cdot \frac{1}{6} + 2 \cdot \frac{2}{3} + 3 \cdot \frac{1}{6} = 0.3 + 1.33 + 0.5 = 2.13$$

For a given value $x$ it is often necessary to compute the probability that the value of the random variable $X$ will be at most $x$. This can be accomplished by a cumulative distribution function (CDF), denoted $F_X$. The CDF of a discrete random variable provides the probability $P(X \leq x)$ (Bertsekas and Tsitsiklis 2000):

$$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k)$$

(3)

Since the two events $\{X \leq x\}$ and $\{X > x\}$ are mutually exclusive, the sum of their probabilities is equal to 1. Therefore:

$$P(X > x) = 1 - F_X(x)$$

(4)

For the example above, the probability that the Response time will be at most 2 sec is calculated as follows:

$$F_X(x) = P(X \leq 2) = \frac{1}{6} + \frac{4}{6} = \frac{5}{6}$$

The probabilistic models usually concern several random variables. For example, in the context of web services, the QoS requirements of the clients consist of definitions for several QoS properties that can be seen as random variables belonging to the same experiment. The joint CDF of two random variables is as follows:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = F_X(x)F_Y(y)$$

(5)

For example, if the probability that the Response time is at most 2 sec is 5/6 and the probability that the Throughput is at most 160 kbps is 2/5, then the joint CDF of these two QoS properties is calculated as follows:

$$F_{X,Y}(x,y) = P(X \leq 2, Y > 160) = P(X \leq 2) \cdot (1 - P(Y \leq 160)) = \frac{5}{6} \cdot \frac{3}{5} = \frac{1}{2}$$

The equation 5 can be used for calculation of probability that the QoS properties of a particular web services will satisfy the client’s requirements. Such calculations can be accomplished for all candidate web services that meet functional requirements of the client. After that the client will be able to select a web service with a maximum probability of having QoS properties that are as close as possible to the preliminary defined requirements. The differentiation among web services with the same probability to meet client’s requirements can be accomplished by comparison of their expected values of QoS properties calculated according to equation 2.

3. A QoS-AWARE APPROACH FOR WEB SERVICE SELECTION

This section presents a QoS-aware approach for web service selection based on probability evaluation. The approach consists of eight steps that are described below in detail.

Let the candidate web services, namely those that satisfy functional requirements of the client, form the set $S$:

$$S = \{S_1, S_2, \ldots, S_{\ell}, \ldots, S_n\}, \ell = 1 + n$$

where $n$ is the number of candidate web services.

Let the QoS properties, whose values are stored in history log, form the set $Q$:
where \( l \) is the number of QoS properties.

**Step 1 Data extraction from history log.**

As mentioned in Section 2.2 each QoS property of web services can be represented with a discrete random variable with several possible values that correspond to the recorded values in the history log. Thus, the values of QoS properties of a particular web service form the following matrix:

\[
V_{si} = \begin{bmatrix}
q_1 & q_2 & \cdots & q_l \\
x_{11} & x_{12} & \cdots & x_{1l} \\
x_{21} & x_{22} & \cdots & x_{2l} \\
\vdots & \vdots & \ddots & \vdots \\
x_{kl} & x_{k2} & \cdots & x_{kl}
\end{bmatrix}, \quad l = 1 + n
\]

where \( x_{ij} \) is the \( k \)-th value of QoS property \( q_i \) obtained from history log and \( k \) is the number of values obtained.

**Step 2 Extraction of events and priorities from client’s requirements.**

The QoS requirements of the client are considered as events. They form the following vector:

\[
R = [r_1 \ r_2 \ \cdots \ r_l]
\]

where \( r_i = [q_i \leq y_i] \) or \( r_i = [q_i > y_i] \).

The client also specifies a priority of each QoS property from his/her perspective. The vector \( W \) presents these priorities and is defined as follows:

\[
W = [w_1 \ w_2 \ \cdots \ w_l]
\]

where \( w_l \) is the priority of QoS property \( q_l \). The priorities are integer values from \( l \) to the number \( n \) of QoS properties, so that \( l \) is the maximum priority and \( n \) is the minimum priority.

**Step 3 Calculation of PMF for each value of each QoS property for all candidate web service.**

The probabilities of the values that the QoS property takes are calculated according to equation 1. They are presented with the following vector:

\[
P_{q_{x_{si}}} = [p(x_1) \ p(x_2) \ \cdots \ p(x_m)], \quad j = 1 + l
\]

where \( p(x_m) \) is the probability that the value of QoS property \( q_i \) will be \( x_m \). Note that \( m \) is the number of unique values obtained from history log for a given QoS property.

**Step 4 Calculation of CDF for each QoS property for all candidate web service.**

The CDF of each QoS property is calculated according to equation 3. In case of \( r_i = [q_i \leq y_i] \) the CDF of QoS property \( q_i \) is as follows:

\[
F_{q_i}(y_i) = P(q_i \leq y_i)
\]

In case of \( r_i = [q_i > y_i] \) the CDF of QoS property \( q_i \) is as follows:

\[
F_{q_i}(y_i) = 1 - P(q_i \leq y_i)
\]

The CDFs calculated for each QoS property of all candidate web services form the following matrix:

\[
F = \begin{bmatrix}
F_{11} & F_{12} & \cdots & F_{1l} \\
F_{21} & F_{22} & \cdots & F_{2l} \\
\vdots & \vdots & \ddots & \vdots \\
F_{n1} & F_{n2} & \cdots & F_{nl}
\end{bmatrix}
\]

**Step 5 Calculation of joint CDF for all candidate web services.**

Finally, for each web service the joint CDF is calculated according to equation 5. The result is a vector \( F \) of probabilities:

\[
F = [F_{s_1} \ F_{s_2} \ \cdots \ F_{s_n}]
\]

where \( F_{s_n} \) is the probability that web service \( S_n \) will satisfy the client’s requirements.
Step 6 Finding of web service with the maximum probability to satisfy client’s requirements.
The web service with the maximum probability to satisfy the client’s requirements is then proposed to the client. Here, several well known algorithms for finding largest element in an array are applicable.

Step 7 Calculation of expected values of each QoS property for all candidate web services.
Suppose that there are several web services with the same probability to satisfy client’s requirements. They form the set $S'$, which is the subset of $S$:

$$S' = \{S_1, S_2, ..., S_{t}, ..., S_t\}, t = 1 + t$$

where $t$ is the number of web services that have probability to satisfy client’s requirements.

The expected value for each QoS property of those web services is calculated according to equation 2. These expected values form the following matrix:

$$E = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1t} \\ E_{21} & E_{22} & \cdots & E_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ E_{n1} & E_{n2} & \cdots & E_{nt} \end{bmatrix}, t = 1 + t$$

where $E_{ni}$ is the expected value of QoS property $q_i$ of web service $S_n$.

Step 8 Finding of web service with the best expected values of QoS properties.
The rows of the matrix $E$ are sorted according to the values of vector $F$, so that the web services with the maximum probability to satisfy the client’s requirements are shifted at the top of the matrix $E$. Next the columns of the matrix $E$ are sorted according to the values of vector $W$. Practically, each value of the vector $W$ shows the target position of a particular column in $E$ during sorting. Then the rows of the matrix $E$ are sorted according to the values of the first column. For any rows that have equal elements in a particular column sorting is based on the column immediately to the right. The sorting is in descending or ascending order depending on definition of the requirement for the QoS properties identifying the corresponding column. In case of $r_l = \{q_l \leq y_l\}$ the sorting is in ascending order. Otherwise, in case of $r_l = \{q_l \geq y_l\}$ the sorting is in descending order. As a result, the web service with the best expected values of QoS properties is shifted on the top of the matrix $E$ and is proposed to the client.

The pseudo code of the algorithm that implements the proposed approach is presented on Figure 1.

```
1 FOR EACH web service in S DO
2     Extract data from history log
3     Create matrix V
4 END FOR
5 FOR EACH QoS property DO
6     Extract an event from client’s requirement
7     Extract a priority from client’s requirement
8 END FOR
9 FOR EACH web service in S DO
10    FOR EACH QoS properties in P DO
11        FOR EACH unique value of QoS property DO
12            Calculate PMF
13        END FOR
14    END FOR
15    Calculate CDF
16 END FOR
17 Calculate joint CDF
18 Find web service with maximum joint CDF
19 IF there are web services with the same CDF THEN
20    FOR EACH web service in S' DO
21        FOR EACH QoS properties in P DO
22            Calculate expected value
23        END FOR
24    END FOR
25 END IF
26 Sort the matrix E
27 Extract the web service with the best expected values
```

Figure 1. A pseudo code of the algorithm for QoS-aware web service selection.
4. ILLUSTRATIVE EXAMPLE

In this section, an example that demonstrates the proposed approach is presented. Three web services, named $S_1$, $S_2$ and $S_3$, are considered. The QoS properties that will be used in the example are Successful execution rate ($q_1$), Reputation ($q_2$), Price ($q_3$), Availability ($q_4$) and Execution time ($q_5$).

The values of QoS properties of the web services $S_1$, $S_2$ and $S_3$ are obtained from history log used by Qi et al (2010) to demonstrate a method for QoS-aware web service selection based on credibility evaluation.

According to Step 1 of the proposed approach, the values of QoS properties of each web service form the following matrices:

$$V_{S_1} = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.90 & 0.10 & 0.80 & 5.00 \\ 0.95 & 0.05 & 0.95 & 4.10 \\ 0.80 & 0.12 & 0.90 & 4.50 \\ 0.75 & 0.10 & 0.80 & 4.20 \\ 0.90 & 0.12 & 0.85 & 4.20 \\ 1.00 & 0.12 & 0.90 & 4.00 \\ 0.95 & 0.15 & 0.95 & 4.00 \\ 0.95 & 0.10 & 0.80 & 4.60 \end{bmatrix}$$

$$V_{S_2} = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.85 & 0.13 & 0.85 & 4.70 \\ 0.90 & 0.15 & 0.90 & 4.40 \\ 0.95 & 0.18 & 0.95 & 4.20 \\ 0.75 & 0.13 & 0.95 & 4.00 \\ 0.80 & 0.18 & 0.95 & 4.00 \\ 0.95 & 0.15 & 0.85 & 4.40 \\ 0.80 & 0.13 & 0.80 & 5.10 \\ 1.00 & 0.15 & 0.90 & 10.00 \end{bmatrix}$$

$$V_{S_3} = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.90 & 0.13 & 0.85 & 4.20 \\ 0.80 & 0.16 & 0.95 & 4.00 \\ 0.80 & 0.15 & 0.90 & 5.00 \\ 0.75 & 0.13 & 0.85 & 4.10 \\ 0.60 & 0.15 & 0.90 & 10.00 \\ 1.00 & 0.11 & 0.80 & 10.00 \end{bmatrix}$$

On the Step 2 the client’s requirements are extracted. The events that are obtained from the requirements form the following vector:

$$R = \{ q_1 = 0 \} \quad \{ q_2 > 0.8 \} \quad \{ q_3 < 14 \} \quad \{ q_4 > 0.85 \} \quad \{ q_5 < 5 \}$$

The priorities of QoS properties are presented as follows:

$$W = [1 \ 3 \ 4 \ 5 \ 2]$$

According to Step 3, the probabilities of the values that the QoS property of each web service takes are calculated. They are used to compute the CDF of each QoS property of all candidate web services on the Step 4. As a result, the following matrix is defined:

$$F_q = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.875 & 0.750 & 0.750 & 0.625 & 0.625 \\ 0.875 & 0.750 & 0.375 & 0.875 & 0.625 \\ 0.750 & 0.750 & 0.625 & 0.750 & 0.500 \end{bmatrix}$$

On the step 5 the joint CDF is calculated. The vector $F$ of probabilities is as follows:

$$F = [0.192 \ 0.135 \ 0.132]$$

Finally, on the step 6 the web service with the maximum probability to satisfy the client’s requirements is obtained. As can be seen from the vector $F$ presented above this web service is $S_1$.

In order to present all steps of the proposed approach, suppose that web services $S_1$ and $S_2$ have the same probability to satisfy the client’s requirements. According to step 7, the expected value of QoS properties for those services need to be calculated. They are presented in the following matrix:

$$E = S_1 \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.850 & 0.12 & 0.869 & 5.075 \end{bmatrix}$$

$$S_2 \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 0.825 & 0.15 & 0.881 & 5.250 \end{bmatrix}$$

The matrix $E$ is sorted according to the algorithm that is described in the step 8. As a result, the web service $S_1$ is proposed to the client.

5. RELATED WORK

QoS properties are crucial for selecting a web service that best meets the client’s requirements, especially when there are several web services with the same functionality. Therefore, many approaches for QoS-enabled web service selection exist.
Li et al (2010) propose a web service selection approach that is based on the weight of client’s satisfaction from QoS properties of a particular web service. Also, a mathematical model for definition of QoS constraints according to the preferences given by the client is designed. The main drawback of the approach is the inability to ensure that the service recommending algorithm is open, fair and trustworthy. Also, only measurable QoS properties are considered. In contrast, a web service selection mechanism, proposed by D’Mello et al (2008), deals with all types of QoS properties expressing those that are not measurable in terms of integer values. Here, the constraints QoS are presented in a weighted AND-OR tree.

Sha et al (2009) describe QoS based selection model. It is based on the overall QoS of web service that is obtained from a weighted sum of the normalized values of QoS properties. One of the drawbacks of the model is that it considers only measurable QoS properties that can be directly monitored. Also, the normalization of the values of QoS properties in the interval of [0,1] leads to losing valuable information.

Qi et al (2010) have designed a QoS-aware selection method based on credibility evaluation of the web services. The QoS properties are classified in two categories, i.e. negotiable and nonnegotiable. The values of nonnegotiable QoS properties are obtained from historical records of web service execution and cannot be modified by the provider. The negotiable QoS properties can be changed according to the client’s requirements. The credibility evaluation depends on whether the QoS property is negotiable or nonnegotiable. Li et al (2009) propose a web service selection algorithm, based on quantitative QoS prediction method applied to a dynamic environment. The QoS measurement of composite web services is modeled by structural equation in order to measure accurately the changes of QoS properties.

Shao et al (2007) have designed a collaborative filtering based approach that predicts QoS of unused web services taking into account the similarity among experience of clients. The similarity of clients is considered also in a web service selecting model proposed by Wang and Chen (2008). The model is based only on measurable QoS properties of client-side.

Xiong and Fan (2007) propose a QoS-aware selection method of web services based on adoption of fuzzy multiple criteria decision making. The subjective weights of QoS properties reflecting human rating and objective weights representing reliability of evaluation are used to form synthetic weights. Then the overall quality score of a particular web service is calculated using the synthetic weights of QoS properties. Similarly, Tran et al (2009) adopt Analytic Hierarchy Process (AHP) method as an underlying mechanism for QoS-based web service ranking.

Xu et al (2007) present QoS-based web service selection approach that is based on ratings indicating the level of client’s satisfaction with a web service following certain interaction with it. A drawback of the approach is the assumption that the ratings are objective and valid. Thus, the web service selection process becomes untrustworthy. Al-Masri and Mahmoud (2007) define web service relevancy function in order to measure the relevancy ranking of a particular web service. The function calculates the distance between a particular QoS value and the maximum normalized value in its corresponding set. The approach proposed in the paper has some advantages compared with those described above. Firstly, it considers measurable as well unmeasurable QoS properties. This is due to the fact that the values of QoS properties are presented as events with a probability to happen. Further on, some approaches are based on the ratings that are specified by clients after web service consumption. Other approaches use values of QoS properties that are claimed by the web service providers. In contrast, the approach proposed in the paper relies on values of QoS properties obtained during web service execution and stored in a history log. Therefore, this approach appears to be more reliable. Finally, the proposed approach allows clients to specify the priorities of the QoS properties, and this leads to the possibility of distinguishing the services with the same probability to satisfy clients’ requirements.

6. CONCLUSION

In this paper, a QoS-aware web service selection approach is presented. The QoS data collected during web service executions is used for prediction of future values of QoS properties based on probability evaluation. Thus, the client will be able to select a web service with the largest probability of having QoS properties that are as close as possible to the preliminary defined requirements.

Future work includes further validation of the proposed approach with different QoS data. The big challenge is to deal with composite web services. Therefore, we plan to design an approach predicting the
overall quality of a web service composition, based on the expected quality of the web services participating in that composition.

ACKNOWLEDGEMENT

The work presented in this paper was supported by the Human Resource Development Operational Programme 2007-2013 of the European Social Fund under Grant No. BG 051PO001-3.3.04/13 and the National Scientific Fund of Bulgarian Ministry of Education and Science under Grant No. MU-01-143.

REFERENCES


AGILE DATA WAREHOUSING: THE SUITABILITY OF SCRUM AS DEVELOPMENT METHODOLOGY

Roelien Goede
North-West University: Vaal Campus - Hendrik van Eck Blvd, Vanderbijlpark, South Africa

ABSTRACT

Data warehouses are used as data sources for strategic business decision making or business intelligence. Users of information systems today are more used to rapid response from developers than previously. Traditional data warehouse methodologies do not produce rapid results, as vast amounts of time and effort are spent on data integration and quality issues. This paper aims to investigate the suitability of agile development methodologies, specifically Scrum, for the development of data warehouses. An interpretive experiment with case comparison is used to identify suitable and unsuitable characteristics of agile systems development methodologies and specifically Scrum for data warehouse development.

KEYWORDS

Scrum, agile systems development methodology, data warehousing.

1. INTRODUCTION

The aim of this paper is to investigate the suitability of scrum as a methodology for agile data warehouse development. As more and more emphasis is put on rapid response to business problems, the business intelligence (BI) world becomes more aware of agile software development methodology (ASDM). ASDMs were developed for use in traditional online transaction processing systems (OLTPs) which differ substantially from BI systems such as data warehouses (DWs). The question investigated by this paper is whether one can use scrum for data warehousing?

The paper starts with a short discussion on DWs highlighting the differences from a development perspective between DWs and OLTPs in section 2. Section 3 provides background on ASDMs in general and the specifics of Scrum as ASDM. A short literature review on agile data warehousing is given in section 4, focusing specifically on Scrum data warehousing. An interpretive experiment was conducted in which Scrum was used to develop data warehouses. This experiment is reported in section 5. It also contains the findings of the experiment and recommendation for the use of Scrum in DW projects. Conclusions are given in section 6.

This paper aims to contribute in the wider usage of ASDMs by demonstrating the suitability for usage in DWs. DW methodology has been overshadowed by industry role players and this is an attempt from an academic research perspective to contribute to the success of DW projects.

2. BUSINESS INTELLIGENCE AND DATA WAREHOUSING

2.1 Definitions

Business Intelligence (BI) is viewed in this research as the umbrella term describing the actions of management to make strategic business decisions based on the available data in the organisation. Kimball et al. (2008:596) defines BI as “A generic term to describe leveraging the organisation’s internal and external information assets to support improved business decision making.” A data warehouse is information system
(IS) used to integrate different sources in order to provide reliable data to business users. Various definitions for data warehouses exist but the definitions by Inmon (1996) and Kimball et al. (1998, 2008) are widely used. Inmon (1996:33) defines a data warehouse as a subject oriented integrated, non-volatile, and time variant collection of data in support of management decisions. Kimball et al. (1998:19) simply define a data warehouse as “the queryable source of data in the enterprise.”

2.2 Data Warehouse Methodology

There are two main activities in a data warehouse namely data integration and end-user applications. Data integration is a highly technical activity where data from different source systems are extracted, transformed to common definitions and loaded in the data warehouse. Kimball et al. (1998) argues that up to 80% of the total effort of the project is spent on this process. End-user applications are developed to provide access to the data to end-users in a flexible format.

Inmon (1996) and Kimball et al. (1998) are the most influential authors in data warehousing design methodologies (Mailvanganam 2007, Jukic 2006). They differ on various concepts in data warehousing, one of which is the development lifecycle of a data warehouse. Inmon (1996) advocates a lifecycle that he calls the CLDS (reverse of SDLC: systems development lifecycle) with the following phases: 1. Implement data warehouse; 2. Integrate data; 3. Test for bias; 4. Program against data; 5. Design DSS system; 6. Analyse results; 7. Understand requirements. This is a data-driven lifecycle methodology. Kimball et al. (1998, 2008) advocates the use of a requirements-driven lifecycle methodology. Their methodology begins with a data warehouse readiness test, where after user requirements are gathered, followed by modelling, data staging, end-user application design, and maintenance. Kimball et al.’s (1998) notion of a data mart is used in this paper. A data mart is a subsection of the DW associated with a specific business process. Data staging is a term Kimball et al used in 1998 to refer to what is known today as the extract, transform and load (ETL) process.

It is this ETL process and the considerable effort it requires that distinguishes DWs from OLTPs. Business users will only use the DW if they can trust the accuracy of the information it delivers. “One version of the truth” is the heartbeat of a DW. The aim of the DW development team is to identify and correct all inconsistencies in data in the organisation during the ETL process. This is a very time consuming and complex activity since the source systems (OLTPs) are sometimes very old and dependent on specific architectures. The purpose of the source systems is to capture the daily activity in the organisation and the owners are not always motivated to adapt their systems to standards set by the DW development team. All this mean that it is extremely difficult to provide users with quick usable solutions.

Different authors identify success factors in data warehouse design. Ferranti (1998) quotes Inmon: “Building data marts before developing a data warehouse can be one of your biggest mistakes.” Mimno (2001) argues that the most important success factor is to make your data warehouse business-driven. He argues that a technology-driven approach is much more likely to fail than a business-driven approach. Chenoweth et al. (2006) argues that “the success of a data warehouse depends on the interaction of technology and social context”. They promote the idea of user involvement throughout the development process of the data warehouse.

3. AGILE SYSTEMS DEVELOPMENT METHODOLOGY

3.1 Background

A group of software development representatives came together in 2001 and created the following agile manifesto: "We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
• Responding to change over following a plan
  That is, while there is value in the items on the right, we value the items on the left more.” The list of participants can be found at www.agilemanifesto.org. Various specific methodologies were developed in an attempt to practice these principles. According to Hislop et al. (2002:177), there are nine principles that reflect the common core values of all agile processes:
  • “Users must be actively involved throughout the development process
  • Teams (including both users and developers) must be empowered to make decisions without explicit approval from higher management
  • Frequent delivery of products has highest priority
  • Deliverables are evaluated primarily with respect to their fitness for business purposes
  • Rapid iterations and incremental delivery are key to converging on acceptable business solutions
  • No changes are irreversible – backtracking to or reconstructing previous versions must be possible
  • High-level requirements are frozen early to allow for detailed investigation of their consequences
  • Testing is integrated throughout the development life cycle
  • Collaboration and cooperation among all stakeholders is the key to success”

It is difficult to follow agile principles without following a specific agile methodology. The aim of agile methodologies is to give practical guidelines for applying the principles.

Many ASDMs are used today but seven are well-documented and widely used. These include: Dynamic Systems development Methodology (DSDM), Scrum, Extreme Programming (XP), Feature driven development (FDD), Crystal Methods, Adaptive software development (ASD), and Lean Development (LD). According to (Highsmith 2002a) these ASDMs are the core group, as set out in the Agile Software Development Manifesto (Highsmith 2002a, Lindstorm & Jeffries 2004). These 7 methodologies were used to investigate their suitability in data warehousing in a pilot study to the one reported in this research. Scrum along with XP demonstrated potential for the use in data warehousing and a second study was launched to investigate the suitability of Scrum for data warehouse development. A similar study will be conducted using XP.

3.2 Scrum

Ken Schwaber and Jeff Sutherland developed Scrum in 1995. Amudha (2010) gives a concise description of the main characteristics of Scrum. The name is derived from the sport of Rugby where a scrum may be viewed as a group effort to move quickly to counter the opposite team, adjusting the move as it progress. Activities are organized in “sprints” which is a 30 day period or iteration cycle to deliver a working section of the overall project. There are three identifiable roles namely the customer or product owner (prioritisation is one of his key roles), the scrum master (who resolves impediments and facilitates the scrum process) and the team (a self-organising group of 5±2 members). Fifteen minute stand-up daily meetings are held to communicate and monitor progress. Documentation includes: product backlogs, product and sprint burnout charts as well team and organisational impediments reports.

Limitations of agile processes listed by Amudha (2010) include:
1. Limited support for distributed development environments
2. Limited support for subcontracting
3. Limited support building reusable artefacts
4. Limiting support for development involving large teams
5. Limited support for developing safety-critical software
6. Limited support for developing large, complex software.

Some developers argue that agile software development is about small iterative deliverables with small unit tests and early customer involvement. Kruse (2009) argues in contrast that true agility is when a development department reserves the right to change the rules at any time in order to facilitate a process improvement which will increase the rate of passed tests and functionality of the delivered units.
4. AGILE DATA WAREHOUSING

Since very little peer reviewed research reports on the use of ASDMs in DW, non peer-reviewed literature is included in this section. Graziano (2006) formulates 12 principles of agile methodologies applied to data warehousing. These are summarised in table 1.

Table 1. Application of ASDM principles in DW adapted from Graziano (2006)

<table>
<thead>
<tr>
<th>Principle</th>
<th>Application in DW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Satisfy the customer through early and continuous delivery of valuable software.</td>
<td>Valuable software includes BI reports and Dashboard interfaces</td>
</tr>
<tr>
<td>2. Welcome changing requirements, even late in development</td>
<td>Use code generators; start with normalized tables.</td>
</tr>
<tr>
<td>3. Deliver working software frequently.</td>
<td>Deliver one subject area at a time</td>
</tr>
<tr>
<td>4. Business people and developers must work together daily throughout the project.</td>
<td>Good idea, but politics and priorities may interfere.</td>
</tr>
<tr>
<td>5. Build projects around motivated individuals.</td>
<td>Keep units of work small; Don’t deliver a single Enterprise DW</td>
</tr>
<tr>
<td>6. Use face-to-face communication inside development team.</td>
<td>Use daily meetings and keep documentation efficient.</td>
</tr>
<tr>
<td>7. Working software is the primary measure of progress.</td>
<td>Working BI Reports and ETL software is important.</td>
</tr>
<tr>
<td>8. Sustainable development at constant pace.</td>
<td>Keep units of work small. Use XP and Pair programming</td>
</tr>
<tr>
<td>9. Continuous attention to technical excellence and good design enhances agility.</td>
<td>Frequent design reviews leads to faster delivery.</td>
</tr>
<tr>
<td>10. Simplicity – the art of maximising the amount of work not done – is essential.</td>
<td>Use metadata effectively; use code generators; online data dictionary.</td>
</tr>
<tr>
<td>11. The best architectures, requirements, and designs emerge from self-organising teams,</td>
<td>Don’t micro manage staff, Have shared responsibilities.</td>
</tr>
<tr>
<td>12. Team should reflect at regular intervals how to become more effective, and should adjust accordingly.</td>
<td>Use self organising teams with maximum buy-in to the solution.</td>
</tr>
</tbody>
</table>

Arnett (2002) provides the following six architecture principles for an agile data warehouse:

“1. Assemble a staff conditioned to change; 2. Use metadata-driven development tools; 3. Isolate subjects and data tables; 4. Bit-size analysis through denormalisation; 5. Use surrogate keys; 6. Frequent source system feeds.” All of these principles except the first are explicitly advocated by Kimball et al. (1998).

Ambler (2007) presents best practices for agile data warehousing and business intelligence projects. His suggestions include: Do not create a comprehensive, detailed model up front – follow a just-in-time method based on model storming. Do not use a data centred approach but rather use cases or usage scenarios to focus on usage. Active stakeholder participation throughout the project is essential. These suggestions leads to the following list of best practices: “ 1. Take an evolutionary approach; 2. Embrace Change. 3. Deliver working software regularly, 4. Strive for iterations of one to two weeks. 5. Test throughout the lifecycle. 6. Involve operations and support people early.”

Some of these principles of ASDMs are in conflict with the design methodology for data warehousing proposed by Kimball et al. (1998). Since the literature reference in this section is all non peer reviewed it inspires the academic society to investigate the suitability of ASDMs for DW development.
From DW literature it is clear that Kimball’s (Kimball et al. 1998, 2008) ideas of iterative development and dimensional modelling with maximum user involvement supports the core principle of ASDMs. On the negative side of the suitability of ASDMs for DW, Sen and Sinha (2007) reports that very little information is offered in terms of change management by current vendor methodologies. Two aspects of data warehousing hinders the use of ASDMs in development. The first aspect is the large amount of effort and time spent on data integration. It is difficult to deliver meaningful parts in 30 days. The second aspect is that of complexity as ASDMs do not have a proven track record in complex systems.

Since Inmon’s (2002) method is not driven by requirements collection it is difficult to link it with ASDMs. This research did not apply Inmon’s (2002) method in agile data warehousing.

5. CASE STUDY: SCRUM DATA WAREHOUSING

5.1 Research Design

An interpretive experiment spanning three years was conducted in order to establish the suitability of Scrum for the development of DWs. The aim was to identify suitable and unsuitable characteristics of Scrum for developing a DW. Eight data warehousing teams (two in year one and three in years two and three respectively) used Scrum to develop a DW according to the projects Kimball et al.’s (1998) architecture and lifecycle were used where appropriate.

The teams were from a group of fourth year Information Systems students, who are trained in DWs and systems development methodologies. The DW requirements were the same for each group. Students had to use a programming language for data staging. An Oracle database was used to store the dimensional tables. Students had to design their own access tools in combinations with cube generators such as Cognos.

A total of eight interviews (of approximately 20 minutes each) were conducted with each of the teams. The progress of the team was discussed during each of the development stages. Interviews focused on (1) the team’s understanding of Scrum, (2) their ability to apply the Scrum in the current phase of the project, (3) Scrum related problems, (4) technical problems, and (5) any other matters the group wanted to discuss.

The teams had to hand in essential DW documentation such as dimensional models and source-to-target mappings. A final evaluation of the DW was done at the end of the project. In this extensive interview (duration 1 hour) the teams had to demonstrate their DWs. They also had to discuss the suitability of Scrum in terms of positive and negative experiences.

In order to simulate a typical problem situation the requirements were updated and changed frequently. Users shared different insights in the usage of the DW as the project progressed. Data and requirements for the data warehouses were supplied by one of the industry partners of the university. The data was desensitised, but large volumes were used to keep the project as realistic according to industry standards as possible. Specific DW documentation can be viewed as a user requirement with high priority in the system.

The key advantage of using student teams is they have limited experience in other methods and are therefore prepared to follow a method strictly. The students do not have enough practical experience to adapt methodologies to suit the specific problem environment. Although this can be viewed as a disadvantage it can also be viewed as an advantage in this research as they made an attempt to implement Scrum as it is described in literature without any modifications.

5.2 Data Analysis and Representation

Seaman (1999) argues that when data can be divided into “cases” it is appropriate to analyse qualitative data with cross-case analysis.

The data in this study were analysed similarly to the method described by Seaman (1999). At first, one case (all information collected from a specific team) was documented and described in terms of suitable and unsuitable characteristics of Scrum for DW development. Care was taken to distinguish between characteristics applicable to all ASDMs and those specific for Scrum. After this, a second case (another team’s information) was compared with the first case’s data. The lists of suitable and unsuitable characteristics general to all ASDMS were expanded to represent both cases’ information. Some of the
characteristics already in the list were supported while others were refuted. Notes were made on which cases could be associated with specific characteristics. Separate lists were compiled for Scrum, consisting of characteristics that are suitable or unsuitable for DW development. The process described here for the second case was repeated for the other cases. The result was a set of suitable and unsuitable characteristics applicable to all ASDMs and a set of suitable and unsuitable characteristics for Scrum. Excerpts of two of these tables are given as a clarification of the discussion.

Following is an excerpt of the list containing suitable and unsuitable characteristics of ASDMs in general for the development of DWs.

Suitable characteristics:
- The data warehouse design improved as incremental and iterative development progressed.
- ADSMs encourage and improve development of a DW in an agile environment such as the one created by the users of this DW.

Unsuitable Characteristics:
- ADSMs do not prescribe specifics in terms of DW architecture. Teams used Kimball et al.’s (1998) architecture model.
- Business users do not have enough time to be involved as expected by the ASDM.

Following is an excerpt of the list containing suitable and unsuitable characteristics for Scrum. Suitable characteristics:
- The post-game phase is useful for implementation and maintenance.
- The pre-game is very useful for detailed planning.

Unsuitable Characteristics:
- Scrum does not provide guidelines of technical implementation of the ETL process.
- Sometimes difficult to determine exactly how a data mart should be divided into “small releases”

The data analysis of the study was used to verify and extend the initial literature studies resulting in the identification of certain suitable and unsuitable characteristics when using Scrum for DW development. Some of these characteristics are applicable to all ASDMs while others are specific to Scrum. The first part of this section reports on the suitable and unsuitable characteristics that are applicable to all the investigated ASDMs, while the second part focuses on Scrum.

### 5.3 Suitable and Unsuitable Characteristics of ASDMs in General in DW Development

ASDMs general principles are broadly suitable for data warehousing. For example: Communication between team members and stakeholders (source system owners and business users) are critical for a successful DW. This is a key feature of all ASDMs. As advocated by ASDMs, teams should be effective and take ownership of their own problems and the process to solve these problems.

However, it is difficult to decide exactly what constitutes an increment in terms of ASDMs, especially at the start of the project. A large amount of work must be done before the first data mart can be delivered to users. The initial data mart should not be too complicated. The first data mart can be divided into smaller increments to facilitate the use of ASDMs. The ETL process can for example be viewed as one increment. Some teams developed user applications before data quality problems were resolved in order to have a shorter first increment. This clearly brought to light the trade-off between short first iteration and high data quality. Teams who first solved all the data quality problems took longer to deliver their first working application. They struggled to incorporate the changes made by the users. Some of the changes eliminated data that teams spent a long time on improving the quality.

Requirements collection: ASDMs were suited for Kimball et al.’s (1998) approach since requirements are collected at the beginning of the development lifecycle. Requirements were added as users were shown working programs. It was confirmed that users only understand what they will be doing with the data warehouse once they see the first working concept.

Data modelling: All teams commented on the ease of understanding of Kimball et al.’s (1998) dimensional models for the team members and the users. The use of these models allows the end users to verify the design of a specific data mart. User verification is key to all ASDMs. All teams referred to the “graceful changeability” of Kimball et al.’s (1998) dimensional model as a positive feature to incorporate changes in user requirements.
Data staging: Lack of technical guidance of ASDMs was identified by all teams as a difficulty. This stage is difficult from a user involvement perspective since the users are not knowledgeable on the technical detail of the data. The source system owners should be viewed as users in terms of the ASDMs. These source system owners are sometimes reluctant to be involved in the DW effort in the organisation since they do not experience ownership. ETL tools and technical knowledge of the tools are essential in providing quick response to changing requirement.

Data access and deployment: ASDMs proved most suitable for this phase, as development of access tools and applications are comparable to other more general software development projects for which ASDMs were developed.

5.4 Scrum in DW Development

Most of the findings could be attributed to ASDMs in general rather than Scrum specifically. The first iteration of a specific data mart will most often take longer than the 30 day sprints. After the first 30 days it is difficult to deliver something that users can test. Teams took between 4 and 5 sprints before the fact table was populated with high quality data. Some of the teams worked during this time on end-user applications. It was explained to end-users that the data in the systems were not trustworthy.

The moral in the teams were high – something they contributed to their daily meetings. On the negative side some team members argued that the rushed nature of meetings and sprints caused that great ideas were never implemented.

Three teams reported that they could have spent less time in the beginning on planning as everything changed from their initial plans.

6. CONCLUSIONS

This paper explored the suitability of using Scrum in Data warehousing. Data analysis reflected on suitability of ASDMs in general for the development of data warehouses as well as the suitability of Scrum’s characteristics.

The academic contribution the authors hope to make with this paper is to identify shortcomings in the Kimball et al. (1998) methodology as well as to provide practical guidelines for the development of agile data warehouses.

The traditional waterfall method of Kimball et al. (1998) advocated the collection of all requirements at the start of the project. All data quality issues may then be addressed before any working data mart is delivered. However in industry users only understand their requirements as the first working data marts are delivered. Only then are they able to articulate their needs. Data quality issues are more prevalent in data warehouses than in operational systems. This research showed that a just-in-time approach to data quality is usable and advisable. Developers should not spend weeks to rectify data quality in fields that may become redundant as users better understand their own requirements.

This research proposed that a subgroup of developers continue to work on data quality discrepancies while the rest of the group delivers data marts. The users must be clearly informed that the initial data marts are not yet fully trustworthy as not all data quality issues are resolved. End-users should be explained the specifics of the outstanding data quality problems as they as often in a better position to resolve political issues amongst source systems owners than the developers are.

The data warehouse methodology by Inmon (2002) whereby all technical work is completed before requirements are identified is against the nature of ASDMs of rapid delivery and user involvement.

From this study it became clear that ASDMs are indeed suitable for the development of DW. However, care should be taken not to neglect the importance of technical accuracy since it is difficult for users to verify. The two important features of agile development are “speed” and “flexibility”. Most teams participating in the study found changes in the requirements easier to incorporate than rapid development of user testable units.
ACKNOWLEDGEMENT

The researcher wants to acknowledge the study of J Grey that served as a pilot study for this study. Some of the data analysis done by him was used in this study on Scrum.

REFERENCES


INTEROPERABLE PROCESS MONITORING USING CLINICAL IT-STANDARDS AND HEALTHCARE INFORMATION SYSTEMS

Eva Gattnar¹, Okan Ekinci¹ and Vesselin Detschew²

¹Clinical Competence Center Cardiology, Siemens Healthcare - Allee am Röthelheimpark 3a - 91052 Erlangen, Germany
²Institute of Biomedical Engineering and Informatics, Ilmenau Technical University - POB 100565 - 98684 Ilmenau, Germany

ABSTRACT
The current interaction between medical devices and IT-systems in hospitals is characterized by inadequate interoperability and integration. One challenge is the monitoring of patient-related process lead times across the involved IT-systems. Therefore we propose a novel way of retrieval of quality and performance metrics along the clinical pathway of time-critical diseases in the context of various clinical standards and systems. For this reason, we investigated the hospital information systems and available clinical communication standards regarding the collection and communication of timing and process information. We found that the widely used clinical communication standard Health Level 7 (HL7) and the XML-based Clinical Document Architecture (CDA) can be extended to provide a valuable service for the acquisition and collection of process data including timestamps. Therefore we propose an integrated technical and process-oriented approach which provides the basis for an automated and standardized process monitoring, quality controlling and benchmarking and therefore represents a promising opportunity to recognize and eliminate workflow bottlenecks. In time-critical diseases this typically directly translates into an improved quality of care.

KEYWORDS
Clinical document architecture, process monitoring, key performance indicators, clinical pathways.

1. INTRODUCTION
Optimization of clinical pathways is one of the key elements of quality improvement initiatives in hospitals. The management of time-critical diseases such as heart attack and stroke – although highly standardized – can be more or less time-consuming depending on the workflow bottlenecks in a particular hospital. To identify bottlenecks in a process that overall can be accomplished within one or two hours, requires a thorough understanding of both workflow and durations each of the sub-processes take. During a hospital stay, a patient crosses several departments, which are based on various information system-catchment areas. This set of departments and IT-system boundaries represent process interfaces that must be bridged. On object-level (patient, blood sample etc.) the interfaces are connected temporarily by a local transformation of the objects. At the level of IT-systems, however, these interfaces very often do not exist. The lack of interoperability, i.e. the ability of an information system to participate in a complex information management process in a concerted fashion with a number of other information systems (Channin, 2009), is a major problem for any type of standardized and automated process monitoring because it complicates a digital information exchange, thus, hinders the full exploitation of the potential use of data.

The current interaction between medical devices and IT-systems in hospitals is characterized by inadequate interoperability and integration. One challenge is the IT-department and cross-system detection of patient-related process lead times. We propose a novel way of retrieval of quality and performance metrics along the clinical pathway of time-critical diseases in the context of various clinical standards and systems. In our approach the retrieval of KPIs is achieved using the example of a dedicated set of time-based process indicators. For this reason, we investigated the hospital information systems and available communication standards regarding the collection and communication of timing and process information. We found that the
widely used clinical communication standard Health Level 7 (HL7) and the XML-based Clinical Document Architecture (CDA) can be extended to provide a valuable service for the acquisition and collection of process data including timestamps. Therefore we propose an integrated technical and process-oriented approach which provides the basis for an automated and standardized process monitoring, quality controlling and benchmarking. This represents a promising opportunity to recognize quality lacks e.g. in form of bottlenecks over several clinical department and system borders and to further optimize processes by eliminating them. In time-critical diseases this typically directly translates into an improved quality outcome for patients.

2. IT-BASED INFORMATION TRANSFER IN HOSPITALS

Healthcare processes are nowadays heavily dependent on Information Technology (IT). With the enormous impact and rapid evolvement of IT, there is a major demand for standardization in health care IT (Wirsz, 2000). DICOM (Digital Imaging and Communications in Medicine) as one of the clinical standards defines the format and mechanism for exchange and storage of information in radiological environment (DICOM, 2011). DICOM includes a large number of specific services and the producers are committed to their implementation because of so-called „Conformance Statements“. By comparing the declarations of different systems, basically it is possible to determine, whether they can interact with each other (DICOM, 2009). Manufacturers have at least a minimum level of general conformance requirements to meet, in order to demonstrate general conformity to the DICOM Standard. Conformance Statements are an integral part of the DICOM standard.

So-called DICOM worklists are generated by the Radiology Information System (RIS). They contain patient data, details of the examination orders, the procedure parameters and appointments and can be accessed by image acquisition devices (modalities) for planning examinations. In return, a modality is able to generate so-called Modality Performed Procedure Steps (MPPS), a structured information (such as start time, end time, status, dose, material consumption etc.), which can be sent to the RIS at the beginning (“MPPS in progress”) and the end (“MPPS completed”) of an examination (Noumeir, 2005).

Outside the radiology department of a hospital the second important clinical standard, Health Level 7 (HL7), standardizes the transfer of patient information between IT systems (HL7, 2011). The communication in HL7 follows the message-based communication principle. Accordingly, the various systems interact with each other via messages. Data exchange is initiated between two or more systems through an event from the real world (called a trigger event), such as the admission of a patient. This event implies the exchange of data between the various systems, distinguishing between specific and general requests. The patient’s master data request initiated by a laboratory information system (LIS) would correspond to a specific request. General requests are on the contrary usually caused by short-term events in the real world. If such a trigger event is registered in a system, a transaction is started, which in turn causes an event and respectively a transaction on the target system (Bärwolff et al, 2006). While DICOM is a standard for image transmission in radiology, HL7 is a standard for electronic data exchange outside of radiology in the catchment area of the hospital information system (HIS). A HIS is an IT-based management system. It supports medical and clinical patient care and documentation, organization, administration and communication in a hospital (Haas and Kuhn, 2007; Huang, 2010). Another important task of the HIS is to evaluate the costs and the efficiency of hospital organization for a longer period of time (Huang, 2010).

A radiology information system (RIS) is one of the IT-systems in the field of radiology, which controls the daily tasks of one or more radiological departments. It is responsible for both the administrative and clinical support in the radiology to relieve the staff from administrative tasks. On the other hand, it improves the quality of radiological examinations (Huang, 2010). PACS (Picture Archiving and Communications Systems) as another component represent the technological core of a modern, digital radiology department (Ralston and Coleman. 2009). The main task of the PACS is to store, distribute and display medical images for interpretation and evaluation. Further IT-systems in radiology departments are so-called modalities. Modalities are known as imaging systems in the medical industry. Modalities represent combined with PACS the technological core of a modern, digital radiology department (Ralston and Coleman, 2009). Table 1 summarizes the interoperability issues between the described IT-systems including HIS, RIS, modalities and PACS.
Table 1. Interoperability of clinical systems (cf. (Wirsz, 2000)).

<table>
<thead>
<tr>
<th>Systems</th>
<th>HIS</th>
<th>RIS</th>
<th>PACS</th>
<th>Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIS</td>
<td>HL7</td>
<td>HL7</td>
<td>__________</td>
<td>__________</td>
</tr>
<tr>
<td>RIS</td>
<td>__________</td>
<td>HL7</td>
<td>HL7/DICOM</td>
<td>DICOM</td>
</tr>
<tr>
<td>PACS</td>
<td>__________</td>
<td>__________</td>
<td>DICOM</td>
<td>DICOM</td>
</tr>
<tr>
<td>Modality</td>
<td>DICOM</td>
<td>DICOM</td>
<td>DICOM</td>
<td>DICOM</td>
</tr>
</tbody>
</table>

3. PROCESS BASED TIME ACQUISITION

Healthcare organizations such as hospitals are interested in measures about their quality of care. A common way to retrieve quality information is to examine the performance of organizational structure (structural performance) and processes (process performance) as well as outcomes (outcomes performance) by means of Key Performance Indicators (KPIs) (NHOP, 1997). The challenge is to develop measurable, significant, appropriate and quality-relevant indicators. Such sets of indicators are being provided by healthcare organizations in many countries (e.g. Joint Commission on Accreditation of Healthcare Organizations (JCAHO, 2011), Agency for Healthcare Research and Quality (AHRQ, 2011), National Health Service (NHS, 2011) etc.).

Clinical IT-systems act in our approach as enabler for end-to-end process monitoring and therefore for clinical quality improvement as well as risk and cost reduction. The assignment of KPIs to the clinical IT-systems is presented in Figure 1 by means of dedicated KPIs using the example of a simplified inpatient care process including four basic sub-processes (admission, diagnostics, treatment and discharge). The affected
IT-Systems are the hospital information system (HIS), in the field of radiology department: the RIS, Modalities and PACS and outside the radiology the laboratory information system (LIS). The information about the process metrics necessary for process monitoring is stored in our approach in so-called KPI-cards (see Table 2). These are structured and short descriptions of KPIs including further information about the start- and end-events, the IT-systems activating the events, the calculation of the KPIs, practical examples as well as related KPIs. In addition, the related process modules are referred to. For modeling purposes, Event-driven Process Chain (EPC) is used. EPC is a process modeling language which represents the temporal and logical dependencies of events and processes, and also allows the explicit notation of events, where process performance measurements can be done.

Table 2. KPI-Card: “Order to Imaging”.

<table>
<thead>
<tr>
<th><strong>Order to Imaging</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
</tr>
<tr>
<td><strong>Calculation</strong></td>
</tr>
<tr>
<td><strong>IT-systems</strong></td>
</tr>
<tr>
<td>$e_{start}$ Hospital Information System (HIS) (direct)</td>
</tr>
<tr>
<td>$e_{end}$ Radiology Information System (RIS) (indirect: DICOM* Worklist**)</td>
</tr>
<tr>
<td><strong>EPC-module</strong></td>
</tr>
<tr>
<td>$e_{start}$ Module Order</td>
</tr>
<tr>
<td>$e_{end}$ Module PerformModality</td>
</tr>
<tr>
<td><strong>Example</strong></td>
</tr>
<tr>
<td><strong>Related KPIs</strong></td>
</tr>
</tbody>
</table>

---

* DICOM (Digital Imaging and Communications in Medicine) as one of the clinical standards which defines the format and mechanism for exchange and storage of information in a radiological environment.

** DICOM Worklists contain patient data, details of the examination orders, the procedure parameters and appointments. They are sent by a RIS-system and can be accessed by image acquisition devices (modalities) for planning examinations.

*** DICOM "MPPS" is a structured information (such as start time, end time, status, dose, material consumption etc.), which can be sent by a modality to the RIS at the beginning ("MPPS in progress") and the end ("MPPS completed") of an examination.

---

4. CLINICAL INFORMATION INTERCHANGE USING IT-STANDARDS

Health Level 7 (HL7) is – as described above – a widely used and supported standard for the interchange of clinical, financial, and administrative information among heterogeneous information systems in the healthcare environment. (HL7, 2011; Chronaki, 2001; Campbell, Stetson 2003). In conforming to the HL7 standard, it is possible to share healthcare information between the hospital information system (HIS) and other departmental systems (e.g. radiology (RIS), laboratory (LIS) etc.) (see Figure 2). HL7 addresses the highest level (level 7) of the OSI model (Open Systems Interconnection model) and facilitates data communication in hospitals by providing rules to convert abstract messages associated with real-world events into strings of characters comprising an actual message. In December 2001, HL7 Version 3 was initially released. It uses an object-oriented methodology and a Reference Information Model (RIM) to create HL7 messages (Huang, 2010). RIM provides a coherent shared information model at the entire scope of health care IT that contains all data content relevant to HL7 messages (Paterson, 2002). It includes more than 100 classes and more than 800 attributes and defines the relationships of each class. RIM provides thus an explicit representation of the semantic and lexical connections between the information in the field of HL7 messages (HL7, 2011; Huang, 2010).
Considering intensive standardization efforts in the health care domain, the widely used extensible markup language (XML) is increasingly used, to exchange medical records (XML, 2011). XML is a system-independent language describing information data in a standard format and making data portable. Especially for clinical purposes, the XML-based Clinical Document Architecture (CDA) was developed. It is a document markup standard for the structure and semantics of exchanged clinical documents. CDA is envisioned as a hierarchically organized set of document schemas and document type definitions (DTD), which define the semantics and structural constraints necessary for the exchange of clinical documents (Paterson, 2002). A CDA-document is a defined and complete information object that can exist outside of a message and can include text, images, sounds, and other multimedia content. The CDA supports shared care between hospital-based and community-based physicians, knowledge integration by permitting external links to other documents, and outcomes research through the capture of discrete and coded clinical data (Chronaki, 2001).

CDA documents derive their meaning from the HL7 RIM. The elements and attributes and the relationships among these elements and attributes are drawn from the RIM and expressed in XML (Paterson, 2002). As XML-based documents, they must all have a header section that gives details of the patient, event, date of creation etc. The document body involves three levels and is defined by a DTD (Chronaki, 2001). CDA-Level 1 is thereby the root of the hierarchy and unconstrained. Each additional level adds further specificity and constraints to the architecture. CDA-Level 2 has structured sections and allows constraints to be imposed. Level 3 contains furthermore structured entries within the document body (Paterson, 2002).

5. IT-BASED CLINICAL PROCESS MONITORING

Most clinical departments in a hospital (e.g. cardiology, radiology, laboratory, pathology, neurology etc.) have their own specific operational requirements that differ from the general hospital operations. For this reason special information systems with different workflow environments may be needed in these departments (e.g. radiology information system, laboratory information system etc.) (Huang, 2010). Often the hospital information system (HIS) doesn’t support their operations and must develop mechanisms to integrate data between these systems and the HIS. The integration of the information systems involved in a clinical process is extremely important for a hospital to monitor the process performance.

The HIS is in addition to business and administration processes also responsible for clinical processes. It should provide the access to patient clinical results generated by the various clinical departmental information systems and broadcasts patient data with the HL7-standard (see previous section). However, a standardized end-to-end clinical process monitoring is not supported.

An end-to-end process monitoring is based on timestamps collected at the start and the end of dedicated sub-processes resp. process steps. This timestamps are collected by several different information systems. For example, the start event (“Door”, i.e. the admission of the patient at hospital) of the KPI “Door to Imaging” is stored in the HIS, while the end event (“Imaging”, i.e. the finishing of imaging e.g. computer tomography) is stored in the radiology information system (RIS) (see Figure 1 and Table 2). Furthermore, several clinical workflow tasks are performed one after another which results are indeed stored in the HIS. However, the historical workflow timing information necessary for the end-to-end process monitoring is unrecoverable lost. Especially the monitoring of medication is not possible. But exactly this is a very important feature and relevant for quality management in hospitals. Medication (e.g. if aspirin is given after arrival at hospital in hospital...)
patients with acute heart attack) is a key parameter in clinical guidelines published by national and international health care organizations (e.g. European Society of Cardiology). Furthermore, the assurance of accurate medication diminishes risks in health care and improves the quality of care. Figure 4 shows the described problem in a demonstrative manner.

Therefore we present in the following a newly approach to enable the end-to-end process monitoring in a heterogeneous health care environment. Our approach provides a new possibility to concentrate information about the clinical process in a single complete clinical information object. For this purpose, the HL7 communication standard is used which ensures a standardized and efficient data exchange between different applications of different departments of a hospital (see previous sections). The Clinical Document Architecture (CDA), one developed as part of the HL7 standard document architecture, offers the appropriate technological framework for the concentration of process and time data in a single information object. CDA is a document standard describing the structure and content of clinical documents in the field of health care for the purpose of electronic exchange (see previous section). However, CDA-documents contain only data such as treatment measures, the roles involved, examination results or information on diagnoses and treatments. They don’t provide information about the clinical process.

In our approach, the XML syntax of CDA-documents is extended for the purpose of process monitoring. The CDA-documents are in our approach also used as a basis for a free and integrated access to relevant clinical process data. We expand the XML-based CDA-document structure therefore to store process information and timestamps within new introduced XML-tags (see Figure 5). Such new CDA-documents can be provided with a “start” timestamp indicating the beginning of the performed process step as well as the process step data describing the current process ID and type. After completion of the step, the “end” timestamp can be stored. Process data information stored in the CDA-document is derived from the HL7 messages sent by the departmental systems or the HIS itself. A special process database is used to collect all relevant clinical process steps. They can be assigned to the appropriate process entry within the CDA-document. The document itself can be referenced by the hospital information system (HIS). After the clinical step is finished, the document has to be updated with the process and timing data and used in the next step.
Following this approach, the CDA-document will grow along the clinical pathway (i.e. the clinical end-to-end process) providing a continuous and accumulated process history involving all performed clinical steps involving data received from crossing departmental IT-systems. Since each CDA-document entry corresponds to a clinical process step-entity, any CDA-document contains the complete information about all previously performed examinations and interventions during a patient stay in hospital. Therefore each document entry is labeled by a unique and accumulated ID. In this manner each document references all other entries generated in history and establishes an end-to-end process data chain. Such a solution allows an integrated and standardized end-to-end process monitoring.

6. CONCLUSION AND OUTLOOK

With the implementation of a CDA-based solution for acquisition and collection of process data including timestamps, the interoperability across the entire process (e.g. between the hospital information system (HIS) and the radiology information system (RIS)) will be ensured. The described method of process monitoring may be very helpful for identification of bottlenecks in clinical workflow. This requires, however, that all involved systems support first HL7 in the XML-based Version 3 and furthermore the CDA standard.

The presented IT-based approach makes it possible that the model can be used as a means of benchmarking and thus enables a more transparent standardized internally (over several department and system borders) and externally benchmarking (e.g. in form of public performance measures or quality reports). In addition, our newly developed clinical process model can be used as an instrument for data collection and clinical pathway monitoring. Clinical pathways are evidence-based treatment processes. They are important in the context of quality management and thus help towards the continuous quality improvement in patient care. Our clinical process model was designed primarily to monitor the clinical process by means of process metrics. Secondly it accelerates the modeling of disease-specific and hospital-specific clinical pathways and the measurement of KPIs in hospitals. In future it could be used to support the clinical pathways construction.

In summary, it is established that a patient's process-oriented view represents a promising opportunity to accurately monitor the patient flow over several clinical department and system borders. The acquisition of KPIs based on established clinical standards like HL7 and CDA is a promising way to support quality improvement efforts in healthcare. Beyond the primary goal of improving quality, KPIs can be published e.g.
as part of quality reports and also can help to improve the reputation of a hospital. The retrieval of KPIs in
the presented way enables standardized process monitoring and measurement of workflow bottlenecks,
which can lead to more accurate analyses and workflow optimization by reduction of lead times and
improving the quality of care in the field of time critical diseases.

Furthermore, the solution enables interoperability in means of supporting and completing clinical
Electronic Health Records (EHR) with process information and providing an automatic retrieval of relevant
clinical process data. In this way, not only the process monitoring will be facilitated but also reporting work
may be reduced. To preserve the idea of the standard, extensions regarding CDA as part of the HL7 standard
document architecture require in the long term however the adoption of the HL7 standard itself.

REFERENCES

January 2011].

Inpatient Care. Springer, New York, USA.

and Applications for PACS Professionals. Springer, New York, USA.

Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society. New York, USA,
pp. 3504-3507.

[Accessed 4 December 2010].


JCAHO, 2011. Joint Commission on Accreditation of Healthcare Organisations. [online] Available at:

NHOP, 1997. Quality and Outcome Indicators for Acute Health Care Services, National Hospital Outcomes Program,


Washington, USA, pp. 1200-1209.

Foundations and Applications for PACS Professionals. Springer, New York, USA.


<http://www.w3c.org/XML>
ALTERNATIVE QUERY DISCOVERY FROM THE WEB FOR DAILY MOBILE DECISION SUPPORT

Shun Hattori

School of Computer Science - Tokyo University of Technology - 1404-1 Katakura-cho, Hachioji - Tokyo 192-0982
JAPAN

ABSTRACT
We daily make decisions on whether or not we should do an activity (behavior) and further on whether or not we had better do its alternative activity. If we can acquire the necessary information for making such decisions situation by situation in mobile/ubiquitous computing environments, we would be able to make better decisions and might be happier. So, this paper proposes a Mobile DSS (Decision Support System) that discovers the alternative activities to a mobile user’s intending (original) activity from the Web by using text mining techniques and then automatically generates her/his activity-based query to the existing location-free information retrieval systems such as mobile Web search engines.

KEYWORDS

1. INTRODUCTION
In our daily lives, we continuously make an absolute decision on whether or not we should do an activity (behavior), and further make a relative decision on whether or not we have a better alternative activity to do than the original activity. If we can acquire the necessary information for making such a decision situation by situation in mobile/ubiquitous computing environments, we would be able to make a better decision.

In recent years, the amount of information available on the Web has been growing exponentially and also mobile computing environments have been improved and maintained. Therefore, the Web might be now the best source of information for daily mobile decision-making. However, the Web would not be useful in practice if we cannot filter the noisy information for making such a decision and collect only the necessary information for making the decision as soon and easily as possible, because we have to make the decision more quickly in mobile computing environments than in immobile ones at home or office.

This paper proposes a novel Mobile DSS (Decision Support System) that allows a mobile user to search the Web for the necessary information for making a decision on whether or not s/he should perform her/his intending activity. When a mobile user or AI robot firstly gives an activity expression, the system discovers its alternative activity expressions and offers them to the mobile user or AI robot’s database. When s/he secondly selects one of them, the system generates the activity-based query and searches the Web by submitting it to the existing location-free information retrieval systems such as mobile Web search engines.

The remainder of this paper is organized as follows. Section 2 introduces expressions of activities (behaviors) in my proposed Mobile DSS. Section 3 gives an overview of my proposed system. Section 4 describes my method to discover typical activity expressions for a given place-name (i.e., typical activities at a given place) from the Web by using text mining techniques. Section 5 explains my method to discover alternative activity expressions to a mobile user’s intending (original) activity from the Web. Section 6 presents related works to this paper. In Section 7, I conclude this paper and mention future works.
2. EXPRESSION OF ACTIVITY IN MOBILE DSS

This section introduces expressions of activities in my proposed Mobile DSS (Decision Support System). The system helps mobile users or AI robots to search the Web for the necessary information for making their decision on whether or not they should do an activity (behavior). Therefore, the system has to receive what their intending activity is in a specific expression from the mobile users.

A mobile user’s intending (or pre-done or now-doing) activity on which s/he requests information would be specified by the following five components:

1. **Action**: what action (verb) does s/he perform?
   - e.g., “buy”, “read”, “enjoy”, “play”, “eat” or “drink”.

2. **Object**: what object does s/he perform the action on?
   - e.g., names of class to which it belongs such as “book” or “novel”, names of its concrete instance such as “Harry Potter” or “Lord of the Rings”.

3. **Place**: where does s/he perform the action on the object?
   - e.g., names of class to which it belongs such as “bookstore”, names of its concrete instance such as “Barnes & Noble” or “Amazon.com”.

4. **Time**: when does s/he perform the action on the object in the place?
   - e.g., date, time of day, time intervals such as “morning” or “night”.

5. **Subject**: who is s/he?
   - e.g., sexuality such as “boy” or “girl”, ages such as “young” or “elder”.

However, we mainly have the following two types of retrieving information for making our daily decision in mobile computing environments:

- **Search for Object and information about it**
  We want to know what object we can perform our intending action on here and also information for going there such as map of surrounding area or route guidance. For example, when a mobile user intends to buy a “Harry Potter” book here, s/he would be thankful to know the following information:
  1. Whether or not should s/he buy the book here?
  2a. Whether or not does there exist a better book than the book?
  2b. Whether or not does there exist a better class (category) of goods named by “Harry Potter” than “book”?
  3a. What is the alternative book? (e.g., “Lord of the Rings”)
  3b. What is the alternative class? (e.g., “movie”, “dvd” or “game”)

- **Search for Place and information about it**
  We want to know where we can do our intending activity and also information for going there such as map of surrounding area or route guidance. For example, when a mobile user intends to buy a book at a certain bookstore, s/he would be glad to know the following information:
  1. Whether or not should s/he go to the bookstore to buy a book?
  2. Whether or not does there exist a better place than the bookstore?
  3. Where is the alternative bookstore?

Therefore, in this paper, my proposed Mobile DSS expresses an activity as a set of a verb of **Action** and three nouns of **Object Class**, **Object Instance**, and **Place**. For example, an activity to buy a “Harry Potter” book at a bookstore is expressed by

[a:“buy” oc:“book” oi:“Harry Potter” p:“bookstore”].

3. SYSTEM OVERVIEW

Figure 1 gives an overview of my proposed system for daily mobile decision support, which consists of the following three steps:

Step 1. Discovering Typical Activities at Current Place
When a mobile user accesses the Mobile DSS via her/his mobile device at a place, the system automatically discovers the typical activity expressions of her/his current place by mining the Web/Blog as described in detail in Section 4, and then offers them to her/him in order to allow her/him to use them as reference to specify an expression of her/his intending activity. For example,
“bookstore” → [a:“buy” oc:“book” p:“bookstore”].

Step 2. Discovering Alternative Activities to Original Activity
When s/he gives an expression of her/his intending (original) activity while referring the typical activities of the current place, the system discovers expressions of its alternative activities by my method described in detail in Section 5, and then offers them to her/him in order to allowing her/him to open up the alternatives from her/his original intending activity. For example,

{[a:“buy” oc:“book” oi:“Harry Potter” p:“Waldenbooks”],
 [a:“buy” oc:“cd” oi:“Harry Potter” p:“Waldenbooks”],
 [a:“buy” oc:“book” oi:“Lord of the Rings” p:“Waldenbooks”],
 [a:“buy” oc:“book” oi:“Harry Potter” p:“Barnes & Noble”].

Step 3. Generating Activity-based Query
When s/he selects one of the alternative activity expressions offered in Step 2, the system generates the activity-based query and searches the Web by submitting it to the existing location-free Web search engines such as [Google Mobile, 2011].

4. DISCOVERING TYPICAL ACTIVITIES OF PLACE FROM THE WEB

This section describes my method to discover the typical activities that people often do at a given place by mining enormous text documents from the Web/Blog, which consists of the following two steps:

Step 1. Extracting Candidates for Typical Activities from the Web
When a name of class or instance of place p is given (e.g., “bookstore” or “Barnes & Noble”), my proposed Mobile DSS (Decision Support System) firstly collects sentences from the Web/Blog, each of which contains the expression “at ... p” and whose subject is either “I” or “we”, by submitting [“at * p”] as a wildcard (*) query to [Google Blog Search, 2011].
Next, the system extracts a set of a verb and its objective noun (if any) from each sentence and regards the set as a candidate for typical activities at the place named by \( p \).

Step 2. Weighting Candidates for Typical Activities

Secondly, the system assigns the following weight, to each candidate which is a set \( s \) of a verb \( v \) and its objective noun \( n \) (if any) for the given place-name \( p \):

\[
weight_{p}(s) := \frac{df(q_{p}^{at})^2}{df(q^{at}) \cdot \log_{2}df(q^{at})}
\]

\[
activity(s) := \begin{cases} "v_{ed}" & \text{if } n \text{ is null} \\ "v_{ed} \ast n" & \text{otherwise.} \end{cases}
\]

\[
q^{at} := ["activity(s) at"]
\]

\[
q_{p}^{at} := ["activity(s) at \ast p"]
\]

where \( df(q) \) stands for the number of documents retrieved by submitting a query \( q \) to Google Blog Search and \( v_{ed} \) stands for the preterit of a verb \( v \).

Table 1 shows the discovered typical activities of a place-name “bookstore”. The candidates with their weight than a certain threshold value (e.g., 0.01) are accepted as the typical activity expressions for the place.

<table>
<thead>
<tr>
<th>Candidate ( s )</th>
<th>activity(s)</th>
<th>( df(q_{p}^{at}) )</th>
<th>( df(q^{at}) )</th>
<th>weight( p(s) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>v:“buy”, n:“book”</td>
<td>&quot;bought * book&quot;</td>
<td>64</td>
<td>1507</td>
<td>0.257446</td>
</tr>
<tr>
<td>v:“see”, n:“book”</td>
<td>&quot;saw * book&quot;</td>
<td>26</td>
<td>572</td>
<td>0.129021</td>
</tr>
<tr>
<td>v:“pick up”, n:“book”</td>
<td>&quot;picked up * book&quot;</td>
<td>25</td>
<td>1095</td>
<td>0.056530</td>
</tr>
<tr>
<td>v:“work”, n:null</td>
<td>&quot;worked&quot;</td>
<td>437</td>
<td>275477</td>
<td>0.038360</td>
</tr>
<tr>
<td>v:“find”, n:“book”</td>
<td>&quot;found * book&quot;</td>
<td>15</td>
<td>855</td>
<td>0.027018</td>
</tr>
<tr>
<td>v:“get”, n:“book”</td>
<td>&quot;got * book&quot;</td>
<td>8</td>
<td>673</td>
<td>0.010122</td>
</tr>
<tr>
<td>v:“meet”, n:null</td>
<td>&quot;met&quot;</td>
<td>139</td>
<td>224058</td>
<td>0.004851</td>
</tr>
<tr>
<td>v:“spend”, n:“hour”</td>
<td>&quot;spent * hour&quot;</td>
<td>6</td>
<td>2102</td>
<td>0.001551</td>
</tr>
<tr>
<td>v:“drink”, n:“coffee”</td>
<td>&quot;drank * coffee&quot;</td>
<td>1</td>
<td>192</td>
<td>0.000686</td>
</tr>
<tr>
<td>v:“meet”, n:“girl”</td>
<td>&quot;met * girl&quot;</td>
<td>4</td>
<td>3501</td>
<td>0.000388</td>
</tr>
</tbody>
</table>

5. DISCOVERING ALTERNATIVE ACTIVITIES FROM THE WEB

A mobile user or AI robot sometimes searches the Web for making her/his decision on whether or not s/he should do an activity (behavior). However, s/he misses better activities than her/his original intending activity, because s/he is usually unaware of the alternative activities to her/his original intending activity.

[Google Sets, 2011] is a web-based tool for discovering comparable phrases of a user-given objective phrase. But we cannot condition the object by one of its various contexts. Meanwhile, [Ohshima, 2006] proposed a method for allowing us to do so. Their approach is based on the following two assumptions:

1. the conjunction “or” connects two comparable phrases.
2. if two phrases \( p_1 \) and \( p_2 \) are comparable, there exist both “\( p_1 \) or \( p_2 \)” and “\( p_2 \) or \( p_1 \)” pattern in the target corpus of documents such as the Web.

This section describes my method to discover the alternative activities to a given activity by mining the Web/Blog, which consists of the following two steps:

Step 1. Extracting Candidates for Alternative Activities

When an activity expression which is a set of components such as Action, Object and Place is given, my proposed Mobile DSS (Decision Support System) discovers the alternative phrases to each component \( p \) where the other components are regarded as its context \( c \). Firstly, the system retrieves the top \( k \) retrieval items, each of which includes its title and snippet (brief summary), by submitting
"p or" AND c as a query to [Google Web Search, 2011] where c is a keyword-based Boolean expression which concatenates the other components except p by AND operators. Next, a text which is between “p or” and its immediate following separator such as “,” “,” “?” “)” and “or” in the top k retrieval sets of title and snippet and whose number of words is less than certain threshold value, is regarded as a candidate for alternative phrases to the component p in the context c. Step 2. Weighting of Candidates for Alternative Activities
Secondly, the system assigns the following weight, to each candidate pi:

$$\text{weight}_{p,c}(p_i) = \min_{p,c}(p_i) + \left(1 - \frac{\min_{p,c}(p_i)}{\max_{p,c}(p_i)}\right)$$

$$\min_{p,c}(p_i) = \min\{\text{df}_{p,c}^A(p_i), \text{df}_{p,c}^B(p_i)\}$$

$$\max_{p,c}(p_i) = \min\{\text{df}_{p,c}^A(p_i), \text{df}_{p,c}^B(p_i)\}$$

$$\text{df}_{p,c}^A(p_i) = \text{df}(["p or p_i" \text{ AND } c])$$

$$\text{df}_{p,c}^B(p_i) = \text{df}(["p_i or p" \text{ AND } c])$$

where df(q) stands for the number of documents retrieved by submitting a query q to Google Web Search, and weight_{p,c}(p_i) = 0 if min_{p,c}(p_i) = 0. Finally, a new set of a candidate with higher weight and the other components is regarded as an expression of alternative activity to the given activity.

Table 2 shows the discovered alternative phrases of an objective phrase “Harry Potter” in various contexts. The candidates with their weight than a certain threshold value (e.g., 1.00) are accepted as the alternative activity expressions to the original activity.

When a mobile user or AI robot at a “bookstore” is intending to “buy” a “book” named by “Harry Potter” and thus inputs her/his intending activity expression,

[a:“buy” oc:“book” oi:“Harry Potter” p:“bookstore”]

to my proposed Mobile DSS, the system offers the following alternative activity expressions to her/him.

1. [a:“buy” oc:“book” oi:“Lord of the Rings” p:“bookstore”]
2. [a:“buy” oc:“book” oi:“Star Wars” p:“bookstore”]
3. [a:“buy” oc:“book” oi:“Da Vinci Code” p:“bookstore”]

Meanwhile, when s/he at a “theater” is intending to “buy” a “movie” named by “Harry Potter” and thus inputs her/his intending activity expression,

[a:“buy” oc:“movie” oi:“Harry Potter” p:“theater”]

to my proposed Mobile DSS, the system offers the following alternative activity expressions to her/him.

1. [a:“buy” oc:“movie” oi:“Lord of the Rings” p:“theater”]
2. [a:“buy” oc:“movie” oi:“Star Wars” p:“theater”]
### Table 2. Alternative Phrases of the Object Named by “Harry Potter”

<table>
<thead>
<tr>
<th>Context (c = {\text{“buy” AND “book” AND “Waldenbooks”}} )</th>
<th>Candidate (p_i)</th>
<th>(d^A_{P_c}(p_i))</th>
<th>(d^B_{P_c}(p_i))</th>
<th>weight(_{p_c}(p_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>“LOTR”</td>
<td>1</td>
<td>1</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>“Lord of the Rings”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“latest issue of Guns and Ammo”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“very popular items”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“anything”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“J.K. Rowling”</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“The Da Vinci Code”</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context (c = {\text{“buy” AND “book” AND “bookstore”}} )</th>
<th>Candidate (p_i)</th>
<th>(d^A_{P_c}(p_i))</th>
<th>(d^B_{P_c}(p_i))</th>
<th>weight(_{p_c}(p_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>“(The) Lord of the Rings” (7+)</td>
<td>31</td>
<td>39</td>
<td>38.03</td>
<td></td>
</tr>
<tr>
<td>“Star Wars”</td>
<td>11</td>
<td>15</td>
<td>11.27</td>
<td></td>
</tr>
<tr>
<td>“(The) Da Vinci Code” (11+)</td>
<td>5</td>
<td>10</td>
<td>10.38</td>
<td></td>
</tr>
<tr>
<td>“Lemony Snicket”</td>
<td>12</td>
<td>4</td>
<td>4.67</td>
<td></td>
</tr>
<tr>
<td>“Artemis Fowl”</td>
<td>4</td>
<td>5</td>
<td>4.20</td>
<td></td>
</tr>
<tr>
<td>“The Wizard of Oz”</td>
<td>5</td>
<td>2</td>
<td>2.60</td>
<td></td>
</tr>
<tr>
<td>“John Grisham”</td>
<td>2</td>
<td>3</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>“Nancy Drew”</td>
<td>22</td>
<td>1</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>“Dan Brown”</td>
<td>4</td>
<td>1</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>“none HP”</td>
<td>13</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“Pokemon”</td>
<td>7</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“Hogwart’s witches and wizards”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context (c = {\text{“buy” AND “movie” AND “theater”}} )</th>
<th>Candidate (p_i)</th>
<th>(d^A_{P_c}(p_i))</th>
<th>(d^B_{P_c}(p_i))</th>
<th>weight(_{p_c}(p_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>“(The) Lord of the Rings” (8+)</td>
<td>105</td>
<td>80</td>
<td>80.29</td>
<td></td>
</tr>
<tr>
<td>“Star Wars”</td>
<td>30</td>
<td>37</td>
<td>30.19</td>
<td></td>
</tr>
<tr>
<td>“James Bond”</td>
<td>5</td>
<td>1</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>“Frodo”</td>
<td>3</td>
<td>1</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>“(The) Chronicles of Narnia” (2+)</td>
<td>0</td>
<td>1</td>
<td>1.50</td>
<td></td>
</tr>
<tr>
<td>“Mission: Impossible”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“ROTK”</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>“White Chicks”</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

### 6. RELATED WORK

#### 6.1 Mobile Decision Support

Decision Support Systems (DSS) are generally a class of computer-based information systems or knowledge-based systems which support decision making activities in very different ways. With the recent advances in wireless and mobile computing environments, Mobile DSS have been also needed and thus developed for more urgent decision-making situations, such as Mobile Recommendation [Heijden, 2005] for consumers’ making decisions on purchasing in retail stores. Especially, various systems for mobile decision support in healthcare have been proposed [Fortier, 2003; Tsumoto, 2005]. For example, iTriage [Pedro, 2005; Padmanabhan, 2006] can help a clinician to identify the urgency of medical intervention when her/his patient presents with ambiguous triage case, based on a heuristic approach which selects the best triage category, identifies corresponding discriminating attribute of the patient and allows her/him to attach a level of confidence in the decision.
6.2 Alternative Query Generation

Query Refinement systems offer the Alternative Queries to a user’s original query for solving the so-called “mismatched query problem” in Information Retrieval that her/his information demand does not match the results retrieved by her/his original query which s/he has composed in order to express it to an IR system. For instance, Query Expansion [Xu, 1996; Cui, 2002] by inserting another retrieval condition for disambiguating a user’s original query, Query Relaxation [Jones, 2003; Mirzadeh, 2004] by deleting a part of its conditions for relaxing it, Query Substitutions [Terra, 2004; Jones, 2006] by replacing a part of its conditions with another condition for shifting it, or Query Recommendation [Baeza-Yates, 2004; Zhang, 2006] for offering its related (nearby-surrounding) queries to her/him. Meanwhile, Query Optimization systems transform a given query into its equivalent Alternative Queries for selecting the optimum query from among them based on their execution costs.

6.3 My Previous Work

My previous work [Hattori, 2006] has proposed a method to expand a mobile user’s original query by its related contextual words such as names of Action, Object or Place, based on the typical activities at her/his current geographic location in the real world, for enhancing context-awareness to the existing location-free IR (Information Retrieval) systems such as [Google Mobile, 2011]. This paper has proposed a Mobile DSS which discovers the alternative activities to a mobile user’s intending one from the Web and then automatically generates her/his activity-based query to the existing location-free IR systems, for supporting her/his decision-making process in her/his daily life.

7. CONCLUSION

This paper has proposed a Mobile DSS (Decision Support System) that allows a mobile user or AI robot to search the Web for the necessary information for making decision on her/his intending activity. When a mobile user or AI robot at a place firstly gives an activity expression while using the typical activities of the place as reference, my proposed system discovers its alternative activity expressions and offers them to the mobile user or AI robot’s database. When s/he secondly selects one of them, my system generates the activity-based query and searches the Web by submitting it to the existing location-free information retrieval systems such as mobile Web search engines. The system would allow a mobile user or AI robot to select a better activity from among the alternative activities to her/his original intending activity.

In the future, I plan to develop and evaluate a prototype system based on my proposed methods in more detail. And moreover, I challenge to invent a method for discovering the alternative activities dependent on not only Action, Object and Place but also Time or Subject, that is Temporized or Personalized.

REFERENCES

Conference paper or contributed volume


AN IMPROVED ENERGY EFFICIENT APPROACH FOR WSN BASED TRACKING APPLICATIONS

Tareq Alhmiedat¹ and Shuang-Hua Yang²
¹Zarka University - Zarka, Jordan
²Loughborough University - Loughborough, United Kingdom

ABSTRACT
Tracking systems using a high number of low cost sensor nodes have been proposed for use in diverse applications including civil, military, and wildlife monitoring applications. In tracking applications, each sensor node attempts to send the target’s location information to a sink node. Deploying a tracking system with a high number of sensor nodes results in the following limitations: high packet dropping rate, high congestion, transmission delay, and high power-consumption. Data aggregation schemes can reduce the number of messages transmitted over the network, while prediction schemes can decrease the number of activated beacon nodes in the tracking process. Consequently, data aggregation and prediction approaches can reduce the energy consumed during the tracking process. In this paper, we propose and implement an energy efficient approach for WSN-based tracking applications by integrating both a novel data aggregation method with a simple prediction approach. Three metrics are utilized for the evaluation purposes: total number of messages transmitted in the network, overall power-consumption, and the quality of the tracking accuracy. The proposed system is simulated using the NS2 simulation environment.

KEYWORDS
Tracking, Localization, ZigBee network, Data Aggregation, and Prediction.

1. INTRODUCTION
The node localization and tracking applications that have received much attention recently in the area Wireless Sensor Networks (WSNs) have focused on the need to achieve high localization accuracy without incurring a large cost, the form factor, and power consumption per node. The authors focused on localization and energy network standard issues in [1, 2] respectively. In this paper, we focus on the power consumption issue for tracking applications.

Tracking systems including numerous low-cost sensor devices have been used in many applications including those for the military, emergencies, wildlife monitoring, and border security. Each sensor node is equipped with one or more sensors for detecting physical characteristics, a battery as an energy source, a processor for performing computations, and a wireless transceiver for two-way communications. The communication range for each sensor node is limited and therefore, sending a message from a beacon node to the sink node normally results in a series of hops through the network. Each one of these hops produces a consumption of energy power, which is limited, and therefore results in failures within the network as soon as the sensor nodes run out of energy.

One of the most critical technical issues which has to be addressed in developing sensor networks for target tracking applications is the conservation of energy. Sensor nodes are usually supported by batteries which could be difficult to replace. A few existing researches [3, 4, 5] focused on the data aggregation concept for tracking applications using WSNs. Thus, in this paper, we study the problem of how to reduce the communication and power consumption for tracking applications by focusing on the research challenge related to the data aggregation and prediction techniques in WSNs for real-time surveillance applications. The idea of utilizing a data aggregation scheme with a prediction approach is to reduce the overheads of mobile computing systems.

Our contributions lie in the following features: 1) We present and compare several data aggregation and prediction schemes for target tracking sensor networks, and point out how a power-efficient tracking system
might be designed; 2) We propose an energy efficient approach for WSN-based tracking applications; 3) Our proposed system is validated through simulation experiments; 4) Trade-offs are investigated between tracking efficiency and network lifetime.

This paper is divided as follows: In Section 2, the related data aggregation and prediction methods are presented. In Section 3, several aggregation strategies are designed and implemented while Section 4 presents the simulation setup. The proposed aggregation strategies are evaluated in Section 5, and finally, Section 6 offers a conclusion and suggestions for future work.

2. RELATED WORK

Target tracking systems require a high number of messages to be transmitted over the network from beacon nodes to a sink node. In this section, we present various research works which have focused on data aggregation and prediction techniques and which are applicable to wireless sensor tracking systems. We divide the discussion into two approaches: an energy efficient approach and a tracking through sensor networks approach.

2.1 Energy Efficient Approach

Power consumption is a critical issue in designing a tracking application for WSNs as data aggregation and prediction methods can reduce the power consumption for the whole WSN. In this section, we briefly describe previous data aggregation and prediction methods.

Several existing researches focus on data aggregation methods, as the work presented in [3] which includes investigating four diverse aggregation strategies for tracking applications, concentrating on the trade-offs between the amount of communication in the network and tracking accuracy. The implemented strategies aim to reduce the total number of messages transmitted over the network while offering reasonable tracking accuracy by aggregating the localization information from beacon nodes.

The proposed work in [4] involves a data aggregation technique in a real-time surveillance application focusing on timeliness, power consumption and information availability. It was agreed that tracking solutions with data aggregation method can reduce the amounts of energy consumed. Three aggregation methods are presented in [5]: in-network, grid-based and hybrid schemes. The proposed hybrid scheme tries to combine features from both previous techniques, and also offers low time delay.

Prediction systems using a WSN are actively researched and addressed in several works. For example, the works proposed in [6, 7, 8] includes investigating several basic energy-saving methods for target tracking in sensor networks.

2.2 Tracking Approach

Target tracking and localization in sensor networks have been dynamically researched and studied in several works. For example, the works presented in [9, 10, 11, 12, 13, 14] focus on diverse aspects of tracking using sensor networks, such as tracking accuracy, real-time implementation, network standard, and the required application. The work presented in this paper is based on the binary detection method.

The work proposed in [15] includes a novel method for tracking the movement of people or vehicles in outdoor environments based on acoustic sensor devices. The presented work includes identifying each target with its unique acoustic signature, a feature sound pattern or a set of frequencies unique to that target. The work presented in [16] offers a binary sensor model, where each sensor’s value is converted constantly into one bit of information. Additionally, the solution proposed in [17] includes a tracking method for use in networks with binary proximity sensors. The presented method finds a straight line which estimates the path of a target during a short period of time, and uses a line to find out the target’s current position.
3. THE PROPOSED MODEL

The proposed system tries to optimize power consumption by minimizing both the sampling frequency rate and the number of nodes involved in target tracking, while offering reasonable localization accuracy. In this section, we first define the system requirements for target tracking applications and identify three performance metrics. Then, we investigate several data energy-saving strategies to meet these requirements. A system design is illustrated in detail next.

3.1 System Requirements and Metrics

In this section, we define the application requirements and identify numerous parameters needed for energy-saving schemes. The information of interest regarding the mobile targets includes: position \( p \), direction \( d \), and velocity \( v \). As soon as this information is found, only a small set of sensor nodes are activated while other sensor nodes stay in sleep mode.

**Requirements**: consider a network \( R = \{ b_1, b_2, ..., b_n \} \) with a total number of beacon nodes \( n \), and \( M = \{ m_1, m_2, ..., m_z \} \) with a total number of mobile targets \( z \). \( nb \) represents the total number of observer nodes which sense the mobile target \( m \). \( ns = n - nb \) is the total number of beacon nodes in sleep mode. The application requires the sensor network to report the mobile target’s location to the sink node with low power consumption and reasonable positioning accuracy.

**Problem Definition**: knowing the requirements for the mobile tracking application, we need to develop energy-saving strategies in order to minimize the overall power consumption while maintaining realistic tracking accuracy.

**Performance Metrics**: three performance metrics have to be taken into consideration: first, the total amount of messages transmitted over the network; second, the overall power-consumption needed by the tracking application; and third, the tracking accuracy for mobile targets through a distributed WSN.

3.2 Energy-saving Strategies

The main goal of this paper is to design and implement different energy saving approaches in order to reduce the total number of messages which need to be transmitted over WSNs, while achieving reasonable tracking accuracy.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ob )</td>
<td>Observer node</td>
<td>A beacon node which senses any mobile target</td>
</tr>
<tr>
<td>( s )</td>
<td>Sink node</td>
<td>A node which need to be informed about the targets’ positions</td>
</tr>
<tr>
<td>( sf )</td>
<td>Sampling frequency rate</td>
<td>How often the sink node has to be informed about the targets’ position</td>
</tr>
<tr>
<td>( l_k )</td>
<td>Leader node</td>
<td>A node which has the role of aggregating the collected localization information</td>
</tr>
<tr>
<td>( v )</td>
<td>Velocity</td>
<td>The mobile target’s speed</td>
</tr>
<tr>
<td>( p )</td>
<td>Power consumption</td>
<td>The total power consumed over a period of time</td>
</tr>
</tbody>
</table>

In this section, we propose 4 energy efficient strategies. The first three strategies are aggregation-based, while the latest strategy includes a simple energy-efficient prediction system integrated with an aggregation approach in order to reduce the power consumption for the whole WSN. Table 1 includes definitions for several parameters.

**Strategy One (Naive)**: All beacon nodes in the network have to be in active mode in order to sense the mobile target. Beacon nodes which sense the presence of the mobile target \( m_j \) broadcast \( \{ p_i, q_i(t) \} \) to the
sink node, where \( p_i \) is the position of beacon node \( i \), and \( q_i \) is the quality of the observation at time \( t \). There is no local processing or data aggregation in this strategy. The sampling frequency rate is \( (sf = 2) \).

**Strategy Two (Differential-based):** Like the previous strategy, all beacon nodes stay in active mode. Once the mobile target \( m_j \) enters the beacon node’s detection range \( b_i \), \( b_i \) then reports the observation to the sink node. From this time, it is assumed by the sink node that the beacon node \( b_i \) can observe the mobile target \( m_j \). As soon as the mobile target \( m_j \) leaves the detection range for sensor \( b_i \), then \( b_i \) reports this to the sink node. The observer node \( b_i \) broadcasts \( \{p_i, q_i(t), \Delta t\} \), where \( \Delta t \) is the total duration time for tracking the mobile target \( m_j \) by the beacon node \( b_i \). In this strategy, the sink node is responsible for computing the location of each mobile target. The sf is presented in Equation 1. The same idea for this strategy is proposed in [3].

\[
sf = \frac{tr}{v} \quad (1)
\]

where \( tr \), \( v \) are the transmission range, and velocity respectively.

**Strategy Three (Leader-based):** In this strategy, a leader node has to be elected each time in order to aggregate the relevant location information and transmit the latest position information to the sink node. Unlike in the previous strategy, as soon as any mobile target \( m_j \) enters the sensor’s range \( b_i \), \( b_i \) reports this to the leader node instead of the sink node. The leader node \( l_k \) aggregates the position information collected from the beacon nodes \( \{b_i, b_{i+1}, ..., b_w\} \), where \( w \) is the total number of observer nodes which cover the mobile target \( m_j \), and informs the sink node. As soon as the mobile target \( m_j \) moves away from any beacon nodes’ range \( b_i \), then \( b_i \) will report this to the leader node. The leader node \( l_k \) will then re-compute the target’s position and report that to the sink node. The sampling frequency rate is presented as follows:

\[
sf = \frac{tr}{v} - agt \quad (2)
\]

where \( agt \) is the time needed to collect the localization information from observer nodes.

**Strategy Four (Prediction-based):** This strategy includes the aggregation strategy 3 with a simple prediction system in order to reduce the power consumption of the whole network. Most of the energy is consumed during the idle period waiting for possible targets. Beacon nodes should spend a minimal time in the active mode and stay in sleep mode as long as possible. Therefore, a large amount of energy is saved if the system does not require all the beacon nodes to be activated for the whole tracking period.

The prediction system works by computing the velocity and direction of each mobile target, based on the localization information collected from the observer node. The leader node has to make these computations and then activate only the essential beacon nodes required to track the mobile targets; other beacon nodes stay in sleep mode.

### 3.3 System Design

The main stages are presented in this section for the 4th strategy as it offers the lowest power-consumption, as will be shown in the evaluation section.

1. **Initial Activation:** We divide the beacon nodes, based on the activating mode, into two main types: sentry and non-sentry nodes. Sentry nodes include a small number of beacon nodes which have to stay active all the time in order to sense any approaching mobile target. Non-sentry nodes are end-device nodes which stay in sleep mode until they are woken up by a leader node. In this phase, all beacon nodes stay in the sleep mode when there are no targets to be tracked in order to save power.

2. **Target Detection:** Unlike localization sensors which are used in traditional tracking systems, binary systems use binary sensors which offer 1-bit of data, representing the presence or absence of the mobile
target in the sensing range. Binary sensors are unable to produce any other information, such as distance to the beacon node or direction of arrival.

The mobile target position $q$ is estimated through the binary sensor as $q \in \{0, 1\}$, where $q = 0$ when the target is out of the beacon node’s detection range, and $q = 1$ when the mobile target is within the range of the beacon node. A centroid method is used to find the final position for each mobile target.

3. Group Aggregation: The proposed system is designed to work with a ZigBee network standard. Therefore, every end-device node in each group belongs to one, and only one, router node. The proposed model includes dividing the wireless sensor area into small groups. Each group includes a number of end-devices and one or more router nodes; the group is represented by a single leader node which has the responsibility of aggregating the localization reports from observer nodes, and transmitting the target’s position to the sink node.

Each group can be in either the active or the inactive state at a specific point in time. The active group includes activating all the beacon nodes in that group, while the inactive state includes deactivating the beacon nodes in that group.

4. Group Leader to Sink Report: After the group aggregation, the leader node aggregates the collected readings from end-device nodes and reports a single message to the sink node. The aggregation function is typically simple as it needs only to find out the average of the mobile target’s position based on the location information collected from observer nodes.

The sink node is responsible for computing the mobile target’s trajectory. A Kalman filter is used to establish the mobile targets’ trajectories. This is a recursive filter that assesses the state of a dynamic system from a series of noisy measurements. The Kalman filter takes advantage of the dynamics of the target to remove the consequences of noise and to offer a good estimation.

5. Prediction and Activation: Tracking based on a single beacon node offers low localization accuracy; a group of sensors have to be collaborated in order to track and locate the target’s position. End-device nodes need to be activated as soon as one of the group’s members detects a mobile target. Before activating a set of beacon nodes, a prediction method has to be applied first in order to predict the future movement of the mobile target and then activate the right set of beacon nodes. This stage includes three main phases: A) Prediction mechanism, B) Activation mechanism, C) Recovery mechanism.

4. IMPLEMENTATION

In the previous section, we outlined the design and development of an energy-efficient WSN-based tracking system. Simulation experiments were carried out in order to evaluate the proposed strategies. This section presents the simulation environment, and the network topology. The proposed model was evaluated using an NS2 simulator.

4.1 Network Topology

Our simulation model includes 145 beacon sensor devices in a $120 \times 120 m^2$ monitored area. The sensing coverage for end-device nodes is 15 meters and is 30 meters for router device nodes.

The proposed tracking system is implemented using ZigBee network standard. ZigBee is a low-power, low-data rate, and low-cost wireless communication standard which aims to be used in home automation and remote control applications. The ZigBee network includes three main roles: coordinator, router and end-device nodes. The coordinator has the role of starting and controlling the network. The router is responsible for routing messages between nodes and provides backup in case of network congestion or device failure. The end-device has the ability to transfer and receive messages.

In our simulation, we used these three components. The coordinator node works as a sink node to collect information from router nodes while router nodes have the responsibility for routing messages between nodes; in addition, each router node might work as a leader for a period of time. End-device nodes collect the localization information about the mobile target and transmit their readings to the router node. Each end-device node can talk to only one router within on hop. ZigBee network uses the Ad-hoc On Demand Distance Vector (AODV) routing protocol.
5. **SYSTEM EVALUATION**

In order to produce a long-term tracking system that meets the requirements of several applications, power-efficiency is both essential and critical. In this section, we evaluate the total number of messages, power-consumption and tracking accuracy.

### 5.1 Total Number of Messages

In this section, the total number of messages is evaluated for the first three strategies. In the first strategy, the observer nodes transmit a single message to the sink node every 2 seconds. The total number of messages which need to be transmitted over the network is presented in Equation 5. The sampling frequency rate is quite high, as the sink node needs to be informed about the mobile target’s location every 2 seconds, hence \( sf = 2 \).

Fewer messages are transmitted in the second strategy as the observer nodes have to transmit their readings only when the mobile target enters and leaves their detection range. The total number of messages is presented in the following equation with sampling frequency rate \( sf = \frac{tr}{v} \).

\[
\text{nmm}_{\text{loc}} = \frac{\text{n}\times \text{f}}{\text{tr}} \quad (5)
\]

where \( n \) and \( f \) are the number of observer nodes and the final time respectively.

In the third and fourth strategies, each observer node transmits its localization report to a leader node and then a leader node aggregates the localization information from observer nodes and transmits a single report to the sink node. Two types of message need to be transmitted: local and global. Local messages include the total number of messages \( \text{nmm}_{\text{loc}} \) which need to be transmitted among observer and leader nodes, as presented in Equation 6. These messages do not reach the sink node as they are required to be transmitted to the leader node. Global messages include the total number of messages \( \text{nmm}_{\text{glb}} \) which reaches the sink node from leader nodes, based on the AODV routing protocol. For both equations, the sampling frequency rate is \( sf = \frac{tr}{v} \). Figure 1 presents the total number of messages which needs to be transmitted for each strategy.

\[
\text{nmm}_{\text{glb}} = \frac{\text{f}}{\text{tr}} \quad (6)
\]

### 5.2 Power Consumption

In the previous section, we evaluated the total number of messages transmitted over the network. Each transmitted packet requires a specific amount of energy; therefore, reducing the total amount of messages transmitted over the network will reduce the total power consumption. In this section, we evaluate the total energy consumed over a period of time for each strategy.

The first strategy offers the worst power-efficient solution, as the sampling frequency rate is high, while the second strategy achieves a reasonably power-efficient system, as the sampling frequency rate is lower than the first strategy. The power consumption \( E \) in both strategies is represented in the following equation:

\[
E = p + ((n - nb) \times p_{\text{act}}) + \left(\text{nm} \times p_{\text{tx}}\right) + \left(\text{nm} \times p_{\text{rx}}\right) \times h \quad (7)
\]

\[
p = \frac{\text{nm} \times p_{\text{tx}}}{h} \quad (8)
\]
where \( p, p_{rx}, \) and \( p_{tx} \) are the power needed to transmit a number of messages \( nm \), the transmission power, and the reception power respectively.

All the position information is transmitted to a leader node in the third strategy; the leader node aggregates the collected information and transmits a single packet to the sink node. The power consumption in the third strategy is based on the target speed and is represented in the following equation:

\[
P = p_{nm} + (n - nb) \times p_{act}
\]

(9)

where \( p_{act} \) is the power needed for sensor nodes in active mode for every 1 time unit.

The first 3 strategies show that most of the energy is consumed while the beacon nodes are in the active mode. The third aggregation strategy reduces the power consumption for the observer nodes but a high level of energy is still consumed during the waiting process. The fourth strategy includes minimizing the power consumption by deactivating beacon nodes that are away from the mobile target.

In the fourth strategy, only the beacon nodes which sense the mobile target are in the active mode; other beacon nodes have to stay in sleep mode. The power consumption is represented in the following equation:

\[
P = p_{nm} + (ns \times p_{act}) + (n - nb) \times p_{dm}
\]

(10)

where \( p_{dm} \) is the power needed for sensor nodes in sleep model for every 1 time unit.

Figure 3 shows the power consumed for each strategy within 3 minutes of tracking time.

5.3 Tracking Accuracy

The accuracy of each strategy is evaluated in this section. Figure 4 presents the tracking trajectory (dotted line) for the third strategy.

6. CONCLUSION

This paper shows the results of a study concerning tracking, through WSNs, multiple mobile targets simultaneously travelling at different speeds, concentrating on the trade-offs between the amount of messages which need to be transmitted to the sink node, power consumption and tracking accuracy. In this paper, we have proposed, implemented and evaluated several energy-efficient schemes for tracking applications. The data aggregation and prediction methods offer power-efficient solutions for sensor networks and achieve good tracking accuracy.

The proposed approach in this paper integrates both the data aggregation and prediction approaches for wireless sensor tracking applications. A novel aggregation strategy is proposed in Strategy 3 and integrated with a simple prediction system in order to reduce the power consumed during the tracking process. Our future work includes implementing the proposed system on real sensor networks and testing the efficiency of tracking multiple mobile targets through a WSN while achieving low power consumption.
REFERENCES


ABSTRACT

The intrinsic property of a MIMC (multi-interface multi-channel) network makes the routing in wireless ad hoc networks more diverse. Traditional ad hoc routing protocols can not make good use of the potential introduced by MIMC. In this paper, a cost-aware multi path routing protocol for MIMC ad hoc networks is proposed. Path cost is measured by link load and interference. The costs spread along the forward direction of RREQ and RREP messages. Then by improving the disjoint path criteria and reverse/forward path process, a RREQ/RREP Waiting Mechanism is proposed. A flow based data forwarding procedure is designed to decrease the interference between different flows. We applied this protocol to some network scenarios. Simulation results show the new protocol can significantly improve flow’s throughput.

KEYWORDS

Ad hoc network, Multi interface multi channel, multi path routing protocol.

1. INTRODUCTION

A mobile ad hoc network (MANET) does not rely on any existing infrastructure, it is very useful in battlefield, emergency and disaster relief where communication network infrastructure may not exit or hard to be deployed. Multi-interface multi-channel (MIMC) wireless networks recently have drawn considerable research attention, in which a node may be equipped with multiple network interfaces operated at different channels. An MIMC MANET can use more spectrum resources, which results in improvement of QoS in the network. On the other hand, MIMC device can provide more robust connection in the network than a single interface single channel (SISC) device.

Various proactive and reactive routing protocols have been proposed for SISC ad hoc networks, such as DSDV, DSR, AODV, and OLSR. These traditional ad hoc routing protocols can’t make full use of the interface and channel resources to improve the QoS of the network when used in MIMC circumstance.

Some researches have been carried on MIMC routing. Gong et al.[5] presented a combined channel assignment and AODV scheme, called Channel Assignment AODV (CA-AODV).But this protocol need a separated control channel. Politecnico et al.[6] proposed optimization framework includes routing, scheduling and channel assignment for MIMC mesh networks. However, this method is TDMA based, and needs a centralized schedule. This scheme is not applicable to distributed and self-organized ad hoc networks. Raniwala et al.[7] present a distributed routing/channel assignment algorithm. This protocol only focused on solving the routing between wireless router and wire networks based on spanning tree. Draves et al[8] implemented a routing protocol MR-LQSR. A path metric called Weighted Cumulative ETT (WCETT) that explicitly accounts for the interference among links is proposed. The high-throughput path between a source and a destination will be chosen. Anguswamy et al. [9] proposed a MIMC routing protocol MMC that considers various QoS parameters such as throughput, end-to-end delay, and energy utilization as a single unified cost metric and identifies the route that optimizes the cost metric and balances the traffic among the channels on a per flow basis. Cheolgi Kim et al.[10] present a link-state routing protocol tailored for
multichannel networks by minimizing the broadcast overheads. Most of these protocols try to discover one best route, but in MIMC network the routes for a data flow are more diverse.

Multi path is a good extension to improve network’s performance with back up paths. AODVM is a multipath version of AODV proposed by Ye[12] to minimum overhead in multipath routing. Mueller[13] made an improvement of AODV protocol with AODVM/PD to reduce the interference among different paths.Jing et al.[14] made a improvement of AODVM and proposed path selection protocol AODVM-PSP to improve the overall delay performance. Huang and Fang [15] proposed prioritize multipath routing PRIMAR, multiple paths are selected depending upon the type of traffic. AOMDV is another multipath routing protocol using disjoint path proposed by Mahesh[16]. AOMDV uses next hop and last hop to distinguish whether two paths are link disjoined. Since the RREQ is flooded network-wide, a node may receive several copies of the same RREQ. Thus, all duplicate copies are examined in AOMDV for potential alternate reverse paths.

There are many researches on MIMC routing and multi-path routing separately, but there is few research about multi-path routing for MIMC MANET. Multi-path routing in MIMC environment can improve the network’s performance significantly. In this paper, we propose cost-aware MIMC multipath routing protocol based on AOMDV called MIMC-AOMDV. Path cost is measured by link load and interference. The cost will spread along the forward direction of RREQ and RREP messages, every path can get its traffic cost to the destination. By improved the disjoint path criteria and reverse/forward path process, the intermediate node can discover more feasible paths. A new path selection and flow based data forwarding procedure is designed to decrease the interference between different flows.

The remainder of this paper continues as follows: In section 2 path cost estimate methods is proposed; Section 3 gives the details of the cost-aware MIMC multi-path routing protocol. It is then evaluated in Section 4. Finally, Section 5 summarizes our main conclusions.

2. PATH COST ESTIMATION

2.1 Interference Measure

The ad hoc network can be modeled as a directed graph \( G = (V, E) \). Each node \( v \in V \) correspond to individual wireless node, \( I(v) \) denotes the number of wireless interfaces that it has. A wireless interface of node \( v \) operates on a single channel selected from set \( \xi \). We assume there are \( K \) orthogonal channels in \( \xi \) numbered from 1 to \( K \). The interfaces of a node are assigned with distinct channels, so each wireless node \( v \) can be associated with a ordered set \( F(v) \subseteq \xi \) of \( I(v) \) distinct channels, where the \( i \)-th interface of node \( v \) operates on the \( i \)-th channel in \( F(v) \).

Let \( R_i \) denote the transmission range and \( d(u,v) \) denote the distance between the nodes \( u \) and \( v \). A link \( (u,v) \in E \) if and only if \( d(u,v) \leq R_j \), which implies that nodes \( u \) and \( v \) can communicate directly. However, such a communication is only possible if there is a common channel among \( F(u) \) and \( F(v) \). \( R_i \) denotes the interference range. We assume that the \( R_j \) is \( q \times R_j \), where \( q \geq 1 \). We use \( ch_e \) to denote the channel used for edge \( e \). \( I(e) \in E \) denotes the set of edges which are in the interference range with edge \( e \in E \). When there is data transmission through link \( e \) on channel \( ch_e \), all the links \( e' \in I(e) \) cannot carry out data transmission on the same channel at the same time.

Denote the CDC (Channel Diversity Cost) as the number of edges under interference with link \( e \) on channel \( ch_e \), we can get the CDC of edge \( e \) by:

\[
CDC(e) = \sum_{e' \in I(e)} d(ch_e, ch_{e'})
\] (1)
\[ d(ch_s, ch_r) = \begin{cases} 1, & ch_s = ch_r \\ 0, & ch_s \neq ch_r \end{cases} \quad (2) \]

As one node may not know the channel usage of surrounding nodes, a Channel Usage Message should be generated and broadcasted periodically to its neighbors. A node should know the channel usage within its two hop nodes. The Channel Usage Message should include the channels used by this node and its one hop neighbors. As one path is composed of many links from source to destination, the CDC of one path can be achieved by collecting the CDC of each link in that path. We denote the CDC of path \( p \) as following:

\[ CDC(p) = \max_{e \in p} CDC(e) \quad (3) \]

In AODV based or extended routing protocol, a route is constructed by RREQ and RREP message. So the interference information should be included in the RREQ and RREP message, letting the CDC of the path spread away. The \( CDC(RREQ_p) \) denotes the Channel Diversity Cost of RREQ message along reverse path \( p \); \( CDC(RREP_p) \) denotes the Channel Diversity Cost of RREP message along forward path \( p \); \( ch_{send} \) denotes the channel used by the interface to send the RREQ or RREP message. \( CDC(RREQ_p) \) and \( CDC(RREP_p) \) should be attached to RREQ or RREP message separately.

\[ CDC(RREQ_p) = \max(CDC(ch_{send}), CDC(p_{reverse})) \quad (4) \]
\[ CDC(RREP_p) = \max(CDC(ch_{send}), CDC(p_{forward})) \quad (5) \]

When node \( i \) receives RREQ or RREP message, it will update the CDC of the reverse path or forward path as follows. The \( ch_{recv} \) denotes the channel used by the interface receiving the RREQ or RREP message.

\[ CDC(p_{reverse}) = \max(CDC(ch_{recv}), CDC(RREQ_p)) \quad (6) \]
\[ CDC(p_{forward}) = \max(CDC(ch_{recv}), CDC(RREP_p)) \quad (7) \]

### 2.2 Load Measure

In ad hoc networks, there may exit some nodes who forward many flows. These nodes can be referred as hotspots, which may lead to packets congestion for the intensive transmission competition of different flows. So the routing protocol should try to balance the route of different flows to avoid the hotspots.

Here we use the busyness rate of MAC layer to indicate the interface workload, which is more direct and can take the affect of overhearing into account. Let \( t_i \) denote the sample interval, \( N \) denote the total sample number in a statistics period, and \( n_b \) denote the number of sampling points where the interface’s MAC is busy(in sending or receiving). Then we can get:

\[ Load(if_i) = \frac{n_b}{N} \quad (8) \]

Load of an interface can’t be obtained without cross-layer design in MAC layer. MAC layer should provide an interface for IP layer so that IP layer can read load for the interface. Furthermore, the load of a path can be achieved, which is defined as follow, \( if_i \) denotes the \( i - \) th interface in a path \( p \).

\[ Load(p) = \max_{i \in p} \{ Load(if_i) \} \quad (9) \]

Same as the interference information, the load information should be included in the RREQ and RREP message to be distributed. We denote \( Load(RREQ_p) \) as the load parameter at RREQ message along the reverse path \( p \). \( Load(RREP_p) \) is the load of RREP message along the forward path \( p \). They can be defined as follows respectively:

\[ Load(RREQ_p) = \max\{ Load(p_{reverse}), Load(if_{send}) \} \quad (10) \]
\[ Load(RREP_p) = \max\{ Load(p_{forward}), Load(if_{send}) \} \quad (11) \]

\( Load(p_{reverse}) \) is the load of the reverse path, \( Load(p_{forward}) \) is the load of forward path. \( Load(if_{send}) \) denotes the load of the interface sending this RREQ or RREP message. When a node \( i \) receives the RREQ or RREP message, it will update the load of the reverse path or forward path as follows.

\[ Load(p_{reverse}) = \max\{ Load(if_{recv}), Load(RREQ_p) \} \quad (12) \]
3. COST-AWARE MIMC-AOMDV PROTOCOL

In order to make use of the frequency and interface resources of a MIMC network, the MIMC multi path route protocol should try to discover and record feasible paths under transfer delay constraint, then the workload can be balanced by choosing proper path according to its load. Fig.2 shows the route table structure of the MIMC-AOMDV protocol. The interface and load for every path must be specified.

3.1 Disjoint Path Criteria

AODV protocol uses next hop and last hop to distinguish whether two paths are link disjoined. If two paths from a node $s$ to a destination $d$ are link disjoint, then they must have unique next hops as well as unique last hops.

$$\text{Disjoint}(p_1, p_2) = \begin{cases} 1, & \text{if} \ (\text{next hop}_{p_1} = \text{next hop}_{p_2}) \cup (\text{last hop}_{p_1} = \text{last hop}_{p_2}) \cup (if_{p_1} = if_{p_2}) \\ 0, & \text{otherwise} \end{cases}$$  \hfill (14)

3.2 RREQ/RREP Waiting Mechanism

In AODV protocol, the intermediate node only forwards RREQ with same sequence number and destination once. The first received RREQ message is forwarded. But for MIMC network, intermediate node may have several reverse paths. REEQ message from different paths may arrive at different time, so the original RREQ forwarding method can’t reflect the workload of the reverse paths nicely. We put forward a RREQ Forward Waiting Mechanism to make load information in the RREQ message more comprehensive.
(1) When the node receives the first RREQ message, the node will check its route table to see if it has the route to the destination. If it has, it replies with RREP to the source node along the reverse route;
(2) If the node doesn’t have the route, it will check the load included in the RREQ message;
(3) If the load is below threshold $\text{Load}_{\text{thresh}}$, it means the load of reverse path is light, and the node can forward the RREQ immediately;
(4) If the load is above threshold $\text{Load}_{\text{thresh}}$, which means the reverse path is quite busy, the node will put this path into candidate path set, then waits for a random period of time $t_{\text{wait}} \in [t_{\text{min}}, t_{\text{max}}]$, to see if some other reverse paths could be discovered.
(5) When the waiting timer is expired, the node will select one reverse path from the candidate path set, and its load will be recorded in RREQ message and be distributed.

As for the RREP message, similar process is applied.

### 3.3 Path Selection

This process provides path information for RREQ/RREP when multi candidate paths exist. The best path will be the one with shortest hops, minimum workload and minimum interference. If one path satisfies the best path criteria, then this path will be selected. However, sometimes no path can satisfy all properties. We accord the following procedure to select the proper path. Suppose $\text{Hop}_{r_{\text{min}}}$ denotes the minimum hop number of all the paths belong to route $r$.

1. Paths with hops less than $\text{Hop}_{r_{\text{min}}} + \Delta$ is considered, where $\Delta$ is the max permissible hop difference.
2. Find out all the paths whose load below $\text{Load}_{\text{thresh}}$, probably select a path according to load of each path;
3. If there is no path satisfying the load threshold $\text{Load}_{\text{thresh}}$ constraint, then use probability selection according to the interference of each path.

According to the decision above, the workload is considered as higher priority. If the workload threshold can’t be satisfied, the interference is considered.

### 3.4 Path Discovering

Every intermediate node only forwards the RREQ once. But the intermediate node may receive several RREQ messages from different interfaces. The entire disjoint path discovered by RREQ massage will be reserved in the route table.

Destination node will reply to every RREQ message with RREP message. The intermediate node will forward RREP along all the reverse paths in its route table. In SISC network, the intermediate may not receive RREP over once, because a RREP is sent in unicast mode. But in MIMC network, an intermediate node may receive multiple RREPs to the same source node from different interfaces. Only the RREP indicating a better forward path will be forwarded to update the route information.

### 3.5 Flow based Data Forwarding

As one node may have several paths for one route, it must decide which path is selected to send or forward a packet. Here we propose a flow based stochastic path selection method to solve this problem. A flow $f_{s,d}^i$ expresses data from source $s$ to destination $d$ at node $i$. The procedure is presented as follows.

1. Every node keeps a flow table to save all the flows passing through this node and its output path. The flow item in the flow table has lifetime. Every time the node receives data of a flow, it will update that flow’s expiration time. When a flow item is expired, it will be deleted from the flow table;
2. When a node will send or forward packets from $s$ to $d$, it first extracts the flow information $f_{s,d}^i$;
3. The MIMC-AOMDV protocol will check if $f_{s,d}^i$ has existed in the flow table. If it does, then the protocol directly gets the flow’s output path and sends the packet through that path;
(4) If the flow is new, the protocol will allocate a new item. Let \( \text{Load}(f_i) \) denote the current workload of output interface for path \( p_i \), in route \( r_{s,d} \), and \( \text{Load}(p_i) \) be the workload of \( p_i \). The path selected will be:

\[
p^* = \min_{p \in \mathcal{R}_{s,d}} \max(\text{Load}(f_i), \text{Load}(p_i))
\]

(5) The path information for flow \( f_{s,d} \) will be filled in to the flow table.

According to the path selection procedure, the old flows will keep stable by following existing output paths. The new flows will select its output interface combining path load and current interface load. Because the path load kept in the route table may not reflect current load of the network. A totally update of all the paths load is not feasible, so putting the current interface’s workload into consider can be effective and convenient.

All the other paths belonging to a route will serve as backup paths. If the protocol perceives a link broken of current transmission, it will try to find an alternative path to resume data transferring by path selection procedure again. This will improve the End to End delay by avoiding a new route discovering procedure.

4. PERFORMANCE EVALUATION

The Monarch research group in CMU developed complete physical, data link, and MAC layer models on ns2 for supporting ad hoc networks simulation. So ns2 is a perfect platform to realize this new protocol. The original ns2 platform doesn’t support MIMC wireless network, so we first finished the ns2 MIMC extension. Then we realized the MIMC-AOMDV protocol. The simulation parameters are listed in table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network size</td>
<td>1000m x 500m</td>
</tr>
<tr>
<td>Simulation time</td>
<td>30 sec</td>
</tr>
<tr>
<td>Node transmission range</td>
<td>250 m</td>
</tr>
<tr>
<td>Channel capacity</td>
<td>2 Mbps</td>
</tr>
<tr>
<td>CBR interval</td>
<td>0.008 s</td>
</tr>
<tr>
<td>Total channels available</td>
<td>5</td>
</tr>
<tr>
<td>Interfaces per node</td>
<td>3</td>
</tr>
</tbody>
</table>

Here we focus on the routing protocol evaluation; we won’t consider the channel assignment in the MIMC network. We assign every interface a channel in advance, and the channel will not change in the simulation. We use 2 scenarios to show the performance improvement by the new protocol.

1) Flow throughput improvement

Fig. 4 shows a scenario of two adjacent flow transferring. The channels used for every node are listed aside. Flow \( \text{flow}_{s,d} \) from B to E starts at 4s, and another flow \( \text{flow}_{e,d} \) from F to H starts at 9s. The routing protocol will find paths for these two flows. The channels with dark grey background are used for sending, the channels with light grey background is for receiving. And the channels with slash back ground are affected by surrounding transmission.
Fig. 5 shows the performance of different routing protocols. The single channel AODV has the lowest flow throughput. After the introducing of flow $flow_{F,H}$, the throughput of $flow_{B,E}$ decreases dramatically. MIMC-AODV protocol doesn’t consider the load, the second flow $flow_{F,H}$ may choose affected channel at node F, which will disturb the transmission of $flow_{B,E}$. The MIMC-AOMDV will select a more lightly loaded path, so the two flows can all reach its full capacity without interference.

2) Load balance

Fig. 6 shows a scenario for load balance. There is a flow $flow_{B,D}$ from B to D starting at 4s, and another flow $flow_{A,D}$ from A to D starting at 9s. This two flows converge at node B. Node B has two paths to node D, the first is $B \rightarrow D$, the second is $B \rightarrow C \rightarrow D$. When $flow_{B,D}$ starts, MIMC-AOMDV protocol selects path $B \rightarrow D$ as its forwarding path, channel 3 as its transmission channel. When $flow_{A,D}$ starts, and packet of $flow_{A,D}$ reaches node B. MIMC-AOMDV protocol will choose a lighter load path $B \rightarrow C \rightarrow D$ as $flow_{A,D}$’s forwarding path.

Fig. 7 shows the simulation result of different routing protocols for load balance scenario. As for single channel AODV protocol, the later started flow $flow_{A,D}$ can’t even reach its destination for the heavy load in node B. At the same time the first started flow $flow_{B,D}$ is disturbed when $flow_{A,D}$ trying to compete the channel with it. The performance of MIMC-AODV is much better. The channels used for $A \rightarrow B$ and $B \rightarrow D$ are different, the transmission of these two parts are out of interference. But at node B, these two flows share the same link $B \rightarrow D$. MIMC-AOMDV can achieve the best result. At node B, the new protocol will choose a more lightly loaded path, so the transmission of $flow_{A,D}$ and $flow_{B,D}$ carry on independently.
5. CONCLUSION

In this paper, we put forward a load balanced MIMC multi path routing protocol MIMC-AOMDV. We take some measures to realize the load balance: first, interference and path load estimate are introduced into route discovering, these information is attached to RREQ and RREP message; then we improve the disjoint path criteria and reverse/forward path finding of the protocol, propose RREQ/RREP waiting mechanism, path selection and flow based data forwarding procedure. The protocol is realized with ns2 network simulation tool. We designed some scenarios to demonstrate the improvement of the new protocol. The result shows the new protocol can effectively improve the throughput of data flows.

In MIMC ad hoc network, channel assignment is another important problem need to be solved. The routing protocol presented in this paper is based on the result of prearranged channel assignment. In actual MIMC ad hoc network, every node should have the ability to adjust its channel and keep the network’s connectivity. We will research the channel assignment combined with multi path routing protocol in the future.

REFERENCES

A COMPLETE AND EFFICIENT STRATEGY FOR COOPERATIVE CACHE IN MOBILE INFORMATION SYSTEM

Thu Tran Minh Nguyen and Thuy Thi Bich Dong

IS Department, Faculty of IT, University of Science, VNU, HCMC - 227 Nguyen Van Cu Street, Dist 5, HCM City

ABSTRACT

Cooperative data caching strategies become more and more attractive in mobile environment. Mobile clients cache frequently accessed data items to share and coordinate among these items in the network. These strategies save communication bandwidth, and resource usage as well as data latency and data accessibility. In this paper, we propose a complete and efficient strategy for cooperative caching in mobile information systems. There are four problems which have been solved in our strategy are cache discovery, cache placement, cache replacement, and cache consistency. We evaluate and demonstrate the experimental results on datasets by using NS2 simulator. We also prove the effectiveness by comparing it with the other existing approaches. The experimental results present our proposed strategy is more effective than other approaches.

KEYWORDS

Mobile information systems, cooperative caching, cache discovery, cache replacement, cache consistency, mobile P2P network.

1. INTRODUCTION

The recent development of wireless communications and mobile devices (mobile phones, pocket PCs, laptops) have paved the road for information access at anytime, anywhere, and in any form. Wireless applications are becoming more and more popular and useful in our life. However, the development of this application type has met a lot of disadvantages due to limited battery power, scarce wireless bandwidth, and inevitable mobile.

Cooperative caching has been studied and it has become one of the most exciting technique in mobile environments. Cooperative caching allows sharing and coordinating cached data items, which are stored in mobile clients’ local cache to achieve a better performance. This technique has got lower mobile host communication overhead and energy consumption as well as reduce query latency. However, it has brought up new challenges for researching on routing, resource discovery, data retrieval, and data consistency control, security and privacy management. In this paper, we will propose a cooperative caching in mobile information system. Our proposed model addresses all main problems including cache discovery, cache admission, cache replacement and cache consistency.

A variety of popular cooperative caching strategies have been studied such as: Cach Data, Cache Path, Hybrid Cache (Yin et al, 2004); Zone Cooperative Scheme (Chand, 2007); COCA (Chow, 2004) and GroCOCA (Chow, 2004); Group Caching Scheme (Ting, 2007); Cluster Cooperative Caching (Chand, 2006); COOP (YuDu, 2005 & 2009); Proactive approach for cooperative caching (Kumar, 2010). However, none of all these researches provides a complete solution for cooperative caching. Each of them has both advantages and disadvantages. In this section, we present some works, which were used and compared with our strategy in more detail.

(Yu Du et al, 2005 &2009) proposed a cooperative caching scheme called COOP for MANETs. To improve data availability and access performance, COOP addresses two basic problems of cooperative caching. For cache resolution, COOP uses the cocktail approach that consists of two basic schemes: hop-by-hop resolution and zone-based resolution. By using this approach, COOP discovers data sources, which have
The Asynchronous Stateful (AS) algorithm (Kahol et al, 2001) is proposed to maintain cache consistency. In AS, the Mobile Station Support (MSS) only broadcasts the changed data to the related cache hosts and can avoid unnecessary IRs. It supports cache consistency verification under any arbitrary sleep – wake up pattern. However, the MSS must record the state of all caches, which consumes many source node resources (Jiannong Cao et al, 2007). In the scalable asynchronous cache consistency scheme (SACCS) from (Z.Wang et al, 2004), MSS maintains a flag and a timestamp for every data item to keep minimum state information instead of all mobile hosts’ information as in AS. This can improve scalability and performance, but stale data can exist in the caches. A novel Dynamic Scalable Asynchronous Cache Consistency Scheme (DSACCS) has been developed by (Z.Wang et al, 2006). In this scheme, the authors proposed three types of cache consistency invalidation reports for multi – cell systems using SACCS and AS schemes: homogeneous IR, heterogeneous IR without roaming check, and inhomogeneous IR with roaming check. In DSACCS, the IR of a data object is broadcasted globally or locally depending on which has the minimum consistency maintenance cost. For efficient broadcast of IRs, some cells are grouped into a logical cell. A variation of DSACCS for such grouped cells is DSACCS-G (Z.Wang et al, 2006). (S.Sankara Gomathi et al, 2007) have proposed the stateful cache invalidation algorithm. This mainly focuses on optimizing the bandwidth utilization and energy consumption in ad-hoc network. The major focus is that every node has the capability of broadcasting the updated data items. Each node has the same amount of energy to maintain certain data item. Frequently accessed data items are cached to reduce the bandwidth requirement. SWRCC (Sleep/Wakeup/Recovery/Check/Confirm) + MUVI (Modified/Uncertain/Valid/Invalid) scheme proposed in (Po-Jen Chuang et al, 2008), which aims to solve the validity problem of cached data after client is disconnected from the server. However, as most of these mentioned approaches only focus on broadcasting data, they have three main limitations:

- There is neither any data-broadcast scheduling nor mentioning whether the altered data items are necessary for all clients. Therefore, these approaches are not efficient as the clients may not connect to the server or data items may not be stored in clients’ cache anymore.
- The server does not have the specific time notification for clients to notify when the clients allow sending the confirmation message on the needed data to be updated. Thus, there are some drawbacks in these approaches such as bottleneck situations which will happen when a large number of clients send requests to the server or when they access to the uplink channel to get data items at the same time.
- Moreover, these strategies do not deal with the disconnection problem of mobile clients. Because the disconnection is a constriction in mobile environment, it is necessary to show excessive interest in this problem.

We have presented the overview of some popular works that are relating to cooperative caching management. However, none of all above researches provides a complete solution for cooperative caching management. Each of them has both advantages and disadvantages. Our proposed approach will overcome the above mentioned drawbacks. Especially, we will focus on the data consistency model in disconnected mobile environment. The cooperative caching management scheme in our study named IMIX-GROUP.
IMIX-GROUP scheme can be considered as an extension of MIX-GROUP (Thu Nguyen & Thuy Dong, 2011). The difference between MIX-GROUP and IMIX-GROUP is IMIX-GROUP has solved data consistency problem while almost other approaches as well as MIX-GROUP have not mentioned yet. In this paper, we will present IMIX-GROUP in more detail and clear.

2. COOPERATIVE CACHING SYSTEM MODEL

2.1 Description of System Architecture

We assume that our IMIX-GROUP cooperative caching architecture has only one server and many mobile users (MU). MUs are mobile devices used to retrieve data items from server or other MUs via wireless link. Because of the similar of structure and data search principle between IMIX-GROUP and MIXGROUP architecture, you can get the detail description of the architecture from (Thu Nguyen & Thuy Dong, 2011).

In this paper, IMIX-GROUP architecture have been completely solved four problems: cache discovery, cache placement, cache replacement and cache consistency. Cache discovery focuses on the way that MUs discovery, select and deliver the requested data items from neighboring nodes in network. Cache placement is also called as cache admission. When a MU receives the response for the requested data items, a cache placement control is triggered at MU to decide whether the data item should be cached. Cache replacement means that MU decides which data item should be removed from its local cache when the caching space is not enough for caching the new data items. Cache consistency guarantees how to keep the cached data items at the mobile clients consistent with the original data at the server. We will describe the processing principles and meta data of these problems in more detail in next section.

2.2 Management Cache

2.2.1 Cache Discovery

As we know that cache discovery shows the way to search requested data items in system. Our cache discovery control is divided into three cache discovery mechanisms includes local cache discovery, In zone discovery control and Out zone discovery control. These mechanisms are described in more detail as below.

a) Local Cache Discovery

Each MU has a local cache (LC). It stores the frequently accessed data items to serve local data search process. Caching data items in the local cache of each MU help reduce latency and increase accessibility. The LC structure includes \{id, f, t, L, TTL, D, S\} where id is the identification of the cached data item, f is the frequency of cached data item accessed by other MUs; L is the label of the cached data item. Basing on L, we
will recognize a data item is the primary data or the secondary data; \( t \) is the timestamp of cached data item. \( TTL \) value is a data item’s living time, \( D \) is data value of data item \( d \), \( S \) is a size of data item \( d \). When a MU requests for data item \( d \), it first will looks for \( d \) from its local cache. A local cache hit occurs if the request is satisfied in local cache.

b) In zone Query Process
When the data request is missed and not responded from MU’s LC, the data request will be sent to neighboring MUs of the requester. The requesting MU links to neighboring MUs through one hop. In this process, each MU has an information profile named \( ZoneData \). \( ZoneData \) is used to store the information of the neighboring MUs and their data items sent to the requesting MU before. The \( ZoneData \) structure includes \( \{id, L, id_{sender}, t \} \) where \( id \) is the identification of requested data item; \( L \) is the label of data item cached at MU sender; \( id_{sender} \) is the identification of MU sender; \( t \) is the time that MU resource gives this data item. Basing on the \( ZoneData \), MU resource can find out the MU destination that is caching the request data in its zone. \( ZoneData \) is as a way to avoid blind broadcasting to all MUs in the region. This way will reduce the communication cost and increase the system performance.

c) Out zone Query Process
When the data request is not responded from neighboring MUs, it will be sent to networking MUs. The requesting MU link to networking MUs through at least two hops. This process used \( OutZoneData \) table to discover the MU destination. The \( OutZoneData \) structure includes \( \{id, ID_{sender}, m, t \} \), where \( id \) is the identification of the data item \( d \); \( ID_{sender} \) is the identification of a networking MU which is storing the data item \( d \); \( m \) is the number of hops from MU requester to that MU network; \( t \) is the time that MU requester gets the data item \( d \). The data discovery path selects MUs network that \( m \) value is minimum. This approach will increase data response with the least hops.

2.2.2 Cache Placement
When a MU receives responded data items, a cache admission control is triggered at MU to decide whether the data item should be cached. The idea of cache admission control is known as each cached data item will be marked as the primary data (PD) or the secondary data (SD). This process must always ensure that there is only one primary data copy in cooperation zone. That means once a MU fetches a data item, it labels the data item as the primary copy if the data item comes from a MU beyond the zone radius. Otherwise, if a data item comes from within the zone radius, we need to consider whether the data provider labels the data item as primary or secondary. If one of the providers already labels its copy as primary, the new copy would be secondary since we do not intend to have duplicated primary copies in cooperation zone. On the other hand, if all providers tag its own copy as secondary, the new copy is primary copy. The idea of classifying cached data items is the same way that Y. Du et al used in COOP scheme(Yu Du et al, 2005 & 2009). However, the difference between COOP and our mechanism is COOP considered the marking of data label based on multi hops zone while as our approach only acknowledge this problem on one hop cooperative zone.

An example for marking data label in our approach is illustrated in Figure 1. When MU 3 requests data item A, MU 4 replies and marks A’s label as primary. Because MU 4 is in MU 3’s zone, so data item A is copied at MU 3 as the secondary data. Then MU 1 requests data item A, MU 3 responds A to MU 1 as the secondary data. MU1 will caches A as the primary data because MU 1 has the same zone with MU 3, but it has other zone with MU 4.

2.2.3 Cache Replacement
Cache replacement means that MU decides which data item should be removed from its local cache when the caching space is not enough for caching the new data items. The policy of IMIX-GROUP’s cache replacement is based on three factors: \( L, f, TTL, t \) and \( S \). These factors, as described above \( f \) is the access count of the data items that reflects data popularity in the region; \( L \) is used to classify the data items as primary or secondary (primary data item is more priority than secondary one); \( TTL \) is the living of the data item, \( t \) is timestamp of data item, \( S \) is size of data item.

When the local cache has not enough space for storing new data item, cache replacement policy will choose the data item that has invalid \( TTL \) value. In the case, there is not invalid data item; we choose the secondary data item whose \( \beta \) value is smallest. This decision is selected because cooperating caching is not only on the behalf of the caching MU but also basing on the other MUs need. \( \beta \) value is evaluated as follow: \( \beta = \frac{f}{t} \times S \)
2.2.4 Cache Consistency

In our consistency model, we assume that MUs are not allowed to update cached data items. This means MUs are allowed to delete old data items or add new data items but they are not allowed to edit data items in its local cache. Editing data items is only done by the server and these new data items are only sent to clients by the server. The principle of consistency processing is presented as follow:

At the beginning of each updated data broadcast cycle $T$, the BS will send invalidation report ($IR$) to clients to notify its data status. The client bases on $IR$ to determine which its cached data items are influenced, then it send the confirmation requests to the server for requiring data updating. The BS bases on these confirmations requests to schedule data broadcast report ($DBR$). $DBR$ will send to clients and then they base on this $DBR$ to schedule sleep/wake up for receiving data. Updated Data Broadcast Cycle $T$ (UDBC $T$) is illustrated in Figure 2 and the processes of UDBC $T$ are presented at Table 1.

However, MUs are frequently disconnected in cycle $T$ due to unreliable links, and battery saving. Under these circumstances, a MU’s local cache may become inconsistent and it can update its local cache by using the searching principle of IMIX-GROUP to discovery lost information or data items from its neighbors. We divided our data consistency model into two main consistency models. The first consistency model is the client-server consistency model (is called cache consistency model between BS and MUs). The second one is the peer-to-peer consistency model (is called cache consistency model among MUs) which is used for disconnection occurring between BS and MUs. The first model has been published in (Thu Nguyen & Thuy Dong, 2010). In this paper, we only mention to the second model as the extension of the first model to solve for disconnected cases basing on IMIX-GROUP principle.

In our strategy, we assume that MU loses communication with the server at $t$ ($T_i < t < T_{i+k}$) and divide this scenario into three main missed cases: (i) the MU is lost the $IR$ information issued from server (at Figure 3), (ii) the MU is lost the $DBR$ information broadcasted from server (at Figure 4), and (iii) the MU is missed broadcasted data items (at Figure 5). These lost information will be researched from MUs in system based on the processing principle of IMIX-GROUP.

(i). The MU loses the $IR$ information: while the server broadcasts $IR$, the MU can be disconnected and misses $IR$. So, it does not have any information of updated data items. If they reconnect at time $t$ with $T_{i+1} < t < T_{i+2}$ (at Figure 2) then it can request $IR$ from other MUs in the network. Otherwise, it must remark invalidation data items in its local cache and wait for relieving $DBR$ and broadcasted data items within cycle $T$ to check cache validation.

(ii). The MU loses the $DBR$ information: the MU can be disconnected while the BS is broadcasting the $DBR$. So, the MU does not have the information of the data broadcast schedule. If the MU reconnects at time $t$ with $T_{i+1} < t < T_{i+2}$ (at Figure 2) then it can request $DBR$ from other MUs in network. If the MU receives $DBR$ from other MUs then it can determinate when updated data items will be broadcasted to the downlink channel. In contrary, the MU must request updated data items from other MUs in network.

(iii). The MU misses broadcasted data items: we assume that the MU can be disconnected while the BS is broadcasting updated data items to the downlink channel. And the MU misses some or all broadcasted data items when it reconnects again. In this case, the MU can request data items lost from other MUs in network. Figure 5 presented the illustration of missing data items.
Table 1. Detail description of UDBC T’s processes

<table>
<thead>
<tr>
<th>Time period</th>
<th>BS’s processes</th>
<th>MUs processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>From $T_i$ to $T_i+1$</td>
<td>The server broadcasts the invalidation report ($IR$) to the downlink channel</td>
<td>Clients get $IR$ and determinate which cached data items are affected.</td>
</tr>
<tr>
<td>From $T_i+1$ to $T_i+2$</td>
<td>The server receives the confirmation messages from clients and schedules the data broadcast</td>
<td>Clients send the confirmation requests to the server for confirming that they want to update data.</td>
</tr>
<tr>
<td>From $T_i+2$ to $T_i+3$</td>
<td>The server broadcasts the data broadcast report ($DBR$) to the downlink channel.</td>
<td>Clients get $DBR$ and schedule sleep/wake up to receive broadcasted data.</td>
</tr>
<tr>
<td>From $T_i+3$ to $T_i+k$</td>
<td>The server broadcasts the data items to the downlink channel.</td>
<td>Clients get updated data items and refresh their local cache.</td>
</tr>
</tbody>
</table>

Table 2. Simulation parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation area</td>
<td>950m x 510m</td>
</tr>
<tr>
<td>Total #MU</td>
<td>10 – 100</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>2 Mb/s</td>
</tr>
<tr>
<td>Radio range</td>
<td>250 m</td>
</tr>
<tr>
<td>Cooperation zone radius</td>
<td>2-3 hops</td>
</tr>
<tr>
<td>Total # data items</td>
<td>1000</td>
</tr>
<tr>
<td>Client cache size</td>
<td>2000 – 20000 KB</td>
</tr>
<tr>
<td>Data item size</td>
<td>1 – 1000 KB</td>
</tr>
</tbody>
</table>

3. SIMULATION RESULTS

3.1 Simulation Environment and Experiments

We built a simulation scenario and implemented by NS2 simulation in (Sinh & Quang, 2011), (Thi & Tuyen, 2011). The simulation parameters are configured as in Table 2. The goal of the simulation is to test the quality of the IMIX-GROUP scheme. The performance metrics used for testing the quality of IMIX-GROUP scheme include: average response time for a data request, average number of messages, and cache hit ratio. These metrics are evaluated basing on the MU’s cache size and the number of MUs varying in system. The experimental results are compared with those of other schemes such as GROUPCACHING (Ting, 2007) and COCA (Chow et al, 2004). The plot of results in (Thu Nguyen & Thuy Dong, 2011) showed that our approach is more effective than other ones (GROUPCACHING & COCA). In this section of paper, we add the experimental results of data consistency problem to our approach. These results are also compared with SWRCC+ MUVI scheme (Po-Jen Chuang et al, 2008) to evaluate the effective of our consistency model. The experimental results shown in the plots prove the effectiveness of our proposed approach.

3.2 Experimental Results for Cache Consistency Scheme

3.2.1 Effect of AT and TT based on Number of Updated Data Items

We evaluate the effectiveness of our cache consistency model through performance metrics such as Access Time (AT), Tuning Time (TT) (J. Xu et al, 2002) in three varying cases. The first scenario, we evaluate AT, and TT in the mobile environment that has no disconnection occurring. The second one is in the mobile environment having disconnection without cooperative cache among MUs. The third one is same to the second one but we applied IMIX-GROUP cache cooperative strategy to search need information. The result plots in Figure 6, Figure 7 shown that AT and TT in disconnected environment using IMIX-GROUP cooperative cache scheme are closely effective with the case that has no disconnection occurring.

3.2.2 Effect of AT and TT based on comprising between IMIX-GROUP Consistency Scheme and SWRCC+MUVI.

Our IMIX-GROUP data consistency scheme is also compared with SWRCC+MUVI scheme (Po-Jen Chuang et al, 2008) based on invalid data average ratio updated, uplink ratio and access time. These metrics were evaluated based on the difference disconnection probability of clients in system (Po-Jen Chuang et al, 2008). The experimental result plots in Figure 8, Figure 9, and Figure 10 show that our scheme is more effective than SWRCC+MUVI scheme.
4. CONCLUSION

Cooperating caching is more interesting problem in mobile environment. In this paper, we proposed a cooperative caching model named IMIX-GROUP, which addresses all basic problems of cooperative caching: cache discovery, cache admission, cache replacement and cache consistency. Experimental results show that IMIX-GROUP improved data availability and access performance. A supplemental evidence for our success is the average number of messages, the average response time and cache hit ratio that IMIX-GROUP has archived is more effective than COCA and GROUPCACHING. Besides, our consistency model is also more effective than SWRCC+MUVI scheme. However, we still have not taken care about the mobility of MUs in IMIX-GROUP model yet. This will be considered in our future work.

REFERENCES

Journal

Conference paper or contributed volume

Yin, L.Cao,G,“Supporting Cooperative Caching in Ad hoc Networks”, In 2004 Infocomm, The 23rd Conference on the IEEE Communications Society.


Jiannong Cao, Yang Zhang, Guohong Cao, “Data Consistency for Cooperative Caching in Mobile Environments”, Published by the IEEE Computer Society, 2007.


QOS PROVISIONING IN CONTENTION AWARE MANETS USING FLOW-AWARE ADMISSION CONTROL PROTOCOL

Muhammad Asif, Zhili Sun, Haitham Cruickshank and Naveed Ahmad

CCSR, University of Surrey, UK

ABSTRACT
The popularity of MANET based applications is on the rise by the day and this includes the use of multimedia application over MANETs. The existing routing protocols provide best effort service, but do not provide any guarantee of Quality of Service (QoS) provisioning. Admission control based approach is desirable and plays a vital role in maintaining QoS for MANET-based applications. In this paper, we present a novel Flow–Aware Admission Control (FAAC) protocol that will maintain guaranteed throughput to the applications requiring QoS. FAAC protocol is designed to utilize the caching mechanism of the Dynamic Source Routing (DSR) protocol. It will be implemented in two stages: the first stage is searching the cache for untested paths from source to destinations and initiating the route search before checking the nodes resources. The second stage will include checking of local and carrier sensing neighbors’ resources. The protocol is implemented using C++ within NS-2 simulation environment and validated to check the effect of newly admitting traffic over admitted data traffic. The newly arrival traffic was blocked when there is no enough network resources to support the existing and newly arrival traffic.

KEYWORDS
Mobile ad hoc networks (MANETs), Throughput, Quality of Service (QoS) and Flow-aware admission control (FAAC) protocol

1. INTRODUCTION

The applications of Mobile Ad-hoc Networks (MANETs) are gaining momentum by the day in every field of life including military, business and social gathering applications [8]. This expanding use of MANETs is made possible by the easily available and inexpensive IEEE 802.11 devices. The advancement in technology and the emerging use of multimedia applications have triggered the quest to use MANETs to support these applications [2]. To provide QoS assurance in MANETs is a challenging task and have a lot of research opportunities. The QoS provisioning in MANETs cannot be guaranteed alone at network layer but needs to be addressed by different layer co-operation [10]. Most of the routing protocols (DSR, AODV etc) are designed for MANETs to provide best effort services. Mobility and the lack of centralized control in MANETs make it difficult for these routing protocols to support QoS applications. To address these challenges and provide guaranteed throughput to these applications, Flow-aware Admission Control (FAAC) is designed and implemented.

FAAC protocol is partially coupled with Dynamic Source Routing (DSR) protocol [7] and is designed to use the basic routing functionality of DSR protocol. FAAC protocol is designed to ensure: the sessions are admitted according to the capacity of the network; and the intended route of the session have sufficient capacity to accommodate new sessions without affecting already admitted sessions. The session completion ratio of the protocol determined, how effective the protocol is in carrying out the admission procedure and constant monitoring of the session’s throughput. The implementation of FAAC protocol is in two phases: first to find a route from source to destination by searching the route cache and initiate Route Request (RReq) if no route is found and second to test the local as well as two hops neighbors’ resources during Session Request (SReq). The implementation of this protocol in two phases is to take advantage of the frequent paths availability to intended destination.
This paper discusses FAAC protocol design and demonstrates how the protocol is used to test the resources at the time of Session Request (SReq). The restriction of resource testing to the second phase (SReq) is due to the availability of the full knowledge of Contention Count (C_count) during this phase. The design is modeled and simulation studies conducted to show the effect of Admission Control (AC) protocol on improving the QoS provisioning in MANETs. The remaining part of the paper is structured as follows: section II discusses the background work; section III presents the FAAC protocol; section IV shows the simulation environment and results analysis while section V presents the conclusion of the findings of the research.

2. BACKGROUND REVIEWS

The work in Contention Aware Admission Control (CACP) protocol [12] is considered something of landmark in the design of Admission Control (AC) Protocol. It states that the data session not only affect the resources of routes nodes but their neighborhood nodes as well. It has devised three different methods to test the resources of neighborhood nodes. The protocol maintains only single path from source to destination which becomes difficult to cope with the mobility. CACP do not take into consideration the effect of newly admitted data traffic and then due to increase in collision, session completion ratio is very low and also do not take advantage of DSR caching mechanism [1]. Perceptive Admission Control (PAC) protocol [3] enables the nodes to estimate their available bandwidth for admission decision by sensing the transmission. PAC uses passive monitoring to estimate the available capacity at the current node and its neighbors. PAC has devised new transmission sensing range and checks the resources using carrier sense method that is not near to practical deployment of the MANETs. Staggered Admission control (StAC) protocol [4] used to test the local available resources of nodes more than twice to get admission into the network which is the waste of resources and its delays the admission of sessions. All these studied protocols do not take into consideration the caching mechanism of Dynamic Source Routing (DSR) protocol. The mechanism of these protocols to find the suitable path for new data session is different from our designed protocol [5].

3. FLOW-AWARE ADMISSION CONTROL (FAAC) PROTOCOL

FAAC protocol incorporates both routing and admission control aspects of operation. Its purpose is to provide end-to-end guaranteed throughput services to application data sessions that have a strict constraint on the minimum level of throughput they require. FAAC includes features to discover routes that nominally have adequate capacity to support admission of data sessions, as well as to admit only those new sessions that would not have a derogatory effect on the throughput of the previously-admitted sessions and finally to uphold the level of throughput that it has promised to sessions by way of admitting them. FAAC protocol is partially coupled with the dynamic source routing (DSR) protocol, which performs the basic routing functions. By implementing the FAAC protocol on the top of DSR protocol; the issue of using stale routes in DSR cache has been solved, because every route must have to be tested for throughput before data transmission. The novelty of our protocol is the mechanism to find the route which can guarantee the throughput and make use of caching mechanism of DSR protocol and to test the local and neighbor resources during the Session Request (SReq) with the full knowledge of contention count and how to propagate the SReq in the network.

3.1 Protocol Operation

Here, we give a full description of its operation as well as the design choices made. The protocol working mechanism is a combination of application layer, network layer and link layer. We have explained the behavior and characteristics of each layer involved in our protocol.
3.1.1 Application Layer

The application agent defines the notion of a session. A new data session is specified by the following fields: data session ID, start time (s), minimum end-to-end throughput (bps), data packet size (bytes), and session state. The session ID is allocated by the application agent. The throughput requirement defines how many bits, and therefore packets, are generated per second, as well as the desired end-to-end throughput. Traffic is modeled by constant bit rate sources, since this adequately demonstrates the ability of FAAC to handle various traffic loads and to make admission decisions.

When a new session is generated by a user, a blocking timer is set to expire in 10s and a session request (SReq) message is passed to the network at the source node. All source nodes have randomly selected destination in the network. The SReq is passed down to the UDP agent. The UDP agent encapsulates the SReq in a UDP packet, giving it a unique sequential packet ID. The SReq is carried as the application data and passed down to the routing agent, which takes over the handling of SReq.

3.1.2 Network Layer

Application data sessions that are requesting service from and admission to the network are assumed to specify their desired traffic characteristics to the FAAC protocol. In this work, we model these characteristics in the form of Session Request (SReq) packet. The SReq is passed down to the network layer to model the arrival of a session admission request at a traffic source node. The routing agent will find the route in route cache or will initiate the Route Request (RReq). When route is found then the protocol will test the route nodes resources according to Session Request (SReq). The Novelty of our designed FAAC protocol is the method of propagating Session Request (SReq), resource checking and to find the route for throughput sensitive data session.

3.2 Protocol Implementations

FAAC is implemented in two steps:

i) Routing agent generates Route Request (RReq) and finds all the routes between intended source and destination. During this stage of admission, no local and neighbor resources are tested. If the route is available in cache between source and destination, then there is no need for the first phase and the protocol will initiate second phase directly because the task of first phase is to find the path between source and destination.

ii) In second phase of admission control, local and neighbor resources are tested before forwarding the SReq to other nodes. As in the second stage, the full route is known to the source, so FAAC protocol will check the resources with the full knowledge of contention count (C\text{count}).

The capacity of a node (C\text{req}) can be estimated by using the following equation. The session single hop requirement is calculated as:
\[ C_{\text{req}} = b_{\text{req}} * W_{\text{req}} \] (1)

Where \( b_{\text{req}} \) is the required capacity by the session and \( W_{\text{req}} \) is the weighting factor means the overheads of different layers to be included with the data capacity as show in following equation (2) [4].
\[ W_{\text{req}} = \frac{(T_{\text{DIFS}} + T_{\text{RTS}} + T_{\text{CTS}} + T_{\text{ACK}} + 3T_{\text{SIFS}} + T_{\text{backoff}}) + (T_{\text{MAChdr}} + T_{\text{IPhdr}} + T_{\text{SRhdr}} + T_{\text{QoShdr}} + T_{\text{data}})}{T_{\text{data}}} \] (2)

Where \( T_{\text{RTS}}, T_{\text{CTS}}, T_{\text{DATA}} \) and \( T_{\text{ACK}} \) are the times taken to transmit Request-to-send (RTS), Clear-to-send (CTS), Data and ACK frames (along with the physical layer preambles) respectively. \( T_{\text{MAChdr}}, T_{\text{IPhdr}}, T_{\text{SRhdr}}, T_{\text{QoShdr}} \) are the times taken to transmit the fixed size MAC, IP, source route and QoS-specification (SReq contents) headers on each data frame, and \( T_{\text{DIFS}} \) and \( T_{\text{SIFS}} \) are the times taken by distributed coordinated function (DCF) inter-frame space (DIFS) and short inter-frame space (SIFS) employed by the direct sequence spread spectrum (DSSS) physical layer (PHY) specification in IEEE 802.11 standard [6]. Where \( T_{\text{backoff}} \) represents the time for which a node backs off before each packet transmission. So for any node to forward the SReq should satisfy the following equation:
\[ (T_{\text{idle}} - T_{\text{res}} + T_{\text{res}}(i))\beta > C_{\text{req}} * C_{\text{count}} \text{ where } (\text{sres, } i) \in {1, 2, 3, 4, \ldots} \] (3)
Where \( T_{idle} \) is the fraction of channel idle time, \( T_{sres} \) is the fraction of the channel time, which is not yet being used, but which has been reserved by previously processed session admission request (SReq). \( T_{sres(i)} \) is the fraction of the channel time that the current session \( i \), has reserved, if it had recent previous admission attempts and \( \beta \) is the node transmission rate, which specifies the raw channel capacity in bps.

The source node of the data traffic will check its local resources according to equation (3). If it can satisfy the requirement of the new session then it will check the resources of its two hops neighbor resources using the CACP-multi-hop. The node will forwards the SReq when its local and neighbor nodes can accommodate the resources. FAAC protocol checks the node resources during the session request phase with full knowledge of contention count (\( C_{count} \)). Contention count of the node can be calculated by the following formula [12],

\[
C_{count} = (CSN \cap R) \setminus D
\]  

(4)

Here Contention Count (\( C_{count} \)) is the combination of Carrier Sensing Neighbors (CSN) and tentative route (R) of the data traffic excluding the destination node (D). The destination node does not transmit the data further therefore it is not considered in \( C_{count} \).

FAAC protocol is implemented on the top of routing protocol means it is the protocol of network layer in OSI reference model. FAAC protocol receives the SReq from application layer which states the throughput requirement of the session. FAAC protocol cache the SReq information in SReq table. In figure 1 small circle represent mobile nodes, middle and large circles represent the transmission and carrier sensing range of node ‘B’ respectively. Node ‘S’ is the source and node ‘D’ is the destination of the data session. Solid arrows represent the intended data route from ‘S’ to ‘D’ and dotted arrows represent the transmission of Admission Request (AdReq) control packets from node ‘B’ to its two hops neighbors to check their capacity.

Let say the SReq approaches the node ‘B’ and ‘B’ will calculate its residual capacity while considering the Contention Count (\( C_{count} \)). So node ‘B’ will consider nodes \{S,A,B,C,E\} as a Contention Count because these nodes are within the Carrier Sensing Range (CSR) of node ‘B’ as well as part of intended route and also none of these nodes is destination. At this stage node ‘B’ will check its local resources using equation (3). Figure 2 shows the processing of SReq by each individual node. SReqtime is the time for which the source node of data waits for Admission denied (AD) before reserving the resources and propagating the SReq. When the local resources of node ‘B’ satisfy the requirements of the new session, it stores this information and broadcast Admission Request (AdReq) up to two hops neighbors. Figure 3 shows how each node process AdReq packet. AdReq TTL represents the number of nodes to which AdReq packet has to be forward. Two hops neighbor to equation (3).

The two hops neighbor’s node will issue the admission (AD) packet to the source or Admission Request (AdReq) that is node ‘B’, if it does not have enough resources to accommodate the new data session. If node ‘B’ does not receive AD packet then it will forward the SReq to other node on the intended route of the data and reserve the required resources of the data session. Each node will continue the process of checking local and neighbors’ resources and forwarding the SReq till destination node. After receiving SReq by destination, it generates Session Reply (SRep) and transmits back to source of the data session on same route followed by SReq. If a data source node receives multiple SReps, it will select the shortest path to destination. The throughput of the session is continuously monitored and averaged over a period of one second, if the session degrades the throughput of previously admitted sessions or its achievable throughput is below the guaranteed throughput, then the session is dropped. FAAC protocol does not support session pausing because it means that protocol is not upholding the guaranteed throughput. The protocol will divert the flow from current route to backup route. The backup route characteristic and local backup route repair will be implemented and valued in future work.
4. SIMULATION ENVIRONMENT

4.1 Simulation Setup

To evaluate and compare the effectiveness of different protocols in a MANETs, all extensive simulations are carried out using Network Simulator NS-2 [11]. Data files produced by NS-2 are further processed by data driven programming language, AWK, which is designed for text-based data, either in files or data streams. Each simulation is carried out under different traffic load and changing mobility. The simulation parameters are listed in Table 1.

4.1.1 Traffic Generation Model

An application was developed to generate a constant bit rate (CBR) traffic between mobile nodes, since this adequately demonstrates the ability of FAAC to handle various traffic loads and to make admission decisions. Different network loads can be introduced in the network either by changing the number of traffic sources, changing packet size or by changing number of traffic sessions generated by each source. In our simulation, eight different sessions per source were generated to study the behavior of the protocol in different network load. Each session was of 60 seconds.

4.1.2 Mobility Model

There are many mobility models available for the MANETs community to use to generate node position and movement, and the choice of mobility model can greatly affect the outcome of a simulation study. According to a survey, 64% of the simulation papers used random waypoint model (RWM) [9]. Because of the needs of our simulation scenario, we used RWM in our simulation.

FAAC protocol can be used with any other mobility model that is appropriate for other researchers. To study the effect of mobility of nodes network performance, different mobility scenario files are generated with different seeds.

---

**Figure 2. Processing of SReq**

- # Received SReq
- If (Bavail >= Breq) then
  - Broadcast AdReq
- Note: If (AD) then
  - Drop SReq
  - Inform Source Node
- Else
  - If (time>=SReqtime) then
    - Reserve resources
    - Propagate SReq
  - Else
    - Goto Note: End if
  - End if
- Else
  - Drop SReq
  - Inform Source Node
  - Goto Note: End if

**Figure 3. Processing of AdReq**

- #Received AdReq
- # Store travelled route in cache
- If (AdReq is redundant) then
  - Drop AdReq
- Else
  - If (Breq > Bavail) then
    - Drop AdReq
  - Else
    - Reserved Capacity
  - If (AdReq TTL > 0) then
    - Decrement TTL
    - Rebroadcast AdReq
  - Else
    - Drop AdReq
  - End if
- End if
4.1.3 Communication Model

Contention aware MAC layer of the IEEE 802.11b [6], Distributed Coordination Function (DCF) is used as a communication model. DCF uses carrier sense multiple access with collision avoidance (CSMA/CA) technique for channel contention. Mobile node resources are shared within transmission range as well as carrier sensing range. In selection of IEEE 802.11b and IEEE 802.11e, the later one just give the service differentiation not guaranteed quality of service to the application, so with the compatibility with the earlier admission control protocols, we used the earlier 802.11b.

4.2 Performance Evaluation Metrics

Different performance metrics can be used for the evaluation of QoS-aware and Admission control protocols. We presented our results in terms of these metrics: Session Admission Ratio (SAR), Session Completion Ratio (SCR), Packet Loss Ratio (PLR), Aggregate throughput, Average end-to-end delay and Aggregate useful throughput. The most important metric is SCR, which shows the effectiveness of the protocol to maintain the guaranteed throughput throughout the session duration. Simulation runs for 800 seconds but the results of first and last 100 seconds are not used for the purpose of accuracy. The following protocols DSR, CACP and FAAC are evaluated under different mobility and traffic load conditions. We have selected CACP and DSR for comparison because CACP is the well accepted state of the art AC protocol and our protocol based on DSR protocol.

4.3 Simulation Results and Analysis

4.3.1 Session Admission Ratio

Figure 4 represents that session admission ratio of DSR is higher than any other Admission Control protocol. Actually in DSR there is no mechanism to check the resources before getting the session admission into the network. FAAC protocol admission ratio is lower than DSR and CACP protocol due to through checking of resources, as the main contribution of our protocol is to maintain the throughput requirements of the session till completion. The admission ratio of both protocols FAAC and CACP decreases as the number of sessions (load) increases in the network. Network load have no effect on DSR admission ratio.

4.3.2 Session Completion Ratio

Figure 5 shows a significant percentage of CACP sessions dropping even at lowest load. This is partly because CACP does not consider the increase in collision rate that occurs upon session admission. The dropped sessions also allow more sessions to be admitted to use the freed capacity.

DSR performs better in lower network load because there is less congestion, but as the load increases, the session completion ratio is dropped. The FAAC protocol maintain almost high completion ratio at lower as well as higher network load. This is due to the careful admission of sessions and takes into consideration the effect of new sessions over previously already admitted sessions.

4.3.3 Packet Loss Ratio

Packet Loss Ratio (PLR) is an opposite to Packet Delivery Ratio (PDR). Figure 6 presents high packet loss ratio (PLR) of Dynamic Source Routing (DSR) protocol due to no session pausing mechanism as well as DSR admits all session requesting admission into the network. The PLR of DSR increases from 8% to 40% as the

<table>
<thead>
<tr>
<th>S.No</th>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No. of nodes</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>No. of traffic source</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Sessions per source</td>
<td>5,10,15,20,25,30,35,40</td>
</tr>
<tr>
<td>4</td>
<td>Data packet Size</td>
<td>512 bytes</td>
</tr>
<tr>
<td>5</td>
<td>Mobility model</td>
<td>Random WayPoint</td>
</tr>
<tr>
<td>6</td>
<td>Routing protocol</td>
<td>DSR</td>
</tr>
<tr>
<td>7</td>
<td>Node speed</td>
<td>1-2 m/s</td>
</tr>
<tr>
<td>8</td>
<td>Transmission rage</td>
<td>250m</td>
</tr>
<tr>
<td>9</td>
<td>Carrier sensing range</td>
<td>500m</td>
</tr>
<tr>
<td>10</td>
<td>Channel capacity</td>
<td>2Mbps</td>
</tr>
<tr>
<td>11</td>
<td>Simulation area</td>
<td>1500m * 1500m</td>
</tr>
<tr>
<td>12</td>
<td>Node pause time</td>
<td>801 sec</td>
</tr>
<tr>
<td>13</td>
<td>Simulation time</td>
<td>800 sec</td>
</tr>
<tr>
<td>14</td>
<td>Results averaged over</td>
<td>10 runs</td>
</tr>
</tbody>
</table>
traffic load increases in the network. PLR of CACP is low as compared to DSR but high as compared to FAAC. PLR of CACP changes from 3% to 12% when number of sessions per source increases from 5 to 40. Relatively low PLR of CACP due to the reason that CACP pauses the session readily and does not admit all admission requesting sessions.

FAAC has lowest PLR at different network loads because of low session admission ratio of FAAC. It checks the available resources thoroughly before admitting the session into the network. There is no session pausing mechanism in FAAC because session pausing mean that sessions are not able to transmit at their desired level. But the ultimate purpose of FAAC to provide constant guaranteed throughput to the agreed applications. It increases from 0.8% to 6% when number of sessions per source increase from 5 to 40.

4.3.4 Aggregate Throughput

Throughput is the average successful reception of data at application layer. It is usually measured in bits per second (bps). Figure 7 shows the aggregate throughput of the protocols. The aggregate throughput of all the studied protocol increases relatively more when the sessions per source increase from 5 to 25. FAAC protocol maintains high aggregate throughput in low as well as high traffic load. CACP performs better than DSR at high traffic rate. At low traffic load (upto10 sessions per source) DSR performs better than CACP and FAAC because there is no congestion at that time in the network and enough resources are available for all traffic sessions and PLR is low.

The useful throughput parameter explains the actual position of the protocol as shown in figure 8. This metric is the same as the aggregate throughput, except that it is multiplied by the session completion ratio (SCR). This indicates what fraction of the throughput was useful in terms of allowing an application data session to be completed while upholding its throughput requirements.

4.3.5 Average End-to-End Delay

Figure 9 represents average end to end packet delay of a data session. End-to-end delay include different types of delays including buffer delay, queuing delay, re-transmission, propagation delay. In DSR end to end delay increases with increasing number of sessions due to increasing collision rate and PLR. Session pausing mechanism in CACP also contributes to higher delay other than collision and PLR with high traffic load. Due to thorough admission control, low PLR and absence of sessions pausing mechanism average end to end delay in FAAC remains low. Due to very high delay in DSR, it is not included in that graph.
5. CONCLUSION

In this paper we have presented the design and implementation of the Flow-aware admission control (FAAC) protocol. The FAAC protocol was evaluated with Dynamic source routing (DSR) and Contention Aware Admission Control (CACP) protocols. The main feature of the proposed protocol is to assure guaranteed throughput to the admitted sessions. The completion ratio of FAAC, DSR and CACP shown in the simulation are 85%, 11% and 16% respectively, which represent the assurance of throughput. The average aggregate useful throughput are 157kbps, 18.8kbps and 28 kbps of FAAC, DSR and CACP, which show the higher throughput achievement of the protocol.

6. FUTURE WORK

The FAAC protocol will extend to FAAC-multipath protocol that will maintain multiple paths between every source and destination. Data transmission will take place on primary path and secondary path will be reserved as a back. Secondary path will also be constantly monitored passively for the resource availability. The data session will be shift from primary path to backup path whenever primary route fails either due to mobility or lack of resources. The data session transmission will not be affected with the route failure. Backup Local route repair will also be added to the protocol.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support from the EU FP7 MONET project and Higher Education Commission (HEC) of Pakistan for this research.

REFERENCES

THE SHAPE OF THINGS TO COME: SCENARIOS AND VISUAL STORIES FOR TELECOMMUNICATION IN 2020

Dr. Henning Breuer, Dr. Hartwig Grabowski and Dr. Heinrich Arnold
Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin, Germany

ABSTRACT
The road ahead towards ubiquitous information and communication is not a deterministic path, rather radical alternatives are possible and viable within the upcoming decade. In order to systematically explore alternative future developments in telecommunication and human-computer interaction we investigated futures research reports and trend projections, interviewed experts and conducted a series of workshops with corporate and domain experts. Four scenarios named End of Growth, Postmaterial Lifestyle, Technocratic Corporatism and Bright Digital Future, have been derived. They differ regarding their degree of user impact and technological progress. With a new matching-based approach, nine strategic topics (like machine-to-machine communication and sustainability) have been derived out of these scenarios. Finally the potential of different formats of illustration and animation to advance and communicate futures research results is elaborated upon.

KEYWORDS
Learning Organization, Telecommunication, Strategy, Scenario Planning, Futures Research, Visual Storytelling

1. INTRODUCTION
IT and telecommunications are considered the most innovative and dynamic industries of our times. Over the last 20 years they heavily impacted all aspects of life. Information technologies have developed from a workout exercise for geeks to a multi-trillion dollar business, while the telecommunication industry (telco) has developed from a governmental regime towards worldwide revenues of 3.7 trillion ($10^{12}$) dollars in 2009 (WTIR 2010). In some cases technological breakthrough has been exploited in order to substitute old, and to generate new business. In other cases advances in usability or user participation have driven these changes with impact on all aspects of modern culture including the ways that business is run today. Still, just like human beings the industry needs to look ahead to not be dead tomorrow. Strategic orientation is needed to inform the allocation of resources and the perception and rejection of opportunities. What measures should IT and telecommunication companies take today in order to prepare for an uncertain future, what search fields should they invest in or at least investigate in more depth?

What if? What if, in contrast to the conventional projection, the current exponential growth in processing power, connectivity and online collaboration is followed by a long period of stagnation, or what if increasing computational power and positive feedback accelerating technological progress lead to the emergence of a new singularity (Kurzweil 2005) surpassing human intelligence and taking control of its own actions? Phenomena like exponential growth overstrain our imagination; and in some cases only hindsight bias (the retrospective attribution of predictability to once-future events) provides for a sense of control and of sufficiently informed future orientation.

Insufficient information and reflection upon potential future developments yields fundamental threats to organizations working with mid- and long-term investments. Like other high investment infrastructure companies, telecommunication providers work on five to ten years plans in major parts of their business. Long investment life cycles (10 years and more for infrastructures, 1-5 years for services) for high volume investments into durable infrastructures like cables, satellites and the energy provision to run them pair up with market driving power even within related service fields. Being well prepared for mid-term future developments is of existential importance for all actors in this industry.
In spite of this importance the telecom industry’s track record in predicting future needs is poor. The next big thing was regularly out of sight – not only regarding technological topics (like ATM, ISDN, 3G), but especially for usability and user-driven developments like the World Wide Web, SMS, and Web 2.0. Oftentimes predated mindsets persist and limit future imagination. Tunnels of imagination may then turn out as bottlenecks for innovation. In order to face these threat, to cultivate a sense for potentials within the organization and to inform future oriented research and development we set up a futures research project.

Implicit assumptions about the global future landscape of HCI research and development impact local decisions today. Still, lacking empirical evidence and scientific rigor most authors refrain from futures speculation. Thus, strategic knowledge of orientation remains implicit or missing. Our contribution combines an approach to generate and communicate strategic insight on future developments with exemplary results as a basis for discussion and setting priorities within the development of HCI today. On the one hand this paper focuses on scenarios for 2020 and their impact on managing innovation in the IT and telecommunication industry. On the other hand it focuses on different approaches to (visual) storytelling, sketching, illustration and animation as means to advance and communicate futures scenarios.

2. RELATED WORKS (FUTURES RESEARCH AND SCENARIOS)

Futures research as scientific study of possible, desirable and likely future developments (Kreibich 2000, 9) assumes that different, but not arbitrary or countless futures are possible and viable. Its interdisciplinary and multidisciplinary research approach matches potential futures that also develop across disciplines. As the Greek politician Pericles mentioned in the 4th century B.C. the task is not to predict the future, but to be well prepared for it. Accordingly scenario planning emphasizes decision-making utility as main outcome of inquiry over the production of testable knowledge (Walton 2008).

While futures studies may rarely apply controlled, repeatable and falsifiable experiments with highly standardized methodologies, still, scientific principles apply. Their statements must be plausible and checkable (Opaschowski 2008), being based on transparent, traceable reasoning, and theoretically based methods instead of esoteric insights and beliefs. Recombining knowledge from various disciplines its methodological approach reaches beyond analytical and descriptive paradigms of traditional science to include communicative and participatory accounts (Kreibich 2010). Apart from the differing scenarios as such this paper therefore addresses the communicative visualization and advancement of the results.

Scenarios enable us to anticipate and structure discussion about the shape of things to come. Within the broad field of futures studies scenario analysis and planning represents a privileged methodology (Ramirez, Selsky, & van der Heijden 2010; Ringland 1998). Since the Royal Shell Dutch Group conducted the first systematic scenario studies in the 1970s based on the work of Kahn & Wiener (1967) numerous scenario-processes have been conducted and their results published (e.g. Ducatel et al. 2001; Schoemaker 1995). While the Shell approach gained impetus with the Oil Shock in 1973, dynamic developments in the IT and telecommunications industries and its environments suggest that scenarios are a valuable approach to address and prepare for upcoming uncertainties.

Thinking in alternative futures prepares and informs decisions how to strengthen desirable developments. Scenarios point out alternative and logical consistent development possibilities in the face of abounding uncertainties. Unlike traditional forecasting and even recent Delphi Studies (Münchner Kreis et al. 2009) scenarios do not intend to predict the future. Instead they fuel strategic conversation (Van der Heijden 2005) and challenge conventional assumptions. They prevent from linear extrapolation, and foster thinking in alternatives. Peter Schwartz, one of the founders of this approach, writes: “Scenarios are not predictions. It is simply not possible to predict the future with certainty… Rather, scenarios are vehicles for helping people learn” (1991; also see Schwartz 2009). Other than simulations they identify patterns and clusters among possible futures and include subjective interpretations (Schoemaker 1995, 27). Scenario planning then aims at changing mindsets about external factors antecedent to the formulation of specific strategies.

Principles and best practices of scenario planning have been described in (van der Merwe 2008). In order to identify the most relevant influencing factors and future possibilities an explorative approach was chosen over a normative approach focusing only on the most desirable outcomes (Gaßner & Kosow 2010). The basic methodology of the scenario approach we followed has been described e.g. in Fink, Schalke & Siebe 2000 or Fink & Siebe 2006.
3. METHODOLOGICAL APPROACH

“A constant stream of rich, diverse and thought provoking information” (Schwartz 1991) was needed to feed organizational learning we intended. More than 200 publications have been reviewed and extracted, corporate stakeholders and focus topic domain experts were interviewed, and a series of three workshops has been conducted. All intermediary results have been aggregated and refined within a core team of 4 researchers. Various approaches to advance and communicate results have been explored. The goal was to:

- Identify alternative developments and potential impact in the most vibrant technological focus fields
- Develop consistent and historically plausible future scenarios for the year 2020.

Scenarios describe plausible visions of the future and consist of multiple dimensions (Fink et al. 2000). The dimensions consist of a number of influence areas and relevant trends. In order to identify an undefined number of distinct scenarios we took four steps:

1. System analysis and detection of key factors: System levels and influence areas are developed, influence factors are described, and key factors for future developments are identified.
2. Description of alternative future projections: Future projections are formulated on basis of the key factors, these future projections are development alternatives of the single key factors.
3. Combination of future projections to scenario creation: A manageable number of three to six alternative pictures of the future are developed. They form the basis for the future scenarios.
4. Analysis of the scenarios and their interaction: Scenarios are refined and strategically interpreted; results are condensed and later on used for strategic planning.

Within the first step the subject area is structured into system levels & influence areas with a focus on technology & user respectively human. On the human side we differed between an individual, an interpersonal and a social level looking into attitudes and values, behavior and preferences, social relations and allocation (social structure). On the technology side we started with five areas of vivid research and innovation like usability and security. The specific environment of European telecommunication is dominated by legal frameworks and issues of regulation. Adapting the PESTEC framework (representing Political, Economic, Social, Technical, Ecological and Consumer; Runonen & Mannonen 2009) the global environment was defined to include economy, politics, society, and ecology.

Focusing on the technology and customer side a three months literature review was conducted in order to identify trends and influencing factors within these domains. Besides internal resources for corporate foresight like Customer Foresight and Technology Radar (with reports from an international network of technology scouts; Thom & Rohrbeck 2009) major publications of academic societies like IEEE and ACM have been reviewed. Interviews with 18 experts from research and development in engineering, innovation marketing, and strategic management yielded additional information on potential factors that impact telecommunication within the next ten years. Broadly following a “grounded innovation approach” (Breuer & Steinhoff 2010), 89 factors (e.g. standardization and ubiquitous computing) have been identified and described with evidence from the literature review and interviews.

An influence matrix reveals factors whose development has high impact on the whole system. A core team of five experts from innovation marketing, strategic management and engineering was asked if e.g. developments in standardization impact developments in ubiquitous computing in terms of: “no influence”, “weak / delayed influence”, “regular influence”, or “strong / immediate influence”. The team discussed the influence of each factor on another factor in the system and assigned ratings between 0-3 accordingly. Within the resulting influence matrix the sum of a row indicates the influence a factor has on the system (active sum), the sum of a column indicates how much a factor is influenced by the system (passive sum). A combination of both variables indicated the most dynamic key factors being highly influential and subject to other factors at the same time. For each of the 9 resulting key factors we identified the two main uncertainties associated with each factor and mounted possible developments along these dimensions resulting in two decisive dimensions spanning up a matrix of four alternative scenarios.

One factor for example was convergence. On a vertical axis we differentiated between high and low convergence of content and services (vertical integration). On a horizontal axis low versus high technical convergence of systems (channels, networks, devices / horizontal integration) were distinguished. Four alternative projections resulted (i.e. all in all low level of convergence, high level of horizontal integration, high level of vertical integration, high levels of vertical and horizontal convergence).
The third step started with these 9 key factors and 4 projections for each. A consistency check (matrix) was applied to sort out combinations that are unlikely to occur simultaneously in the future. For each of the 576 potential combinations of projections a consistency value between 1 and 5 was assigned to identify the most consistent combinations. Finally a cluster analysis was performed on the most consistent combinations – grouping combinations with high similarity. As a result, four alternative and consistent sets of projections were identified. These alternative “pictures” of the future span the range of possible developments. Each cluster is interpreted as one future scenario. For each of them we elaborated:

- Names (e.g. “bright digital future”), essential characteristics, driving and reactive forces, winners & losers within the particular scenario, customer needs and technologies,
- opportunities (e.g. dominant players to enter adjacent markets) and threats (e.g. new competitors),
- indicators (e.g. further increase in technical performance and customer demand for high class ICT),
- measures to be taken for each scenario (e.g. opening toward customers and fulfillment partners with flexible supply chain networks and advanced open innovation, fostering a culture thriving openness, creativity, flexibility and customer value orientation).

4. THE SCENARIO SPACE

The four scenarios can be distinguished according to two major development paths: the “technological process” and the “impact of the customer”. The axis “technological process” describes the ongoing increase of computation power, bandwidth, miniaturization, etc. and its extremes can be characterized by the end of “Moore’s Law” (Terman and Lanzori 2006) and “Gilder’s Law” (Gilder 2000) or the ongoing predominance of these laws (According to Moore’s Law transistors placed on an integrated circuit are doubling every 18 months while prices are falling. According to Gilder’s Law bandwidth triples every 12 months). The axis “impact of the customer” describes the involvement and the integration of the customer in upcoming developments. Figure 1 illustrates the two axes with its four scenarios.

Figure 1. Two forces with major impact on the telco industry: “technological progress” and “customer impact”.

The four scenarios span the space of relevant possibilities that ICT players should prepare for. According to the digital value chain, each may be described with respect to the dominating content, available infrastructure, offered services and widely-used devices. For the digital future scenario this could be specified in terms of personalized rich media, user generated content, infrastructure with LTE-A (long term evolution advanced) and more than 100 Mbit/s, FTTH (Fiber to the home) with >200 Mbit/s, and a unique service-
oriented architecture, including network-centric storage, identity-management, and open innovation of context-aware services, and interoperable high-end mobile devices including autonomous robots.

5. FROM SCENARIOS TO TOPICS

The four scenarios roughly describe the space that telecommunication companies should prepare for within the next ten years. These scenarios help to get a better understanding of the future, but on this high level of aggregation they don’t prescribe the most relevant topics for the telco industry. To obtain such a set of most relevant “telco topics”, we developed a three-step approach:

1. Review and aggregation of the 89 factors and contributing information sources initially used to describe and specify the system image into topic fields. Information from internal and external researchers, customer foresight and our internal tool for technology forecasts called “technology radar” (Technology Radar 2009) was leveraged for this purpose. Missing pieces of information were investigated in an iterative fashion (see Breuer & Steinhoff 2010).

2. Filter the topics and topic fields according to three criteria (1) “big” - the topic must lead to a paradigm shift or disruptive change, (2) “new” - it must imply a new insight and (3) “impact” – the topic must have an impact on the telco. As a result, an unordered set of topic fields was derived.

3. Map this set of topics onto the scenarios and judge the impact of each field in the underlying scenario by a team of experts.

Finally, trends with major impact in one or more scenarios were selected, e.g. advanced M2M communication for the Digital Futures and Technocratic Corporatism, and Sustainability as a characteristic of the Postmaterial Lifestyle:

- M2M-Communication: Ever more objects – machines, household devices and vehicles – come equipped with ICT components and communicate via the Internet. The increasing connectivity of devices enables new services, but also creates new requirements for telecommunication enterprises.

- Sustainability: Responsible management of our resources is essential to secure our future. “Going Green” is not enough – sustainable solutions are required which not only delay the exhaustion of resources but stop it through a paradigm shift including regulatory and behavioral aspects.

6. DEEP DIVING AND VISUAL STORYTELLING

The nine topic fields have been investigated in more depth. Simultaneously suitable media for communication of research results have been explored and in some cases the materiality of these media as well as preliminary sketches provided valuable means to advance the subject. Detailing a realistic narrative for instance usually directs attention to aspects that would have else wise been overlooked (Schoemaker 1995, 26). In particular a newspaper with animated images, illustrated deep dives into topics and an animation movie were developed in order to communicate the results in tangible ways.

- An introductory presentation describes the initial motivations for the project, its methodological approach, and the resulting scenarios.

- An animated newspaper exemplifies consequences of consequences of the developments in terms of technological and societal change.

- Deep dives into nine topics describe a reasonable trajectory from observable facts to a potential future within one field of development.

- Finally, an animated movie synthesizes selected aspects of these developments involving advanced animation and artistic irritation.

Thinking through consequences of consequences within the scenarios and focus topics an animated newspaper of March 17th 2020 has been created. The format was chosen to enhance suspense, to present a variety of appealing stories and distribute it not only on paper, but also on ebooks and other mobile devices. Figure 2 shows the chosen illustration – by selecting an image, the related story appears with animated images. Two examples:

- Twin Stalker Law Passed: Berlin – Since DDT (Digital Twin Trauma) was added to the ICD 12 classification of diseases two years ago, more than 1 million cases have been diagnosed worldwide, leading
to overcrowding of mental patient hotels. Victim Tomas D. complains: “My devices are stalking me – I can never relax or sleep. The vacuum cleaner announces sales, the cleaning robot obstacles, the toaster cries for full-grain input. Can’t they just leave me alone?” The Twin Stalker Law will protect humans and their digital twins from obtrusive agents. Privacy and stalking laws are now being applied to smart devices. Deviant programs, devices and robots are now blacklisted at the Federal Systems Control.

- Jan Hog is Eco Star 2020: Cologne – With the lowest, in fact a negative, carbon footprint Jan Hog has beaten all other participants in the popular show „Germany’s Next Eco Star“. His mobile Carbon Calculator showed an astonishing 200KWh, which he has created and uploaded to the power grid at his favourite hotspot over the past 6 months. The New Times has interviewed him on his amazing plus-energy house, new stylish hybrid vehicles, and on his daily „maximum energy upload exercise routine.

Diving deeper into technology-related topics, for each of the nine selected topic a detailed story has been developed. Like future scenarios they describe a reasonable trajectory from observable facts to a potential future within one field of development in a tangible form of narration. Finally, a comic – serving as a teaser – was created. This animated movie (figure 3) synthesizes selected aspects of these developments involving advanced animation and artistic irritation.

Figures 2 & 3. An animated newspaper (illustrated by Gabriele Heinzel), and an animation film created by S. Weikopf (2009), depicting aspects like “smart glasses”, “virtual environment”, “car-2-car communication” and “mobile payment”.

Each illustrative approach has its strengths and weaknesses. The newspaper fosters futures imagination by not only presenting future news but condensed stories of public interest resulting from potential developments. Its strength is triggering controversial discussions among technical laymen, but it is limited by its focus on single influencing factors. The deep dive stories provide for a more comprehensive understanding of a topic and several contributing factors. Not only as if stories, but also interacting factors and systems may be visualized. As stories written from a future point of view they come closest to the original scenarios written by Hermann Kahn. Providing a more complex, almost scientific sounding, but still graspable image of the future and back-casting current developments that led to it they suit for researchers and developers. The animation is most advanced in terms of an interpretable storyline, visualization and specification since use cases are actually shown at work within hypothetical future situations. They are best suited to gather attention and initiate discussion but do not qualify well as shared reference for critical discussion and strategic planning. The most comprehensive communication format for this purpose is the detailed presentation that is available upon request.

In the early days of scenario Kahn and colleagues describes scenarios as stories as if written by people in the future. Other formats of (visual) storytelling have been tried out since, but only rarely been documented. Alternative visualizations that have been documented include for instance a world map of islands – each representing one scenario, or caricature of succinct features of the scenarios (see Fink & Siebe 2006). Still, even though several authors have stressed the importance of suitable ways of communicating scenarios, potentials, strengths and weaknesses of each (visual) approach have never been explored in a systematic way. Such research would need to focus on the different communicative purposes of scenarios (e.g. to initiate public discussion, trigger organizational learning or inform strategic decisions). It would continue works on

7. SUMMARY AND OUTLOOK

For organizational strategy and development the scenario process and its results are capable of several contributions:

- Learning organization: An alternative way of informing experts and employees about innovative technologies and initiating related discussions.
- Challenging given assumptions, new roadmaps of potential development appear. They disclose search fields and opportunities for new business that otherwise might have been neglected.
- The scenarios create a contrasting background for the interpretation of upcoming topics and help to inform strategic decisions.
- Concrete, robust measures may be derived that ensure sustainable growth considering different scenarios at the same time.

In our project, main attention has been paid on the identification of critical topics and on suitable formats for communicating these topics. For developing future scenarios well proven methodologies exist. However, the way from scenarios to strategic topics is still beyond state of the art. Therefore, a new matching approach was developed in which a set of trends is projected into the scenarios and as a result a set of most imported topics is derived.

Advanced communication channels (animated illustrations, movies, storytelling, etc.) have been set up. However, we observed that suitable feedback channels interwoven with the topics and their presentations are still missing. New channels for rich feedback applying e.g. blogs or tweets might enrich communication and the organizational learning intended around the scenarios.

The practical use of scenarios largely depends on the decision making processes following the scenarios. It usually involves strategic planning and budgeting, and setting up suitable monitoring systems (Schoemaker & Gunther 2002). In our particular case, situated in research and development the scenarios outlined above were loosely coupled with top management decisions, but still inform the research and development agenda.

Six of the nine trends could be directly correlated with running research and development (R&D) projects at our laboratory. They are serving as a source for inspiration and are used by the associated R&D teams as a communication tool for motivating their current R&D activities. The “sustainability”-issue was not yet sufficiently covered by ongoing R&D projects – therefore, this trend was used for initiating further R&D activities concerning “sustainability” (see Forum for the Future 2010). Generic trends like “digitization” (of everything that can be digitized) also affect a large set of our R&D projects and indicate the global change in the telecommunication industry. Therefore, they were mainly used for discussion with top-management and build the starting point of several workshops concerning technology strategy.

Since only few systematic applications of scenario processes in organizations have been reported, and given the long timescales being addressed, proving the value contribution of scenarios remains a great challenge. Additional scientific research on the benefit of scenarios (see Schoemaker, 1993), their comparative performance and different formats of visual storytelling and communication is needed. Here the decisive benchmark is not predictive validity, but the contribution of the approach to organizational learning and preparedness for dealing with uncertainty. Essential is an understanding that scenarios may not dissolve uncertainty but should help to get used to it and to prepare decision makers for a variety of developments, even beyond the manageable set of scenarios that may be anticipated.

REFERENCES


ABSTRACT
In this paper we reviewed the distributed query execution architecture and query processing mechanism. Distributed data processing is becoming a reality. In real world complications databases are growing and the need for distributed query processing is also growing. In today’s business world it is highly demanded to stay competitive. Distributed query processing is a complex process because it involves many heterogeneous sites, various client machines and many servers. The load of sites varies over time and new sites keep adding in the distributed environment. All the systems are to be integrated—such legacy systems usually have not been designed for distributed data processing and now need to interact with other (modern) systems in a distributed environment. In this paper we tried to review the art of state of distributed databases query processing.

KEYWORDS
Query Processor, Distributed Database System, Parser, Architecture.

1. INTRODUCTION
There are some specific classes of distributed systems: systems with client-server architecture. In this type of master slave protocols one site send a request as a client to the site, the server, which sends an answer as a response to this request. Using this mechanism, it is possible to implement a variety of different database architectures. In the most general architecture which is Peer to Peer (P2P) all and every site act as a server that stores the parts of the databases and executes the queries and application programs. This P2P systems can initiates the query also. The second category is client server the roles are fixed and either a site must be a acting as client or the site must act as server and all the sites are not allowed to get involved in the communication and clients are not interacting as well as sever is also not allowed to communicate either.

In the multi-tire middleware architecture all the sites can be arranged in a hierarchy and the upper hierarchy sites can act as serve rand lower one should act as client. The sites are allowed to communicate with its clients at the above level or its servers at the level below; typically, a site cannot communicate with sites at the same or any other level.[2]

2. ARCHITECTURE OF A QUERY PROCESSOR
In the given figure the architecture for query processing is described. It is found that this architecture is very general nature and it can be used for all kind of database system including centralized, distributed, or parallel systems. In this the query processor receives an SQL query as input then , translates and optimizes this query in several phases into an executable query plan, and then executes the plan in order to obtain the results of the query. The query execution engine directly executes the interactive ad hoc query (dynamic SQL), and the results are presented to the user. For scanned query the application program part that is embedded SQL , first it will get stored in the database and then executed by the query execution engine, every time the application program is executed.[3]
The distributed query optimization generates the static query plan at compile time. So the optimization of this plan depends upon different parameters. The detail discussion is given for the query processing for different components of the query processor. Each component has its own role in the query processing.

**Parser:** In the distributed database environment parser plays an important role. It parses the query and then this parsed query is translated into query graph which is nothing but internal representation of the query. The parser is used for a centralized and distributed database system and this is the first step in query processing.[4]

**Query Rewrite:** In this phase of query processing optimization is carried out, it rewrite the query in order to perform the optimization. The physical state of the system like table size, indices and table copies and their locations, as well as processor speed is considered in this stage. Redundant predicates should be eliminated and all the expressions must be simplified, all the sub queries must be unnested and all the view is also unnested. In the distributed system, portion of the database table can be selected to answer the query and this is done by the query rewrite. A sophisticated rule engine rewrite this query rewrite operation.[5]

**Query Optimizer:** In this phase of query optimization the physical state of the system is considered by the optimizer. The optimizer uses the indices to execute a query, and it decide which methods like hashing or sorting should be used and which operations of a query like joins and group-bys should be used, and in which order to execute the operations of a query. The query optimizer allocates and decides the memory utilization in each operation. To decide the best optimized alternative plan the optimizer uses the cost estimation models. Query Optimizer uses dynamic programming search to find best alternative to decide the best.[6]

**Plan:** Query execution is specified by the plan component in the query processor, and precisely tells how to execute the query. Most of the databases systems represent plans in the same way as trees. The nodes of a plan are operators, and every operator carries out one particular operation (e.g., join, group by, sort, scan, etc.). The nodes are annotated which indicates the instances where the operator is performing. The edges of a plan represent relationships of operators like consumer and supplier relationship. All the communication activates are encapsulated by the send and receive operators, so that all other operators like Nested Loop Join or scan operator also can be implemented and used in the same way as in a centralized database system.[7]

**Plan Refinement/Code Generation:** Executable plan is generated by this component which is generated by the optimizer. The transformation involves the generation of an assembler-like code which evaluate expressions and predicates efficiently with optimization. Simple optimization techniques are used in refinement of the plan and these optimization techniques are not carried out by the query optimizer in order to simplify the implementation of the query optimizer.[6]

**Query Execution Engine:** This is very important component operators are implemented by this execution engine. Iteration is the foundation of the query engine. Operators are implemented as iterators and will have the similar interface. Two or more than two iterators can be plugged as consumer – producer like manner and this will help in executing the any plan. Results are pipelined and this is another advantage of the iterator’s model and with such approach it will help in improving the performance.

**Catalog:** All the information needed which is must to parse, rewrite, and optimize a query is organized by the catalog. Catalog maintains definitions of the tables and views of the table and other schemas of the databases like used defined types and functions and partition schemas. Catalog is the important component which maintains information about global tables and the table partitions, and physical information such as the location of copies of partitions of tables, information about indices, and statistics required to estimate the cost of a plan. One of the important questions arise in distributed databases is where to store this catalog, so in the most of cases it is stored in the central site but it has still issues like replication and to reduce communication cost it makes sense to replicate the catalog at several sites. It is also advisable to cache the catalog.
information at various sites in the network. Replication and caching both are effective for catalog information because catalogs have smaller size and updates are rare in case of catalogs, but it is not the case sometimes catalogues become larger and require updates regularly. In the case of such environments, it is advisable to partition the catalog data and store catalog data where it is most needed. Distributed object databases need to know where copies are stored (potentially millions), and they need to update this information every time an object is migrated or replicated from one location to another location.[8]

3. QUERY OPTIMIZATION

Query optimization is playing vital role in the query flow of DBMS. Query optimization is the big field and the concept is analyzed and studied with various intentions and aspects and researchers provide various solutions for this concept of query optimization. The query optimizer is handled by the system not by users. Instead when queries are triggered and submitted to database server, parser will parse it, after parsing, the queries are passed to the query optimizer where optimization occurs. Mainly, the size of the tables; the tables location; the indexes availability; the need for procedural logic to support complex requests, that can't be coded using SQL alone; the availability of denormalized structures, (fragments, replicas, snapshots); and consider using common, reusable routines, for each distinct request, simplifying, maintenance and modification are some of the factors that are also considered in the query optimization. Optimization is carried out in two ways and this can be combined also

- **Rule-based Query Optimization**: Rules which are predefined are applied to optimize the query. The Rules are based on projection, selection and on joins. Projection is carried out to reduce the size of the relation, or can reduce the size of the row. The unnecessary attributes should be eliminated. ‘Selection’ also reduces the relation size. Small relations should be joined first and then the larger relations should be joined. Cartesian product operation must be combined.

- **Cost-based Query Optimization**: Number of query plan gets generated and then cost of each query plan is estimated. The lowest cost query plan must be considered.[1]

Performance can be improved by these distributed structures. These structures are very useful and support muti-site and Multi-indexes. These structures maintain physical locations of tables and data. Data items available in the physical locations and tables are provided with information in it. Performance is better in these structures but difficult to maintain. Administration is also difficult due to many sites and the goal of the query optimizer is to produce an efficient execution plan for processing the query, it is a standard and it is canonical query tree. Modular architecture is proposed for query optimizer and all the modules are covered in details [9, 10].

4. QUERY OPTIMIZER ARCHITECTURE

Multiple query execution plans exits when a query is run on given database and the most suitable one should be chosen so that leads for the query optimization. In principle, the best one should be chosen from the alternatives so that performance can be improved. In the given figure we are suggesting a modular architecture for the query optimizer which is tests all alternatives and provides abstraction of the processes running. The two stages rewriting and planning are the most important phases in the suggested architecture and the boundaries shown in the architecture is not tight boundaries, so if a query optimizer is designed using such architecture it will surely improve query processing process.[9] Stage- wise processing is suggested rewriter phase should be performed in the first stage and then other modules should be in the next stage. Each module is analyzed for functionality in the figure given below.
4.1 Module Functionality

4.1.1 Rewriter
All the transformation and equivalent queries are generated with replacements of views and flattening out of nested query is carried out in this phase. This stage should be at the declarative level only to make transformations effective and beneficial. [11]

4.1.2 Planner
Ordering is performed in this module. The entire possible cheapest query plan is evaluated in this stage which is generated in the previous stage. Search is performed to check the space of execution plan. The two Modules of the optimizer, the Algebraic Space and the Method-Structure Space check this execution. Cost of execution considered for execution plan. The Cost Model and the Size-Distribution Estimator are derived from the optimizer.

4.1.3 Algebraic Space
Action execution orders are evaluated by algebraic space with performance criteria. Tree or algebraic formulas are used to represent the algebra space. [10]

4.1.4 Method-Structure Space
Implement choices that exist for the execution of each ordered series of actions specified by the Algebraic Space. Nested loops merge scan, and hash join are evaluated. Duplicates are eliminated and shorting is carried out. All the available indices are evaluated for accessing each relation. This module produces all complete execution plans, which implements algebraic operator and any indices. [12]

4.1.5 Cost Model
Cost estimation is carried out in this module. For joins, indexes and execution plan there is formula for calculating the cost which are simple approximations for CPU, buffers and I/O operations. The size of buffer pool and size of relations or indices accessed are considered, so here DBMS AND SIZE Distribution Estimator plays role in determining the cost. [7, 13]

4.1.6 Size-Distribution Estimates
Size of database relation and query results are estimated and it will be used by cost model which should be maintained in the catalogs.

5. PARALLEL PROCESSING DATABASES
Method-Structure Space module in a sequential environment is one of the choice which offers lot of alternatives in the parallel processing case. [14]. Method-Structure Space module offers two more methods In addition to the sources of alternatives offered in a sequential environment, which are the number of processors that should be given to each database operation (intra-operator parallelism) and placing operators into groups that should be executed simultaneously or parallel by the available processors (inter-operator parallelism, which can be further subdivided into pipelining and independent parallelism). The scheduling
alternatives that arise from these two questions add at least another super-exponential factor to the total number of alternatives, and make searching an even more formidable task. Thus, most systems and research prototypes adopt various heuristics to avoid dealing with a very large search space (space of plans) for optimization purposes[7].

Method-Structure module differentiates centralized and distributed query processing and it provides different processing strategies and opportunities for transmitting data for processing at multiple sites. In earlier distributed systems, only the network cost has more dominance, semi-joins are used for processing in order to only transmit tuples that would certainly contribute to join results. Bloom filters can be used for the same purpose, which are bit vectors that approximate join columns and are transferred across sites to determine which tuples might participate in a join so that only these may be transmitted [2].

6. CONCLUSIONS

Attempt is made to review an overall picture of the architecture of a Database Management System with the architecture for traditional and non-traditional application areas dealing with large volumes of data. All the issues related to parallelism of relational database systems are covered along with their pros and cons. Hardware and software design stuff for the relational query processor is discussed. All the phases of the query processing are reviewed and covered. We found that strength of DBMS lies in the quality, functionality, and sophistication of its query optimizer, since that determines much of the system's performance. Query optimization is one of the important aspects in the distributed databases. Various other researchers presented an abstraction of the architecture of a query optimizer and the functionality of various modules present in it. In every single module of this architecture, there are many questions for which we do not have complete answers, even for the simplest, single-query, sequential, relational optimizations. So, still a great deal of work can be done to know in detail about each of these modules and various optimization methods.

REFERENCES

GRAPH-BASED AUTHENTICATION ALGORITHMS FOR VEHICULAR AD-HOC NETWORKS

P. Caballero-Gil, C. Caballero-Gil and J. Molina-Gil

ABSTRACT
This paper describes a new authentication method for vehicular ad-hoc networks that does not require any certificate authority because public-key certificate is based on the nodes themselves in such a way that it allows sharing trust with issued certificates. This work also includes a graph-based cryptographic protocol that can be used by any vehicle to convince another node about the possession of certain authentication secret without revealing anything about it. Thanks to these tools, collaboration among vehicles can be used to detect and warn about abnormal traffic conditions. One of the most interesting aspects of the scheme is that the required equipment can be simple existing mobile devices.

KEYWORDS
Vehicular ad-hoc network, Collaboration, Authentication, Security.

1. INTRODUCTION
A Vehicular Ad-hoc NETwork (VANET) is a special form of Mobile Ad-hoc NETwork (MANET) designed to provide communications among nearby vehicles with the primary goal of improving road traffic conditions. In many situations, communications among vehicles might be useful to prevent road accidents and to avoid traffic collapses. Consequently, a fast VANET deployment could help drivers to save time and money, and to reduce contamination of the environment and consumption of fuel reserves. There are several general security requirements that must be considered before the deployment of any wireless network. Requisites such as authentication, scalability, privacy, anonymity, cooperation, stability and low delay of communications are even more challenging in VANETs than in MANETs because of specific characteristics such as no fixed infrastructure and rapidly changing scenarios that range from rural roads to cities or highways. Consequently, VANET security may be considered one of the most difficult research topics that need to be taken into account before a wide deployment of such networks (Raya and Hubaux, 2005).

The proposal here presented has as starting point the consideration that the introduction of a complete model of VANET including Road Side Units (RSUs) and On Board Units (OBUs) would be extremely expensive both for users, who would have to buy new cars or install specific devices in their vehicles, and for the state, which would have to deploy a huge infrastructure to support VANET services. Thus, this work proposes a self-managed VANET following a cross-layer approach that does not require any infrastructure, and that might be used as a fast and secure introduction to more complex and complete VANETs.

This paper is organized as follows. Related work and preliminaries are reviewed in Section 2. Section 3 includes a proposal for the generation of key pairs, and characterization of nodes and beacon management. A zero-knowledge authentication protocol is described and simulated in Section 4. Conclusions close the paper.

2. PRELIMINARIES
The main goal of this paper is to define a simple, scalable and cross-layer design for immediate deployment of VANETs that exploits the potential of existing mobile systems. The proposed scheme implies that users can cooperate through their mobile devices and can start to obtain updated information of their interest about
road traffic in order to choose the best updated route to their destinations. Our proposal takes into account that the practical implementation of VANETs will be gradual, without any RSUs or OBUs, and with only a few mobile devices at the beginning. The growth of VANETs will be faster or slower depending on their popularity, acceptance and ease of use. In this paper we focus on the first phase, when the number of cooperating devices on the road will be low. Once VANETs have grown, the model should be checked to avoid unnecessary communications that might degrade the network.

With respect to requirement minimization, several papers focus on different aspects and applications of VANETs. (Yeh et al, 2009) focuses on decentralized vehicular communications without any fixed infrastructure and proposes a method for dynamic establishment of secure communications in VANETs. There are other bibliographic references that propose different approaches to authentication for self-managed VANETs. For instance, (Calandriello, et al, 2007) proposes a scheme that is entirely based on pseudonyms.

Regarding public-key certification, (Laberteaux et al, 2008) presents a method for certificate revocation based on car-to-car epidemic distribution, and (Haas et al, 2009) proposes a mechanism for revoking security certificates that needs a certificate authority and certificate revocation lists. Another related paper is (Dornbush and Joshi, 2007), but its authors do not address the issue of the security of communications. A papers that analyzes security issues in VANETs is (Pohl and Federrath, 2008), which proposes a security infrastructure based on asymmetric and symmetric cryptography to protect the privacy of users.

Neither of the aforementioned works proposes a self-managed and secure approach to VANETs, which is the main objective of this work. In particular, the proposed scheme focuses on allowing the immediate and secure deployment of VANETs through existing devices.

Different tools such as zero-knowledge proofs, hamiltonian cycles, graph isomorphisms and certificate graphs, which are used for the implementation of the proposed self-managed VANETs are now described.

A Zero-Knowledge Proof (ZKP) is a cryptographic protocol that a prover can use to prove possession of a certain piece of information to a verifier without revealing anything about it. ZKPs were first introduced in 1986 by Goldreich, Micali and Wigderson. It constitutes the basis of challenge-response authentication protocols in which users are required to prove the correctness of their secrets without revealing them in order to access a system. During the authentication procedure, a user, usually called Alice (A), must respond to a number of challenges issued by the system, usually called Bob (B). If the protocol is repeated t times, all t rounds must be answered successfully to prove A’s identity so that B is convinced of the validity of the proof with probability $1 - 2^{-t}$. During the execution of a ZKP, the verifier B cannot learn anything from the authentication procedure, which is possible because it always receives only one value from two possible challenge responses and this is not enough to discover anything about A’s secret. The admission control scheme included in the node authentication proposal described below is based on the general scheme of ZKP used in (Caballero-Gil et al, 2006) for the particular case of the Hamiltonian cycle problem.

A Hamiltonian cycle of a graph is a cycle that visits each node exactly once. The determination of whether such cycles exist in a graph is what is called the Hamiltonian cycle problem, which is an NP-complete problem. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ that have the same vertex sets $V_1 = V_2$ are said to be isomorphic if there exists a permutation $\sigma$ of $V$ such that $(a, b) \in E_1 \iff (\sigma(a), \sigma(b)) \in E_2$. The computational problem of determining whether two finite graphs are isomorphic is called the graph isomorphism problem. Such a problem is not likely to be NP-complete, but it is NP as there is no known polynomial time algorithm that solves it. In most cases, ZKPs are based on this type of difficult problems, and just in a few cases they use hard-on-average problems (Caballero-Gil, 2005).

An initial approach to the problem of self-managed public-key certificate in MANETs was described in (Capkun et al, 2003) where its authors proposed the replacement of the centralized certificate authority typically used in wired networks, by a self-organized scenario in which the certification is done through a chain of certificates issued and signed by the network nodes. This scheme is based both on the information stored by each node and on the fact that each node relies on its neighboring nodes. This last feature is also the essence of the self-organized infrastructure proposed here as it is used to build new trust from existing trust. It is said that a graph exhibits the so-called ’small-world phenomenon’ (Kleinberg, 2000) if it is likely that any two vertices are connected through a short path. This phenomenon, also known as ’the principle of six degrees of separation’, implies the idea that everyone is at most six steps away from any other person so that a chain can be made to connect any two people in six steps or fewer. What is interesting of this phenomenon is that it emerges naturally in many different situations such as social networks and self-organized systems.

In the scheme here proposed, based on graph certificates, each node $A$ has a private/public key pair and a key store ($\text{KeyStore}_A$) including a list of all node certificates that $A$ trusts. The set of stored public-keys and
certificates might be represented as an undirected graph \( G = (V, E) \), known as certificate graph in which each vertex represents both a public-key and its owner, and each edge \((a, b)\) symbolizes two public-key certificates: of node \(a\) signed with the private key of node \(b\) and vice versa. A certificate chain is an undirected path in a certificate graph. We define a subgraph \( G_A \) of the certificate graph \( G \) containing exactly the current certificates stored by node \(A\) in KeyStore\(_A\). The growth rate of a certificate graph until reaching the optimal point where it contains enough connections to allow communications between any pair of nodes depends both on the motivation of users to distribute certificates and on their mobility.

3. INGREDIENTS OF THE PROPOSAL

In order to make easier the implementation of the ZKP for the Hamiltonian cycle in the node authentication process described below, the public-key is computed from the decimal value of the binary representation for the upper triangular submatrix of the symmetric adjacency matrix containing the elements corresponding to a Hamiltonian cycle in a graph.

The decimal number corresponding to the binary representation is used as the exponent of the public-key in RSA encryption used by the device to encrypt and decrypt messages and to sign public-key certificates.

In the choice of a key pair by using the Hamiltonian cycle for the generation of the public-key. After choosing the prime numbers \(p\) and \(q\), the public exponent \(e\) is generated from a random Hamiltonian cycle, so that it is lower than and coprime with \((p-1)(q-1)\). Afterwards, the private exponent is generated.

The present proposal assumes that each node in the network is characterized by the following parameters:

- ID, \((KU_{ID}, KR_{ID})\), \(\{ID, KU_{ID}, Cert_{ID}\} \in \text{KeyStore}\)
- ID: Unique IDentifier obtained as the output of a one-way function on a single value. For example, if the used device is a mobile phone the value can be its number, while in other cases an email address might be used. The one-way function could be any hash function.
- \((KU_{ID}, KR_{ID})\): Fixed public/private key pair, called identity keys, which are used in an asymmetric cryptosystem such as RSA.
- KeyStore: Key store containing various IDs and corresponding public-keys and certificates, which the node keeps always updated.

Sending multicast beacons containing variable sender IDs are required both for the active node discovery process and also to avoid vehicle tracking. In the same step where beacons are sent, each node commits to its secret by sending to its neighbours also a witness of its secret. The variable ID of each node that is sent as part of its beacon is the hash of the IDs that are present in its key store at that moment.

In particular, the beacons sent by a node are formed by the following elements:

- Frame-Control (FC), which indicates the type of data being sent.
- Pseudonym (Pseu), which is a temporal ID of the node.
- Timestamp (Time), which allows knowing the moment when the information was generated.
- Pair formed by public-key and timestamp \((KU, Time)\) encrypted with the private-key \((KR)\) of the node, which is used by nodes who have already authenticated it when its Pseu changes.

4. AUTHENTICATION

In this proposal, the device associated to each network node should be able to generate its public/private key pair and also to sign the public-keys of other nodes that want to become part of the network and are trustable. In order to be able to authenticate its public-key, every node must exchange certificates with a number of legitimate network nodes that depends on the width of the VANET. At the beginning, two signatures are enough to prove that the user is reliable and cannot self-sign certificates to compromise the network security, but the number of required signatures must grow with the expansion of the VANET.

Self-organized certification of public-keys makes possible to authenticate a node public-key without knowing it and with no need of any trusted third party. Such a certification is based on trusting the neighbors of your neighbors by forming a certificate graph. In this work we understand that a certificate between \(A\) and \(B\) consists always in two signatures: \(A\) signs the public-key of \(B\) with its private key, and vice versa.
When a node A wants to check the validity of the public-key of another node B, A must find a certificate chain from it to B in the certificate graph that results from merging the subgraphs $G_A$ and $G_B$ corresponding to KeyStore$_A$ and KeyStore$_B$ respectively. In particular, the authentication process of the public-key of a node A by another node B and vice versa. It is based on a chain of correct and not expired certificates between A and B in the graph resulting from the union of the two key stores because:

- The first certificate in the chain can be verified by A (respectively B) because it was signed by itself.
- Each of the other certificates in the chain can be verified with the public-key of the previous certificate.
- The last certificate is B’s public-key (respectively A).

Special packets are sent between users to authenticate each other. Among other information, they contain the data FC, source Pseudonym and destination Pseudonym. Figure 1 shows all the possible phases of interaction included in the proposed self-managed protocol for authentication between two nodes A and B.

![Figure 1. Self-Managed Authentication Protocol](image)

The phases are fully described below. The first phase is the discovering process, which includes part of the beacons sent by node A, and in particular the hash of the IDs stored in KeyStore$_A$. In the second phase of ZKP authentication, a node B who wants to communicate with A, asks A to send the list of IDs in its store. After that, B checks whether a common key X can be found in the joint KeyStore$_A$ ∩ KeyStore$_B$, and in that case both execute a mutual ZKP on the knowledge of X so that in the last phase both nodes are sure they can use the shared key X to exchange their temporal secret keys and key stores.

The above algorithm allows any node to authenticate another node as well as to exchange both secret keys to update both key stores. An implementation of the proposed authentication scheme has been performed using Microsoft Visual Studio in C#.

A client-server capable of multiple connections at the same time is implemented in each device. All signals about authentication and beacons are performed with UDP packets. Each client broadcasts beacons periodically to all connected devices in the network. Each beacon is formed by the following data:

"01," + thisIpAddr + "," + PSEU + "," + Ek1(ID1,KUid1,TimeStamp)

Before starting to use the device, the node needs information to communicate with other devices, and in particular a database with three tables is loaded. These tables keep data for a low number of users whose data (certificates and public key) are generated with the generator 3:

- certificateStore (idcolumn INT PRIMARY KEY, idA NTEXT, idB NTEXT, certAB BIGINT, certBA BIGINT, date DATETIME);
Incoming connections are managed on the server so that when one is received, the server checks the identity of the node that sent the packet. After that, it checks whether the node is already authenticated in the network, and if not, the authentication protocol begins. The procedure the nodes use to send and receive information as indicated in Algorithm 1 (Authentication Scheme).

5. CONCLUSION

This paper proposes a self-managed authentication method for VANETs, which does not require deploying any infrastructure on the road or any special equipment on vehicles, which allows a gradual introduction of VANETs without any investment. In order to make it possible, existing devices like mobile phones or laptops equipped with wireless communication and GPS can be used to implement the software corresponding to the algorithms here proposed. The main contributions of this paper are a self-managed node authentication protocol based on public/private key pairs and certificate graphs, and a vehicle discovering scheme with variable pseudonyms to protect node privacy and prevent vehicle tracking. Our proposal allows the use of a hybrid encryption scheme that combines secret and public-key cryptography, which can be used both to optimize resources and to enforce node cooperation by avoiding passive behavior of nodes.

ACKNOWLEDGEMENT

Research supported by TIN2008-02236/TSI, PI2007/005, BES-2009-016774 and ACIISI (BOC Number 60).

REFERENCES


ONLINE VERSUS OFFLINE CRITICAL GEOGRAPHIC INFORMATION SYSTEM

Jan Hodicky and Petr Frantis
Communication and Information System Department, University of Defense - 65 Kounicova str, Brno, 612 00
Czech Republic.

ABSTRACT

The article deals with problems of information flow drop-outs at Critical Geographic Information Systems (CGIS). CGIS has high demand on availability of input and output data and must provide data analysis in very short time. Online CGIS has all input data available in real time or very close to real time. If there is a data flow drop-out the Online CGIS goes to offline mode. Offline CGIS must use different approaches to overcome the shortage of current data. The experiment depicted in the article use constructive simulation connected with the Czech military CGIS to prove the idea of using simulation method to predict unit position in the real situation. Experiment results reveal the maximum time interval in which the predicted unit movement calculated by simulator is still suitable for operational use in military CGIS.

KEYWORDS

Geographic Information system, Command and Control System, Constructive Simulator.

1. INTRODUCTION

The first historic information about Geographic information System (GIS) domain and its computer support dates back to 1960. Organization such as US Bureau of the Census, The US Geological Survey and The Harvard Laboratory for computer graphics and Environmental Systems Research Institute were involved in the first GIS prototypes solution (Raju, 2006). From that moment many GIS definition was formed and there is still not unanimous consensus about it.

As en example the definition of (Aronoff, 1989) can be taken. GIS is a computer-based system that provides the following sets of capabilities to handle geo-referenced data:

- Input.
- Data management (data storage and retrieval).
- Manipulation and analysis,
- Output.

With the respect of previous GIS definition the term Critical GIS (CGIS) can be defined. CGIS is GIS with high demand on availability of input and output data and analysis from the very short time interval perspective. The main CGIS features are:

- Sensors or agents deployed at the area of interest are responsible for inputs of CGIS. The interval of information updates is very high.
- One of the sensors or agents is objects of high importance. The object position must be visualized almost in real-time.
- Quality and credibility of inputs and output data is key enabler to support decision making process of CGIS users.

As an example of the CGIS non-military application the Virginia Interoperability Picture for Emergency Response (VIPER) is cited. The Virginia Department of Emergency Management launched this system that is capable of monitor and respond to emergencies and disasters cases (ESRI, 2009).

The military CGIS enhanced the non-military CGIS requirements to have common operational picture (COP) that contains the friendly and enemy unit positions.
Czech Command and Control System (C2) is used as an example of the military CGIS (Jindra, 2006). This system was chosen to validate our experiment.

2. ONLINE AND OFFLINE CGIS

One of the unresolved issues in the CGIS domain is the situation with no connection capability between sensors or agents and CGIS server. It can be caused by many of reasons:

- Overloading of bandwidth.
- Destruction of agents or sensors.
- Collapse of communication infrastructure caused by attack, disaster, etc.

From that perspective the online and offline CGIS can be defined. Online CGIS has all input data available in real time or very close to real time. Offline CGIS doesn’t have vital input data currently available. Offline CGIS must use different approaches to overcome the shortage of current data.

Our experiment proves the concept of using the simulation in the case of offline state.

3. CZECH MILITARY COMMAND AND CONTROL SYSTEM

The Czech C2 system is based on military GIS that works with the digital geographic data in standard formats from Czech army geographic service or NATO. This GIS core combines geographic data visualization with military symbols and tactical graphics and offers 2D view as a standard visualization module. As a result of the military research project called “Virtual” there is a 3D presentation layer that can use the digital geographic data from C2 system and generates 3D terrain with 3D geographic objects (Prenosil, 2008). The visual output of this presentation layer is real-time 3D visualization of the terrain with aerial or satellite imagery, 3D models of buildings generated from vector shape files, polygon or full 3D tree representation of woods, forests and tree lines, 3D road network, power lines and obstacles.

This real-time 3D visualization of GIS data can be combined with tactical vector graphics and 3D tactical symbols from C2 system. This merged tactical picture is used by a commander for better understanding the operational situation as is shown on Figure 1.

![Figure 1. Operational situation visualization in C2 presentation layer](image)

4. EXPERIMENT OBJECTIVE

The tactical situation depicted by tactical vector graphics and symbols is get from the C2 system in real-time so if the situation in the battlefield changes – the data in C2 systems are updated. If everything is working
the commander gets a real-time picture of the battlefield situation but if there is a communication line failure the information flow is interrupted and the tactical data are unsynchronized with the real world situation.

To overcome this situation a simple dead-reckoning algorithm can be used but this solution will work only for a short period of time (Yu, 2005). For longer time period it would create a false scenario because units represented by tactical symbols will continue moving according to their last known vectors of movement. In the real world a unit primary moves according to terrain and geographic objects and secondary according their task. So simple dead-reckoning algorithm would not be suitable because units would move through buildings, tree areas and it would not follow their assigned tasks. A solution to this problem would be using an artificial intelligence to drive the units when there is a communication problem. The artificial intelligence would use the known information – unit tasks, vector of movement and unit parameters (vehicle technical parameters, load, damage and so on) and would drive the unit according to this known state in respect of the terrain and geographic features and objects. To develop such an artificial intelligence system would be a huge amount of work but there is a solution that comes right out of a box. The solution is to use a military constructive simulator that would load the last known situation as a starting scenario and would perform a detailed constructive simulation to drive the units in advanced time. This simulated situation would send to our C2 presentation layer instead of real C2 data. The military constructive simulator would have the same terrain database as the C2 system so the unit movements would be affected by the actual terrain. The architecture of this complex system is depicted on Figure 2.

Figure 2. Interconnection between C2 presentation layer and constructive simulator

The difficult task to be solved is the interconnection between our C2 Presentation layer and military constructive simulator. The military simulators have the ability to be interconnected into distributed simulation. For the distributed simulation interconnection there is a standard called High Level Architecture (HLA) (NATO, 2009). We use the HLA architecture to exchange tactical situation data between military constructive simulator and our C2 presentation layer. Using the HLA interface allows using any suitable constructive simulator.

5. EXPERIMENT SOLUTION

To prove the concept described above we had to build an experimental workplace that consisted of:

- C2 system;
- C2 Presentation layer;
- Constructive simulator.

As a constructive simulator the MÄK VR Forces product was chosen. VR-Forces is a powerful and flexible simulation toolkit for generating and executing battlefield scenarios. It has all the necessary simulation features for use as a tactical leadership trainer, threat generator, behavior model testbed, or Computer Generated Forces (CGF) application.
The VR-Forces supports HLA standard for distributed simulation so we were able to connect our C2 Presentation layer with this constructive simulator. The important parts of the solution are correlated terrain databases. We had to generate the terrain database for the constructive simulator from the same source data as the C2 Presentation layer database. The VR-Force simulator contains a software toolkit that allows importing standard GIS digital geographic data and builds a terrain database. We used DTED format for the digital elevation model and ESRI Shapefile files for the vector features. The C2 Presentation layer and the VR-Forces constructive simulator were connected using the HLA standard and were able to interchange data about units and their tasks. We experimented with various scenarios. As an example we can take a moving convoy situation.

6. EXPERIMENT RESULTS

A convoy of 5 vehicles is moving in an area of interest. The convoy unit data are taken from the C2 systems and presented to the user by C2 Presentation layer. If the connection to C2 system is interrupted the C2 Presentation layer loads the last known situation using the HLA protocol to VR-Forces constructive simulator.

As the judging criteria an absolute position difference (APD) between simulated unit position and the real unit position (from the C2 system) was chosen. The designated goal was to get less than a 15% error in position difference. This value was got from the simple questionnaire that was filled by commanders at the battalion level. To be able to define APD variable the vehicle average velocity was set to 50 km/h. Thus the max error distance in the experiment was 4167m. It corresponds to 100% value of APD ratio error (APDRE).

The tested scenario contained only moving units without any interaction with other friendly or enemy units. This scenario was replayed 10 times and the results were compared with the real data from C2 system. Every replay began at different starting point on the terrain. The APD was compared each 30s. The averaged position difference ratio error can be seen on Figure 3.

As we can see from the Figure 3, the average position difference ratio error is 15% in 630 s interval. Time value of 630s was chosen as the maximum interval in which the predicted unit movement calculated by simulator is still suitable for operational use in C2 system. If the longer time interval had been used the
position error would have increased almost exponentially. It is caused by small differences between real terrain and digital terrain data. This small differences are multiplied every simulation step so in increasing time they are becoming more contributing for the position error.

The simulation works well for slow moving units – vehicles, infantry and armored units. For fast moving vehicles is the single simulation step difference more significant, so it cause bigger average position difference ratio error in the specified time period. Simulation can be employed only for ground forces units with relatively slow velocity.

7. CONCLUSION

This approach is more suitable for handling communication drop outs than simple dead-reckoning algorithms. Using the constructive simulator allows considering the terrain features for calculating the predicted unit positions. We allowed only simple moving tasks for the units without any interaction with the opposing or friendly forces in our tested scenarios. If we would allow so the artificial intelligence built-in constructive simulator would try to simulate the interaction and behavior of the unit that would not be in relation with the real orders in the real C2 system. But this drawback could be used as a feature for situations when the commander wants to simulate the situation that is currently in C2 system for longer period of time. In that case the constructive simulator can predict the advance of the battlefield situation and help a commander with decision making process.

REFERENCES


GRANULAR COMPUTING ALGORITHM TO DECREASE THE UNCERTAINTY OF KNOWLEDGE EXTRACTION

H. S. Alinia¹, M. R. Delavar², M. Zare³ and A. Mohseni⁴

¹ M.Sc. Graduate, GIS Division, Department of Surveying and Geomatics Engineering, College of Eng., University of Tehran, Tehran, Iran
² Center of Excellence in Geomatics Eng. and Disaster Management, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran, Tehran, Iran
³ Seismology Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, Iran
⁴ Research Associate, Seismology Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, Iran

ABSTRACT

Pre-determining locations and intensity of a seismic area is considered as a complicated disaster management problem. Since this problem depends on various effective criteria, one of the most important challenges concerned is the existence of uncertainty regarding inconsistency in combining influencing criteria and extracting more consistent knowledge for the next predictions. This paper proposes granular computing algorithm as a data mining approach to search for both more consistent and less uncertain rules from seismologic information table for 18 sample areas.

One of the significant properties of this method is induction of more compatible rules having zero inconsistency from existing databases. Furthermore, in this approach non redundant covering rules will be extracted for consistent classification where one object maybe classified with two or more non-redundant rules. This paper illustrate the advantages of using granular computing to discovery of knowledge from dataset consist of the seismic risk properties of the area between 58˚ 24' E, 60˚ 24' E Longitude and 27˚ 45' N, 29˚ 25' N Latitude around Reygan (Kerman Province), South-East of Iran where a devastating earthquake occurred.

KEYWORDS

Earthquake, Uncertainty, Data mining, Granular Computing Algorithm

1. INTRODUCTION

On December 26, 2003 when a major earthquake with magnitude 6.6 Mw (USGS, 2004) hit the south-eastern of Iran, Kerman province, at 05:26 AM (Iran standard time), the area most affected was the city of Bam where more than 43,000 people were killed, an estimated 30,000 injured and up to 75,000 left homeless, according to official estimates. Seven years later, another deadly earthquake struck its near area. The epicenter of this event which was happened at 22:11(local time) on December 20, 2010 with magnitude 6.2 has been located at 28.36N, 59.22E (IIEES) approximately 52 kilometer of south of Khavari in Mohammad-abad Reygan. According to reports, 1150 residential area around Fahraj village were damaged nearly 30 to 70 percent. Although the resulted devastation of this natural disaster can be mitigated through effective disaster management strategies such as pre-disaster risk assessment and effective post-disaster response, it depends on various effective criteria. One of the most important challenges concerned is combining influencing criteria and extracting more consistent knowledge for the next predictions. To do this, it is important to granulate areas as objects based on their indistinguishability, similarity, proximity, or functionality in which each has similar properties. In this way, an inductive method is proposed and applied to the given values of attributes of some training objects (areas) to determine appropriate rules.

Problem solving with different granularities have been explored in many fields such as Zadeh (Zadeh, 1997) introduced a fuzzy logic view of information granulation and human reasoning, Pedrycz (Pedrycz, 2001) used fuzzy-granular computing in data mining application, Slowinski’s team also have done some researches on multi-criteria decision making with rough sets (Greco et al., 2000; Slowinski et al., 2002)
Yao and Yao presented a granular computing view to classification problems and proposed a granular computing approach for classification (Yao and Yao, 2002). Also this concept was used in assessment of seismic vulnerability of Tehran, Capital of Iran (Samadi Alinia, 2010; Samadi Alinia and Delavar, 2010).

2. OVERVIEW OF GRANULAR COMPUTING

“The basic ideas and principles of granular computing are not entirely new and have indeed been investigated in many disciplines of social and natural sciences” (Yao, 2005). Two essential tasks in data mining are the representation of objects and the identification of forms and types of knowledge to be mined. Broadly, granular computing can be studied based on the notions of representation and process, which were also used by Marr in the study of vision (Marr, 1982). Many computational operations can be performed on granules such as reasoning, inferring and learning. Elements in a granule are grouped together by indistinguishability, similarity, proximity or functionality (Zadeh, 1997).

This paper exemplified data mining especially rule-based mining in two steps; the formation of concepts and the identification of relationship between concepts. From the stand point of granular computing in this research, the concept of seismic risk assessment can be exemplified at two parts, extension, i.e., a set of objects as instances of pre-species category of seismic risk and intension which consists of all properties or attributes with more effective impacts in occurrence of earthquake, that are valid for all those areas where the concept applies. In this approach, each object is represented by the values of a set of attributes and the knowledge mined from a sample dataset is illustrated in the form of rules.

2.1 Information Table

Information tables are base knowledge in granular computing models. An information table provides a convenient way to describe a finite set of objects called a universe by a finite set of attributes (Pawlak, 1991; Yao and Zhong 1999).

Definition of an information table is shown in the following tuple (Pawlak, 1991):

\[ S = (U, At, L, (I_a \mid a \in At), (V_a \mid a \in At)) \]  

Where,  
- \( U \) is a finite non-empty set of objects,  
- \( At \) is a finite non-empty set of attributes,  
- \( L \) is a language defined by using attributes in \( At \),  
- \( V_a \) is a non-empty set of values of \( a \in At \),  
- \( I_a : U \rightarrow V_a \) is an information function that maps an object of \( U \) to exactly one possible value of attribute “\( a \)” in “\( V_a \)”.

“With respect to all attributes in \( A \), \( x \) and \( y \) are indiscernible, if and only if they have the same value for every attribute in \( A \)” (Yao, 2004). Therefore, a language \( L \) is defined for describing objects of the universe in an information table. In this research, the decision logic language (DL-language) studied by Pawlak (Pawlak, 1991) is adopted. This logic language is defined for the information table to provide formal descriptions of various notions.

The information table of this research is constructed by existing knowledge from 18 seismic areas around Mohammad-Abad, arranged in 18 rows of objects and 6 columns of attributes.

The six attributes which are considered in this paper include: the rate of seismic risk of the area, the existence of fault in the area and its length, fault coincidence in that area, occurrence of historical earthquake with magnitude more than 6 Richter and occurrence of historical earthquake with magnitude between 4 and 6 Richter. It is assumed that there is a unique attribute class taking class labels as its value. The set of attributes is expressed as \( At = F \cup \{ \text{class} \} \), where \( F \) is the set of attributes used to describe the objects and \( \{ \text{class} \} \) is the decision attribute. The goal is to find rules from existing training dataset in the form of \( \phi \Rightarrow \text{class} = q \), where \( \phi \) is a formula over \( F \) and \( q \) is a class label. In this paper, four classes based on the occurrences of earthquake during 20th century and the last six months are considered as values of the decision attribute. For simplification, four considered classes are labelled as below:

- \( A \)= Earthquake with Magnitude more than 6 happened during 20th century
B= Earthquake with Magnitude less than 6 happened during 20th century
C= Non occurrence of earthquake
D= Earthquake with Magnitude between 3 and 4 during the last six months
Table 1 illustrates the information table used in this research. For simplification, the titles of all criteria are given as follows:

Seis_Rate  The rate of seismic risk of the area
Fault      The existence of fault in the area and its length (L)
Fault_coin. Fault coincidence in the area
More6.     Occurrence of historical earthquake with magnitude more than 6 Richter
Bet4_6.    Occurrence of historical earthquake with magnitude between 4 and 6 Richter in previous decade

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>Event existence</td>
<td>L&gt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>6 Event</td>
<td>A</td>
</tr>
<tr>
<td>U2</td>
<td>Low occurrence- fault</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>1 Event</td>
<td>C</td>
</tr>
<tr>
<td>U3</td>
<td>Non event with fault in surrounding</td>
<td>L&lt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>1 Event</td>
<td>C</td>
</tr>
<tr>
<td>U4</td>
<td>Low occurrence- fault</td>
<td>3(L&gt;80)&amp;20&lt;L&lt;80</td>
<td>Existence</td>
<td>Event existence</td>
<td>Nothing</td>
<td>A</td>
</tr>
<tr>
<td>U5</td>
<td>Low occurrence- fault</td>
<td>3(L&gt;80)&amp;10&lt;L&lt;20</td>
<td>Existence</td>
<td>2 Event</td>
<td>Nothing</td>
<td>A</td>
</tr>
<tr>
<td>U6</td>
<td>Event existence</td>
<td>2(L&gt;80)</td>
<td>Nothing</td>
<td>Nothing</td>
<td>7 Event</td>
<td>B</td>
</tr>
<tr>
<td>U7</td>
<td>Low occurrence- fault</td>
<td>L&gt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>2 Event</td>
<td>B</td>
</tr>
<tr>
<td>U8</td>
<td>Non event with fault</td>
<td>20&lt;L&lt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>C</td>
</tr>
<tr>
<td>U9</td>
<td>Non event with fault</td>
<td>3(L&gt;80)</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U10</td>
<td>Low occurrence- without fault</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>1 Event</td>
<td>D</td>
</tr>
<tr>
<td>U11</td>
<td>Low occurrence- fault</td>
<td>L&gt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U12</td>
<td>Low occurrence- without fault</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>1 Event</td>
<td>D</td>
</tr>
<tr>
<td>U13</td>
<td>Non event with fault in surrounding</td>
<td>4(L&gt;80)</td>
<td>Existence</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U14</td>
<td>Low occurrence- without fault</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Event existence</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U15</td>
<td>Low occurrence-fault</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U16</td>
<td>Event existence</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U17</td>
<td>Non event with fault in surrounding</td>
<td>L&lt;2</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
<tr>
<td>U18</td>
<td>Low occurrence- fault</td>
<td>L&gt;80</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Nothing</td>
<td>B</td>
</tr>
</tbody>
</table>

2.2 Concept Formation

The algorithmic issues of granule construction are obtained by representation and formation of granules. They address the problem of how to put two objects into the same granule. A formula has a meaning if it has an associated subset of objects.

To illustrate the idea developed so far, consider an information table given by Table 1. The following expressions are some of the formulas by the language L:

- Fault =2A,
- Fault_coin = nothing ∧ more6 = Event existence

The meanings of the above formulas are given by:

- m (Fault =2A) = {u6}
- m (Fault_coin = nothing ∧ more6 = Event existence) = {u14}

In these cases, the involved granules are {u6}, {u14}. To achieve these granules has not unique formulas. That is, there may exist two formulas such that m (ϕ) = m (ψ) (Yao et al., 2005). For example; Fault =2A and Fault_coin = nothing ∧ Fault =2A have the same meaning set {u6}.

One need to extract one of the possible solutions based on his/her understanding and preference. In the proposed algorithm, given a set of data, each user will try to make sense of data by observing it from different angles, in different aspects, and under different views. The preference can be stated order in granular computing. In the first step of the proposed algorithm, all possible granules from given database is
constructed by an atomic formula which is given by \( a = v \), where \( a \in \mathbb{A} \) and \( v \in \mathbb{V} \). Once the granules constructed, it is necessary to describe, to name and to label granules using certain languages. Also, a measurement is employed on each single granule that is named generality. It indicates the relative size of the granule. Indeed, a granule defined by the formula is more general if it covers more instances of the universe.

2.3 Relationship between the Concepts

Based on the notions introduced so far, data mining for rules can be viewed as searching for relationship between the overlap concepts (Yao, 2001). By expressing rules with intensions of concepts, we may easily explain them in a natural language, provided that we can explain formulas of the language L. Therefore, a crucial issue is the characterization, classification, and interpretation of rules. It is reasonable to expect that different types of rules represent different kinds of knowledge derivable from a database. To achieve this goal, different quantitative measures and different mining algorithms can be designed. In many studies of machine learning and data mining, a rule is usually paraphrased by an “IF - THEN” statement, “if an object satisfies \( \phi \), then the object satisfies \( \psi \).” The interpretation suggests a kind of cause and effect relationship between \( \phi \) and \( \psi \). Basically, another issue to construct rules from input dataset is to identify relationship between two granules and relationship between a granule and a family of granules. So, to induct more confident relationship between existence concepts, three quantitative measures are implemented in this research as follow:

i) Confidence or absolute support: It is a measure of the correctness or the precision of the inference. The quantity can be computed by a fraction of number of samples in a granule that is equal to objects belong in a class, to the size of the granule (Samadi Alinia, 2010).

ii) Coverage: It is a measure of the applicability or recall of the inference and indicates fraction of data in a class correctly classified by the atomic formula (Yao and Yao, 2002). The quantity can be computed by fraction of number of samples that satisfy the THEN part of the rule, to the size of data with the same class label as the rule consequent.

iii) Conditional entropy: it provides a measure that is inversely related to the strength of the inference. This measurement depends on the confidence and one can identify one equivalent class in which the object belongs with no measure of uncertainty where an object satisfies the formula of attribute-value. In this case, confidence of the formula for at least one equivalent class is 1.

3. INDUCTION OF MORE CERTAIN RULES

To construct a granule network it is required that first dividing the universe into grouping or partitions of the same class with atomic formula of attribute-values. In the proposed algorithm, three mentioned measures on relationship between the concepts are applied automatically. The algorithm will continue until all objects are correctly classified in which, each object is associated with a unique class label. It is important to find a subset of attribute-value with high coverage, confidence and minimum entropy. In the proposed algorithm, construction of the tree are continued until all granules at the end level reach to zero entropy in which, union of all non-active granules be equal to the universe set. After union of all non-active granules constructed from a covering solution of the universe set, construction of the decision granule tree would be stopped. This can identify the information that is provided by the constructed tree in the form of “IF - THEN” statements.

In this research, from 18 seismic areas, 36 different nodes were added to the granule network tree in the 5 levels. Among these extracted nodes, based on non-redundant covering solutions and user preferences, a user can freely explore the dataset according to his/her preference and priority. It is important to note that user can process various skills, intelligence, cognitive styles, frustration tolerances and other mental abilities. To give an example from the extracted rules in this part of research, 4 different rules which classified same objects are presented below:

If Fault=3(L>80) \( \rightarrow \) \{u9\} (class B)
If Fault_coin= Nothing and Fault= 3(L>80)) \( \rightarrow \) \{u9\} (class B)
If Fault_coin= Nothing and More6 = Nothing and Fault=3(L>80) \( \rightarrow \) \{u9\} (class B)
If Fault_coin= Nothing and More6 = Nothing and Fault= L<2 and Bet4_6= Nothing and Fault=3(L>80) \( \rightarrow \) \{u9\} (class B)
However, all of these kinds of rules have minimum uncertainty and provide consistent classification the first rule is considered more general than others. That is, it gives general information. But in some solutions, one needs to examine the problem at a finer granulation level with more detailed information when there is a need or benefit for doing so.

4. CONCLUSION

This paper has proposed a new approach to induct the classification rules with less uncertainty for seismic risk assessment. The extracted rules with zero entropy and high accuracy can be evaluated by experts and selected based on their needs and the existence situations. These inducted rules and their information along may be used for prediction of the next events with near or same circumstances based on the rule accuracy. This paper is the result of part of the researches which are examined to find methods for knowledge representation, searching, and reasoning. It is the first step to use broader meaning for hierarchies, instead of the restricted mathematical notion defined by a partial ordering in which this paper shows the combination of the theory of hierarchy and the systems thinking, as well as taking advantages of both.

REFERENCES

MOBILE DETECTION OF SOCIAL SITUATIONS WITH TURN-TAKING PATTERNS

Georg Groh, Alexander Lehmann and Manuel de Souza
TU-München, Institut für Informatik, Germany

ABSTRACT
This contribution discusses the detection of co-located Social Situations as a form of social context usable in Mobile Social Networking services. We aim at detecting Social Situations on the basis of turn-taking patterns from audio signals which result from conversation recorded with mobile devices. The proposed method exploits the coordinated structure of turn-taking patterns of persons in Social Situations compared to the uncoordinated turn-taking patterns of pseudo-pairs resulting from inevitable cross-talk in crowded social spaces. We use an HMM based classifier and a manually annotated corpus of conversation to distinguish conversation between interacting pairs of persons from cross-talk between non-interacting persons.

KEYWORDS
Mobile Social Signal Processing, HMM, Turn Taking Pattern, Social Situation Detection, Speaker Diarization.

1. INTRODUCTION

Social Signal Processing is an upcoming field aiming at understanding social interaction and social behavior on the basis of measuring and analyzing social signals and behavioral cues using methods derived from signal processing and machine learning [Vinciarelli et al.2009]. The closely related field of Reality Mining also encompasses analysis of individual behavior and suggests the use of mobile devices as sensors for measuring behavioral cues [Pentland et al.2009]. In this contribution, instead of aiming at understanding social interaction, we aim at detecting social interaction in the form of Social Situations. While detection of Social Situations using geometry of interaction has been successfully demonstrated [Groh et al.2010a], we will now discuss Social Situation detection using audio data.

Our model of Social Situation represents a very simple form of social context that can be used (together with e.g. individual context elements) in future (Mobile) Social Networking services such as Social Life-logging, context-aware privacy definitions, etc. [Groh et al.2010b]. In contrast to the large body of other work on Social Signal Processing [Vinciarelli et al.2009], our model / definition of Social Situation and our proposed methods for instantiating this model solely focuses on the (detection of the) existence of social interaction of a certain kind, without aiming at a deeper level understanding of the nature of this interaction (see also [Groh et al.2010a]). A Social Situation $S=(P;X;K)$ is defined as a three tuple of a set $P$ of $n\geq 2$ persons in a state of co-located, ongoing, face-to-face social interaction with full mutual awareness of this social interaction, a connected spatio-temporal region $X$ that this interaction occurs in, and a simple keyword-based model $K$ of the semantics of $S$. Full mutual awareness implies that the existence of this social interaction is common knowledge [Schelling1960] [Lewis1969], [Aumann1976]) among $P$. We exclude social interaction settings where the interaction is e.g. electronically mediated (e.g. a chat) and not co-located or not face-to-face, respectively. It is the overall goal of our research to develop robust methods that use data that can be acquired by sensors in mobile devices such as smart-phones for the detection of the existence of social situations, thus neglecting their semantics for the moment. The demand of full mutual awareness of the interaction in the upper definition allows for a quite precise conceptual demarcation of the sets $P$ and $X$. We assume that, once partial awareness of a looming social interaction has started to evolve, people quickly strive towards either establishing full mutual awareness via “starter actions” (interaction-establishing rituals such as greeting) or more implicit social signals (e.g. certain types of looks) and by referring to a commonly accepted communication mode (e.g. non-verbal flirting) or force denial of interaction (see [Goffman1966] p.
92). That is, human beings are usually very precisely aware with whom they are currently interacting and with whom they are not interacting.

2. RELATED WORK

Concerning the term and the concept of Social Situation, numerous contributions from sociology and social psychology exist (see e.g. [Argyle et al.1981]). Goffman was one of the first authors to emphasize the importance of studying social situations. In [Goffman1961] p. 144, he defines: “By the term Social Situation I shall refer to the full spatial environment anywhere within which an entering becomes a member of the gathering that is (or does then become) present. Situations begin when mutual monitoring occurs and lapse when the next to last person has left”. In [Goffman1966] (p. 33ff, 83ff) he further distinguishes unfocused and focused types of interaction, the latter “[...] concerned with clusters of individuals who extend one another a special communication license and sustain a special type of mutual activity that can exclude others who are present in the situation.”, while the former is “[...] concerned with what can be communicated between persons merely by virtue of their presence together in the same social situation”. Our definition thus predominantly aims at focused interaction at least as far as unfocused interaction violates the condition of full mutual awareness.

Our approach focuses on turn-taking patterns in conversations. The scientific field that provides methods for inferring these patterns from audio recordings is Speaker Diarization [Tranter and Reynolds2006]. The proposed methods are quite complex and often fine-tuned to specific situations such as broadcast news, telephone conversations and meetings [Kadri et al.2010], and to specific conditions such as the number of microphones, recording quality, noise level, etc. Their results may vary considerably if transferred from their original domain to another domain [Istrate et al.2005]. The main steps in Speaker Diarization systems usually consist of change detection, speaker clustering and speaker classification. In change detection, speaker changes are detected, e.g. by using Bayesian Information Criterion (BIC) to decide whether the pattern-vector of audio features (see e.g. [Ajmera2004] [Campbell et al.2003]) of a given frame is better modeled by a single speaker model (e.g. a single Gaussian) or by a model for two (e.g. two Gaussians) or more speakers (see e.g. [Ajmera2004], [Tranter and Reynolds2006], [Kotti et al.2008], [Zhu et al.2005]) with subsequent refinement in order to detect the actual point of change in time more precisely. Other methods use KL-divergence on GMM fitted to the patterns of each frame to detect changes [Barras et al.2004][Tranter and Reynolds2006]. For clustering the found segments, usually agglomerative hierarchical clustering approaches are used, where clusters are connected on the basis of appropriate distance measures. The resulting dendrogram is cut on the basis of cluster validity criteria, e.g. BIC based [Kotti et al.2008][Barras et al.2004] to yield the optimal number of clusters. Speaker classification (if desired) is then performed using training data from each speaker. Many other approaches exist that propose subtle variations and improvements. Other approaches aim at combining the distinct steps described before (e.g. [Anguera et al.2006][Meignier et al.2006], using HMMs). As has been said before, most speaker diarization systems have problems when they are applied to settings for which they have not been fine-tuned. Most systems assume that at most one speaker is speaking at a time. Cross-talk [Tranter and Reynolds2006] creates problems, e.g. in the speaker segmentation step, which has not yet been satisfactorily solved [Boakye et al.2008]. However, for Social Situation detection, especially in crowded urban spaces, we usually face settings with multiple situations happening at the same time, so heavy cross-talk is a common phenomenon. Even in non-parallel conversations an amount of 10% cross-talk is cited as normal on average [Vinciarelli et al.2009]. Most standard methods that worked well for clean mixtures failed in our experiments when applied to realistic recordings with cross-talk (‘Cocktail Party Problem’ [Haykin and Chen2005]. Other systems work very well when n (or almost n) microphones for n audiosources (speakers) are available, such as ICA-based methods [Hyvärinen and Oja2000] [Lee et al.1999]. However, for Social Situation detection with mobile devices the mixing matrix is typically heavily time-dependent because microphones are also moving, so standard ICA methods have not yielded satisfying results on realistic recordings. As of now, no current speaker diarization approach is able to deliver absolutely reliable turn-taking patterns for Social Situation detection for a broad range of settings. For that reason we rely on manually annotated turn-taking patterns for our investigations. However, regarding the performances cited in the mentioned publications, we assume that future systems will be able to deliver turn-taking patterns for our scenarios with the required performance.
A **Turn** may be intuitively defined as a continuous speech utterance by a person or, more precisely, a 'stretch of speech uttered by one speaker that consists of one or more utterances' [Ten Bosch et al.2004], where an utterance is defined as 'the sequence of words between punctuation marks [...]'. A fine-grained discussion of a suitable definition of 'continuous' in the upper definition may be found in [Weilhammer and Rabold2003] and the related works by Bosch et al. (e.g. [Ten Bosch et al.2004]). On the basis of earlier work on turn-taking patterns, Weilhammer & Rabold and Bosch et al. describe a variety of basic types of turn-taking patterns and verify on different corpora that the logarithm of pauses and of overlaps in conversations each roughly follow Gaussian distributions. Depending on the language, [Weilhammer and Rabold2003] report average lengths between 363 ms and 389 ms for pauses and between 155 ms and 331 ms for overlaps, while the average turn length is found to be roughly 8 seconds for English and 5 seconds for German and Japanese for one of the backchanneling turn-taking patterns, which is of three typical patterns which encompass 96 % of all turn-taking cases [Weilhammer and Rabold2003]. These figures give valuable hints for choosing appropriate sequence lengths for our analysis.

### 3. APPROACH

In order to develop and evaluate approaches for detecting Social Situations on the basis of turn-taking patterns won from audio data, **annotated audio corpora** are necessary since (as was discussed in section 3.2) state of the art diarization systems cannot cope well enough with realistic settings to deliver automatic annotations for real-life social interaction audio data. In order to render this demand more precise, the following requirements should be met: (1) Audio data should correspond to realistic, everyday settings which are not too special and contain several, co-located, face-to-face social interactions (Social Situations) with frequent oral conversation and cross-talk. (2) Sound quality and noise-level (while still realistic) should allow for accurate manual annotation. (3) The setting should contain several concurrent social interactions in order to have, at a given point in time, pairs of participants in a Social Situation as well as pairs not in a Social Situation. Most of the annotated corpora that are used in literature such as VERBMOBIL II [Weilhammer and Rabold2003], AMI meeting corpus [AMI2007] or SLAAP [Kendall2007] reflect special properties and violate at least demands (1) and (3) from above. Thus we decided to record and annotate a small **dataset fulfilling the requirements** from above. We used a single standard voice recorder. The basic setting (a typical time frame) of the social interaction setting we recorded is shown in figure 1 (right part). The interaction lasted for 9 minutes. Social Situation 2 was stable throughout the whole duration whereas situation 1 at the table temporarily dissolved and reformed again with short participations from speaker 5. Another 7 minute setting with one continuous Social Situation involving three speakers was also recorded to provide more examples of speaker combinations in a Social Situation. We manually annotated the audio data using a frame-length of 50 ms. The average length of speech pauses is crucial for deciding on the frame-length, since a fine-grained annotation should be able to accurately map these pauses. Complementing the cited results of section 3.3, e.g. [Campione and Veronis2002] found an average pause length between 490 ms and 587 ms for different European languages. Furthermore, the frame-length should reflect the minimum length of pauses which can still be human-annotated in a reasonable way. This minimum length is claimed to be 60 ms [Kendall]. Our frame-length should thus allow for annotating all pauses and speaker-changes in a sufficiently accurate way, without claiming total accuracy beyond the cited limit of 60 ms.

Settings with many concurrent social situations and thus with lots of cross-talk will, however, have to analyze the correlation patterns in the resulting turn-taking patterns to be able to separate those situations. **Figure 1** (right part) visualizes the problem: Although person 2 and person 4 do not socially interact and thus are not in a common Social Situation, the interpersonal distance between both persons is smaller than to some of those persons interacting with them. A microphone of a mobile device in person 2’s pocket may thus record substantial cross-talk from situation 2 (involving persons 4 and 5). However, the turn-taking patterns of persons within a Social Situation are usually correlated, while persons not socially interacting will show uncorrelated turn-taking patterns. Instead of investigating all $2^m$ subsets of $m$ audible speakers as Social Situations, our approach investigates each **pair** $(s_{i1}, s_{i2})$ of audible speakers and proposes probabilities $p_{\oplus}(s_{i1}, s_{i2}, t)$ of the pair being in a Social Situation and $p_{\ominus}(s_{i1}, s_{i2}, t)$ of the pair not being in a Social Situation at time $t$. We can then use standard graph clustering on the resulting weighted graphs to determine the Social
Situations (possibly integrating in each edge evidence for social interaction between each pair won from other data sources such as geometry of interaction [Groh et al.2010a]).

In order to determine for a point in time \( t_{jm} \) whether two persons are likely to socially interact at that given point, we need to monitor the immediate interaction history of these two persons in \([t_{jm-s}, t_{jm}]\) with a suitably chosen sequence length \( s \). The time interval should be short enough to allow to accurately model dynamic changes in Social Situations and it should be long enough to allow for an accurate distinction between social interaction and non-interaction, which are conflicting issues. Therefore (considering the results cited in section 2) we varied \( s \), allowing for several turns to occur and to map the most important turn-taking patterns described in [Weilhammer and Rabold2003] while still providing enough sequences considering the limited extension of our dataset. The results of this variation will be discussed below. Choosing \( s=10s \) and considering each possible pairing of persons yields \( 10^{3} \oplus \) sequences entirely corresponding to pairs of people in Social Situations, \( 444 \ominus \) sequences entirely corresponding to pairs of people not in a Social Situation and 70 mixed sequences. Due to the small number of available sequences we used a 65-35 cross validation-style schema with 50 runs and random selection of the partition for our experiments.

We trained a standard Hidden Markov Model \( \lambda \oplus \) with the \( \oplus \) training sequences and another HMM \( \lambda \ominus \) with the \( \ominus \) training sequences, using the EM-based Baum-Welch algorithm [Rabiner1989] available through JaHMM [Francois2011]. We then used the Forward-Backward algorithm [Rabiner1989] to provide the probabilities \( p\oplus(O) = p(O \mid \lambda\oplus) \) and \( p\ominus(O) = p(O \mid \lambda\ominus) \) for each test sequence \( O \). To estimate the basic performance, we assigned the class on the basis of which probability was larger. The HMMs use 4 hidden states, intended to model "\( s_{1} \) speaking", "\( s_{2} \) speaking", "both speaking", as well as "no one speaking". For the initial state probabilities \( \pi \) of the states we used the relative frequencies of the overall experiment. For the initialization of the 4 x 4 transition-probability matrix \( A \), we used the relative frequencies of the corresponding transitions in the \( \oplus \) sequences (or \( \ominus \) sequences, respectively). We chose 4 observation states with the same semantics as the hidden states, effectively using convergence of the EM-procedure to a normal Markov-model with a trivial sensor model \( (B = 1) \) as a means of control. For the initialization of the 4 x 4 sensor model we used a matrix with 0.9 on the diagonal and evenly distributed off-diagonal entries, which indeed after learning (modulo a permutation of the states) converged to \( B = 1 \) as desired. As a measure of convergence we computed the KL-divergence between \( A_{t} \) and \( A_{t+1} \), revealing quick convergence after a few EM-iterations while still maintaining a non-zero KL divergence between \( A_{\oplus} \) and \( A_{\ominus} \) which is a hint that both structurally equivalent HMMs do not converge to an identical state which would render them useless for classification.

Averaging our 50 runs, for this pairwise social interaction classification, we get Accuracy=0.713, Precision=0.702, and Recall=0.774. While this result does not allow to distinguish our two classes with absolute reliability, it is nevertheless able to give non-trivial evidence, which may be combined with evidence from other data-sources (such as geometry of interaction) to yield a more complete picture. If we also use the mixed sequences (if more than 50% of the time slices in a mixed sequence correspond to one class, then the sequences is assigned to that class) the overall averaged accuracy does not change.

Figure 1. Right part: Basic setting of recordings (typical time frame). Left part: Problematic setting for non-turn-taking based based approaches
significantly. Varying the length of the sequences shows a saturation effect of the averaged accuracy towards longer sequence lengths. For our case, a sequence length between 10 and 20 seconds is a reasonable compromise.

As we have seen, with the HMM based method it is possible to provide probabilities $p(\oplus (s_{i1}, s_{i2}, t) = p(O(s_{i1}, s_{i2}, t) \mid \lambda(\oplus))$ and $p(\ominus (s_{i1}, s_{i2}, t)$ that allow to decide for a given pair of persons $(s_{i1}, s_{i2})$ at a given point in time $t$ on the basis of sequences of turn-taking patterns with a reasonable accuracy whether those two persons are in a Social Situation or not. We will now finally shortly discuss the question having been mentioned before, namely whether it is possible to reconstruct n-ary Social Situations (Social Situations with $n = |P| > 2$) from these probabilities. For that task we use the dataset obtained in an earlier experiment, where the goal was to investigate and provide these probabilities on the basis of observing the geometry of interaction of pairs of persons (relative body angles $\delta\theta(s_{i1}, s_{i2}, t)$ and interpersonal distances $d\theta(s_{i1}, s_{i2}, t)$) [Groh et al.2010a]. In that experiment an average pair-wise classification accuracy of $\approx 0.75$ using a GMM-based classifier was achieved. We trained separate GMMs $g(\oplus)$ and $g(\ominus)$ for the $\oplus$ and $\ominus$ cases, each providing probabilities $p(\oplus (s_{i1}, s_{i2}, t) = p(O(s_{i1}, s_{i2}, t) \mid \lambda(\oplus))$ and $p(\ominus (s_{i1}, s_{i2}, t)$ accordingly. Compared to the present experiment where $O(s_{i1}, s_{i2}, t_{ml})$ is given by the respective turn-taking pattern in $[t_{ml-1}, t_{ml}]$, the geometry experiment used point-like observations $O(s_{i1}, s_{i2}, t_{ml}) = (\delta\theta, d\theta) (s_{i1}, s_{i2}, t_{ml})$ because using time series of observations and HMM based classifiers did not turn out to provide an increased accuracy. Body language will intuitively allow for point-wise estimations quite accurately while a certain duration of analyzing turn-taking is necessary to provide a meaningful decision ground. We chose to use this dataset (annotated in a similar way) for the investigation of the question of this section because its significantly larger extension (30 minutes of interaction, nine participants, 34 different Social Situations). At each point in time, a complete weighted graph $G = (V,E,w,t)$ may be constructed with the set of participants $f_{\{s\}}$ as nodes $V$ and the weights $w(s_{i1}, s_{i2}, t) = p(\oplus (s_{i1}, s_{i2}, t) \mid (p(\oplus (s_{i1}, s_{i2}, t) + p(\ominus (s_{i1}, s_{i2}, t)$). Aside from being proportional to the strength of positive evidence, the weight-expression thus also takes into account the strength of the negative evidence and the proportion between the two. Using average link graph clustering [Jain and Dubes1988] and Maximum Modularity [Newman2004] as a cluster validity criterion to decide upon the number of clusters on this graph yields a partition of the graph into social situations where nodes in singleton clusters are not considered to be in a Social Situation. Comparing this partition at each ‘point’ in time (frame length was $\approx 180$ ms) to the manually annotated true partition in that frame with the help of standard RAND index $R$ [Rand1971] yields a mean over all frames of $<R> = 0.766, \sigma=0.198$. Random partitions as a base line yield $<R>_{\text{random}} = 0.524, \sigma=0.233$ and the Adjusted RAND index which compensates for the non-zero expectation for random partitions (see [Hubert and Arabie1985]) yields $<R_{\text{adj}} > = 0.529, \sigma=0.374$. While the increase for $<R>$ against the baseline does not seem to be very great at first glance, the high value for the Adjusted Rand index (which has a baseline of $<R_{\text{random}} > = 0.022$ for random partitions) shows that indeed a very good match between the computed partitions and the annotated partitions exists. This shows that the detection of larger than binary Social Situations is possible with satisfactory accuracy if $p(\oplus)$ and $p(\ominus)$ probabilities for pairs with a predictive accuracy of roughly 0.75 are used. We can thus expect a similar performance for an analogous Social Situation detection process using the $p(\oplus)$ and $p(\ominus)$ won on the basis of turn taking as described in this contribution.

4. CONCLUSION

We proposed and investigated an HMM-based method to successfully distinguish turn-taking patterns within Social Situations from turn-taking patterns resulting from cross-talk. From our investigations, it can be concluded that the $\lambda(\oplus)$ HMM (in comparison to the $\lambda(\ominus)$ HMM) (both possibly trained on larger annotated corpora) can capture sufficient features of essential turn-taking patterns for this distinction and may readily be used in mobile devices to deliver evidence for the detection of Social Situations. In future work we will continue to investigate methods for integrating multiple evidences and specialized methods for speaker diarization for providing the necessary turn-taking patterns in crowded social spaces.
REFERENCES


A SERVICE DISCOVERY PROTOCOL FOR WIRELESS MESH NETWORKS

Omowonuola Saka and Chris Guy
School Of Systems Engineering - University Of Reading - Reading, RG6 6AY, UK.

ABSTRACT
Wireless Mesh Network (WMN) are an emerging wireless technology which will play a vital role in the next generation wireless network. Wireless mesh networks can be viewed as a specialized type of wireless ad hoc network. This paper focuses on background study of Wireless Mesh Networks as a whole to identify the major research issues in the network architecture with emphasis on SDP (Service Discovery Protocol), to identify the strength and weaknesses of existing SDP architectures proposals and standards in order to propose a suitable SDP architecture for WMN. The proposed SDP which is IETF's Service Location Protocol (SLP) is adapted to WMN using NS2 network Simulator in order to measure its suitability and performance.

KEYWORDS
Wireless Mesh Networks (WMN), Service Discovery Protocol (SDP), Service location Protocol (SLP), mesh Enhanced SLP (mSLP).

1. INTRODUCTION

Wireless Mesh Networks (WMN) have emerged as a flexible and low-cost alternative to the traditional wireless networks via multi-hop communications, supporting applications from last-mile Internet delivery, broadband home networking, enterprise networking, community wireless networking to building automation. According to [1], a WMN can be defined as being dynamically self-organized, self-configured and self healing, with the nodes in the network automatically establishing and maintaining mesh connectivity among themselves (creating, in effect, an adhoc network). Wireless mesh networks are expected to provide solutions to the drawbacks and improve the performance of existing wireless networks like WLAN, MANET, WPAN and MAN. Wireless mesh networks aim at ensuring connectivity despite user mobility and problems in the network media. For an introduction to WMNs see [1].

2. SERVICE DISCOVERY

Service discovery in simple term refers to the process through which services are discovered in a network. A service in this context can be a software service like an application hosted on a particular device which is available for use by the other devices or a hardware service like a printer, a scanner or a fax machine [3] [4]. Other services could be data operations like data backup or data transfer and internet services. A typical service discovery scenario is a client server network where a client establishes a connection with the server and either searches for the services the server offers or automatically discovers these services and capabilities which it can benefit from. In a typical network there are several servers and several clients communicating either via wired or wireless technology, services need to be discovered and accessed via a certain process or procedure this brings about the need for Service discovery protocols. Service discovery protocols have been developed to simplify and automate service discovery in networks.

There are two types of service discovery architectures; the directory-less service discovery architecture where clients send requests for services directly to the server device providing a service and the directory based architecture where information about devices offering services and the services they offer are stored in.
a centralized directory, clients do not send requests directly to the server but requests to the service directory to get information about available services. The majority of the existing service discovery protocol proposals for wireless networks are targeting WLANs, MANETs with little or no proposals for WMNs. Since WMNs are most similar to MANETs, the SDPs to be analysed in the course of this work will be the standard service discovery protocol and service discovery protocols proposed for MANETs analyzing their suitability for use in Wireless mesh networks.

3. PROPOSED SERVICE DISCOVERY PROTOCOL

IETF’s Service Location Protocol (SLP) was selected after the examination and analysis of the existing SDPs carried out. Some of these protocols are standardized and commercialized while others are proposals which are still under construction. SLP seemed the most appropriate service discovery standard to base the proposed solution on due to the fact that it met most of the requirements identified for an appropriate SDP for WMNs.

- SLP in comparison to the other standards is a simple and lightweight protocol which does not depend on any other protocol to work and it preserves the modular OSI approach.
- SLP is a mature protocol with several extensions and implementations with its complete specification defined in RFC 2608.
- SLP eliminates the need for a user to know the name of a network node hosting/supporting a service, users search for services based on the service type and attributes.
- SLP can be deployed in small to large enterprise type networks and allows requesting clients to search for services based on their attributes or attribute values.
- SLP is has a hybrid service discovery architecture, this basically means users can send request for services either directly to the server hosting the service or to a Directory Agent. The servers must accept either multicast service requests or unicast service requests.
- SLP has an extension which is known as Mesh Enhanced SLP (mSLP) which can be adapted to work in WMNs particularly Infrastructure and hybrid WMNs.

3.1 Service Location Protocol (SLP)

This is a service discovery protocol developed by IETF. The Service Location Protocol provides a scalable framework for the discovery and selection of network services [10]. In these protocols agents are assigned to both the clients/users and servers, these agents act on their behalf. Services are searched for using string based querying, alternatively services could be browsed. Typically user agents send out request for services via multicast and the service agents which have the service send replies via unicast. SLP can work in either directory-less or directory centric mode as shown in figure 2, UA can either send requests directly to the SAs or go through the DAs. Directory agents were introduced to improve efficiency and scalability. These directory agents store a list of available service agents and the services they provide, instead of user agents sending multicasts to the service agents, they send unicast messages to the directory agents to acquire information about the available services. Services are accessed via URLs (Universal Resource locator). Also scope awareness is available and can be used to classify services into scopes to simplify service discovery.

User Agent (UA): A process working on the user's behalf to establish contact with some service. The UA retrieves service information from the Service Agents or Directory Agents [10].

Service Agent (SA): A process working on the behalf of one or more services to advertise the services [10].

Directory Agent (DA): A process which collects service advertisements. There can only be one DA present per given host [10].
3.2 Mesh Enhanced Service Location Protocol (mSLP)

Several extensions have been made to SLP, one of which is the Mesh-enhanced Service location Protocol (mSLP). mSLP proposes interaction between the DAs to form an overlay network [8]. This extension was proposed mainly to make the SA lighter. In SLP the service agent has to register with all DAs on the network and reregister whenever a DA is rebooted and also with all new DAs which join the network. With the extension proposed in mSLP the Service registrations have to be sent to other peer DAs in the overlay network formed [11]. This extension is most applicable in hierarchical and clustered large scale networks.

4. SIMULATION

In order to measure the suitability of SLP in WMNs, simulations were carried out in NS2 network simulator. There are two implementations of SLP in NS2, one of which is the Basic SLP implementation as described in the RFC [10], the second is a Cache Enhanced SLP implementation where the user agents which participate in relaying a service reply from a service agent to another user agent cache this service reply as it is relayed. Simulations were carried out using both implementations varying the number of client service discovery requests. WMNs have several areas of application but there was the need to choose scenarios where users/nodes in the network need several network services and need to discover and use these services. Two main areas of application where chosen.

4.1 Enterprise Scenario

Enterprise here refers to a large network among multiple buildings or a medium size network for offices in different floors in a building or a small network within an office. The type of mesh network which will be appropriate here is either Hybrid/Hierarchical mesh network. Figure 3 shows a medium size enterprise WMN, this network is made of 4 mesh routers, one of which is the mesh portal and is connected to a wired network infrastructure and all clients nodes connected to the other mesh routers send packets through it to the wired network infrastructure. A mesh connection is created amongst these mesh routers and the mesh portal.
act as both a base station and a wireless relay node and is responsible for forwarding/routing traffic meant for
the wired infrastructure network it is connected to.

Figure 3. An illustration of the simulated Enterprise Scenario.

4.2 Conference/Meeting Scenario

This basically refers to a scenario where several users come together for a particular event and will need to
use services like internet, printing and faxing services during the period when the event will be taking place
which can vary from a few hours to few days or more but not a permanent situation like an office or any
organization where there is a fixed infrastructure. The type of mesh network which will be most appropriate
in such situations is the Flat/Client wireless mesh network. Figure 4 is a representation of the scenario which
was created in NS2 network simulator, it is made up several clients nodes, where 4 of this client nodes act as
servers with each one attached to a particular service and other clients in the network access these services
via these servers.

Figure 4. A simple illustration of Client Mesh Network [3].

4.3 Simulation Results and Analysis

In order to do a proper comparison of the performance of SLP in the two WMN scenarios simulated, a
smaller enterprise network with the same number of nodes as the conference scenario that is 45 nodes clients
was simulated and the results of simulation were used to carry out this comparison. The comparisons are
carried out for both the Basic SLP implementation and Cache Enhanced SLP implementation based on the
network bandwidth usage, successful discovery rate and latency/delay experienced in the service discovery
transactions.

4.3.1 Network Bandwidth Usage

Figures 5 and 6 below provide a comparison of the enterprise and conference scenarios based on the
aggregate network bandwidth used, with a total of 45 client nodes varying the number of requests. In both
figures we can observe that the bandwidth used for the same number of requests is higher in the Enterprise
scenario. The performance of both scenarios is higher in the Cache Enhanced SLP if both figures are
compared.
4.3.2 Successful Discovery Rate

The rate of successful discovery is calculated in percentage. Figures 7 and 8 provide a comparison of the performance of the Enterprise and Conference scenario based on the percentage of service discovery transactions which were successful. In both figures we can observe that the percentage of successful discovery transactions in the conference scenario is 100% whereas that of the Enterprise scenario varies without a particular pattern. The variation in the rate of successful discovery transactions in the Enterprise scenario can be attributed to the variation in the distance between client nodes and servers.

4.3.3 Latency/Delay

The delay obtained from the simulations with both the basic SLP and the cache enhanced SLP in both the Enterprise and Conference scenarios do not have a specific pattern with respect to the number of requests.
whether decreasing or increasing. But it was observed that the ratio of delay in the enterprise scenario with both basic SLP implementation and cache enhanced SLP is higher than that of the Conference scenario. The major contributing factor to the delay is the distance between the client and the server and also the number of times the service request is retransmitted before a reply is received. In the case of the Enterprise scenario, the delay can be further attributed to the fact that both service discovery requests and replies need to be relayed not only via the mesh router but also the mesh portal.

5. CONCLUSION

In this paper, WMN were introduced with some of the open research issues, one of which is an appropriate service discovery protocol. SLP was proposed as a suitable service discovery protocol based on certain features it has which were identified and simulations were carried out in NS2 network simulator to measure its suitability for WMNs. The results obtained showed that it is suitable to a reasonable extend but more work/research need to be carried out in order to prove its 100% suitability.

REFERENCES


ENHANCED SIMPLE FAIR EXECUTION TIME ESTIMATION (ESFETE) SCHEDULING ALGORITHM FOR GRID SCHEDULER

Rohaya Latip and Adam Malik Sahar
Department Of Technology Communication and Network,
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia,
43400, Serdang Selangor, Malaysia.

ABSTRACT
Grid computing involves sharing and coordinating distributed resources such as computing power, application, data storage or network resources across dynamic and geographically dispersed multiple administrative virtual organizations. Grids fall into the category of distributed parallel computing systems that have a lot of unique characteristics which make the scheduling in Grid highly difficult. Submitted jobs in Grid environment are normally put in queue due to the large number of jobs submission. An adequate Grid scheduling technique used to schedule these jobs and sending them to their assigned resources. The main goal of this paper is to analyze, classify the problems and study the limitations of Simple Fair Estimation Time Execution (SFETE) Grid scheduling algorithm including the studying of improvement on the SFETE. SFETE was selected because the algorithms outperform the existing gLite production Grid middleware scheduling. Problem and limitation extracted from SFETE and enhanced approach of SFETE, called Enhanced Simple Fair Estimation Time Execution (eSFETE) was implemented. The implementation and comparative study were done in simulation Grid written in standard C programming language. The outcome shows that eSFETE produce better performance result in term of total delay due to the unfair utilization of idle resources by SFETE.

KEYWORDS
Grid Computing, Scheduling, Simulation, Delay.

1. INTRODUCTION
Grid is an emerging as a wide-scale, distributed computing infrastructure that promises to support resource sharing and coordinated problem solving in dynamic, multi institutional Virtual Organization [5]. Grids consist of geographically distributed and heterogeneous communication, computation and storage resources that may belong to different administrative domains, but can be shared among users [2].

The most remarkable benefit of Grid computing is collaboration of many computers (resources) in a network for a single problem at the same time, which requires a great number of computing speed or access to large volume of data usually to solve massive and complicated computational problems. Grid Computing can be thought of as distributed and large-scale cluster computing and as a form of network-distributed parallel processing. The aim of the Grid is to provide the ability to manage all the computing resources across different administrative domains which cannot be completed by traditional clusters or distributed computing.

Grid efficiency has become an essential problem due to the Grid structure itself [4]. A Grid is a structure of heterogeneous distributed resources that are offered to it users. Because of it complexity, the performance evaluation approach for the Grid is much more difficult than traditional distributed computing environment [8].

The main purpose of scheduling is to satisfy users and system demands for example to reduce total delay of the jobs. The good scheduler is when the scheduler can pass the job to the resources at lower schedule time and lower delay.

Numbers of scheduling algorithms are used to minimize the number of delayed jobs in Grid computing. This paper concentrates on the performance study of eSFETE (Enhanced Simple Fair Estimation Time
Execution Grid scheduling algorithm in term of total delay. The eSFETE was derived from SFETE (Simple Fair Estimation Time Execution) in [2] Grid scheduling algorithm to overcome some limitation. Computer simulation was used to demonstrate the interesting result.

SFETE algorithm solely relies on the computational capacity of resources and the number of jobs in the resources queue. The computational capacity gives high influence on the computation of “fair execution time” variable.

SFETE totally ignore the resources status whether it idle or busy. From the observation, at a certain queue size, highest speed resource always selected even they are not idle. This will result in significant delay at certain workload when the high speed resource is busy with long queuing jobs.

The paper is organized as follows. In Section 2, we discuss the related work which is SFETE. Section 3, we carry on the discussion with the enhancement of SFETE, called Enhanced Simple Fair Estimation Time Execution (eSFETE). Section 4 simulate the simulation design and Section 5 discussed on the simulation result. The conclusion and future works in the research is discussed in Section 4.

2. RELATED WORK

The SFETE assigns job $i$ to resource $j$ that provides the minimum simple fair execution time $E_{ij}$. The simple fair execution time $E_{ij}$ is an estimation of the time by which job $i$ will be executed on resource $j$, assuming it gets a fair share of the resource’s computational power, without taking into account the fair execution times of the other tasks already assigned to the resource.

SFETE pseudo code:

1: for each job $i$ queued in the scheduler’s ordered list do
2:      for each resource $j$ in the Grid do
3:         Estimate the fair execution time: $E_{ij} = \frac{(N_j + 1)}{C_j}$
4:      end for
5:  Assign job $i$ to resource $j$ that gives the minimum simple fair execution time $E_{ij}$
6:  $N_j \leftarrow N_j + 1$
7:  Send the scheduling decision to the user of job $i$
8: end for

where $N_j$: the number of jobs in resource’s queue; $C_j$: computational capacity of resource

SFETE has some limitations. The scheduling decision made by SFETE is based on the computational capacity and the number of jobs in resource’s queue. SFETE totally ignore the resources status whether the resources are free or busy. The ignorance will result in significant total delay at certain workload when all the resources are busy with long queuing jobs.

3. THE ENHANCED SIMPLE FAIR EXECUTION TASK ESTIMATION (ESFETE)

Based on the SFETE problem statement stated in previous section, the enhanced version of SFETE, called Enhanced Simple Fair Execution Task Estimation (eSFETE) was proposed to overcome the issue. In eSFETE, resources status was determined first before the computation of simple fair execution time, $E_{ij}$. Job assigned to the idle grid resources regardless their computational capacity. If more than one idle resources are available, the first found idle resource will be selected. The assign policy effected only if all Grid resources are busy. Then the job will be executed on the resource with the minimum $E_{ij}$, similar to the original SFETE.

The pseudo code of eSFETE as below:

1: Find idle resource
2: If found, select the idle resource
3: Send the scheduling decision to the user of job $i$
4: If all resources are busy
5: for each job $i$ queued in the scheduler’s ordered list do
6: for each resource $j$ in the Grid do
7: Estimate the fair execution time: $\hat{E}_{ij} = \frac{(N_j + 1)}{C_j}$
8: end for
9: Assign job $i$ to resource $j$ that gives the minimum simple fair execution time $\hat{E}_{ij}$
10: $N_j = N_j + 1$
11: Send the scheduling decision to the user of job $i$
12: end for

where $N_j$: the number of jobs in resource’s queue ; $C_j$: computational capacity of resource

4. SIMULATION AND DESIGN

The simulation model was coded in C programming language. All computing element are connected together and can access the waiting queue. The time $t$, spent on each computing element can be measured separately and the service rate $\mu$ of this unit is calculated by $\mu = 1/t$. Then every computing element finishes a job in a period of time, which follows an exponential distribution with rate $\mu$.

Jobs are submitted to job scheduler in a Poisson stream with rate $\lambda$. In this assumption consecutive jobs arrive after intervals which independently are exponentially distributed. If there are several computing element available, one of them will be selected by the scheduler and execute the job independently. All available computing elements have the same probability. Once a job is assigned, no other jobs will be accepted at the same time. If all jobs keep waiting in the waiting queue which is infinite long until someone finishes its job and becomes available again.

Multi-server queue with different service rates is suitable for this paper with regard to queuing theory [1,3]. Some preconditions including Poisson arrival process [3, 6, 8], exponential distributed service [6, 8], $n$ servers in the system and infinite queue size [8]. Service time for each server follows an exponential distribution with its own rate $\mu$.

In the simulation process flow, the first step, all variables at the start of program must be initialized. Second step, determine the event type of the next event to occur by identify the smallest start time among the events. The event with smallest start time will be executed first. As the third step, the program performs either the arrival or departure routine and updates the statistics accordingly.

For an arrival event, the program checks for the scheduling policy decision to determine the selected server for the job assignment. If the server is free, it is set to busy and scheduled for a departure event. If all servers are busy, the local queue is incremented by one. If the local queue is full, jobs are inserted in the global queue.

For departure event, only the server engaged in this event becomes free. If there is no job in queue, this server remains idle. When the queue is not empty, the server is busy and the next departure event will be scheduled. The local queue is decremented by one. If there are any jobs exist in global queue, the job will forwarded to the selected server accordingly. The job will be processed if the server is idle and put in the local queue if the server is busy.

Termination of program is based on the number of jobs departed and the number of experiments set. If the values of both variables are reaching to the initialized values, the program will be terminated and the result is computed.

All algorithms namely as eSFETE and SFETE use First Come First Server (FCFS) ordering policy and evaluated in non uniform resources or server in which all resources have their own service time. The level of abstraction of Grid computing can be simplified by the queuing theory model. The simulation model has an infinite global queue and multiple non uniform servers with its own local queue and service rate, $\mu$. Arrival process follows Poisson stream with rate $\lambda$ and the service rate for each servers follows an exponential distribution with its own rate, $\mu$.
When the user submits a request to the scheduler, the scheduler hands it on to the lower level system to deal with. The scheduler selects the suitable server or resources to process the job according to the assignment policy implemented on it. If any resources are busy, the job will be kept in the waiting queue until the resource is set to free. The global queue only can be utilized if all local queue is full.

Submitted jobs in Grid environment are normally put in queue due to the large number of jobs submission. To overcome this issue, scheduling technique used in the grid to schedule the jobs but there is delay in process of scheduling these jobs and sending them to their assigned resources.

The buffer capacity for global queue is infinite whereas the capacity for all local queue is set to 100. Five resources represent grid virtual organization computation element implemented to deal with the job processing. Exponential service rates for each resources are pseudo-random between 0.0 and 1.0 (0 < services rate < 1).

The termination of simulation is following the number of jobs processed or departed and set to be 11,000.

The simulation will not address the security and accounting issues and also the communication delay. The simulation is mainly focus on the execution time of the tasks which is the case in Computational Grids and measured by “total delay” performance metric.

Performance matrices evaluated in this paper is total delay which it depends on six other variables listed below. There are six associated time variables for each job [7].

i) The arrival time of job \( a_i \)

ii) The delay of job \( i \) in the queue is \( d_i \geq 0 \).

iii) The time that job \( i \) begins service is \( b_i = a_i + d_i \).

iv) The service time of job \( i \) is \( s_i > 0 \).

v) The wait of job \( i \) in the service node (queue and service) is \( w_i = d_i + s_i \).

vi) The time that job \( i \) completes service (the departure time) is \( c_i = a_i + w_i \).

Based on the definition by [7], scheduling algorithms implemented in this paper was evaluated by the performance metrics as below:

i) Job delay: Delay of the job in the queue, \( d_i \) where \( d_i = b_i - a_i \)

ii) Total delay, \( D \): Total delay of the jobs, \( D = \sum (b_i - a_i) \)

5. RESULT AND DISCUSSIONS

Fig. 1 demonstrates the performance of SFETE and eSFETE algorithms measured by the total delay. The simulation started with job submission rate of 50 jobs/sec. and ended with 500 jobs/sec. which is incremental of 50 jobs for each experiment. Number of total experiments done is 10 experiments.

![Simulation result of eSFETE and SFETE](image_url)
The eSFETE achieve smaller total delay compare to SFETE. At job submission rate 200 jobs/sec., total delay decrease from 130 seconds to 19 seconds or 85% decrement compared to original algorithm performance. 60% decrement at job submission rate 350 jobs/sec and total delay decrease from 1270 seconds to 563 seconds or 56% decrement at job submission rate 450 jobs/sec.

The average total delay of SFETE is 553.38 seconds whereas eSFETE is 223.46 seconds. It is 59.62% decrement for eSFETE average total delay compared to SFETE performance. It shows that eSFETE produce less total delay than the original algorithm; SFETE. This proof that percentage of decrement become smaller as number of jobs submission rate higher. It is because, the more jobs submitted to the resources, the lesser idle resources are to be utilized by eSFETE.

The eSFETE treats the jobs and utilize resources in a fairer manner. It assigned jobs to the idle resource without taking into consideration of the minimum “fair execution time” variable. The eSFETE executed the assign policy according to minimum “fair execution time” variable only if there are no idle resource remains.

6. CONCLUSION

In this paper, the scheduling algorithm (SFETE & eSFETE) implemented in this paper assigns jobs to the computation resources based on the minimum “fair execution time” variable. Based on the result finding, it can be concluded that eSFETE outperform the original SFETE scheduling algorithm in term of total delay due to very high influences from computational capacity variable on the SFETE formula.

For future work, eSFETE use to look up for available computing resources. The searching technique use to select the best resource among idle resources that waiting for the jobs to be assigned. Proper searching technique search the entire solution space for an optimal solution and decrease the time taken to find an acceptable solution.

REFERENCES

A PROPOSAL OF ACCESS-BASED CONTENTS GROUPING ON P2P NETWORK

Takuya Sasaki¹, Jun Sawamoto¹ and Yuji Wada²

¹Faculty of Software and Information Science, Iwate Prefectural University, Takizawa, Iwate 020-0193, Japan
²Department of Information Environment, Tokyo Denki University, Inzai, Chiba 270-1382, Japan

ABSTRACT

As for the usage of the information, the attention of information sharing using P2P network rises, and it is becoming to be used in many fields. P2P networks directly connect computers in an equal relationship and the load is distributed by exchanging information and services among computers. The DHT in P2P network has the drawbacks that the partial matching or the group matching retrieval cannot be achieved. It is difficult to do a flexible retrieval. This becomes a problem that should be solved when the search engine is constructed. In this paper, to improve the convenience of the content access, we propose an efficient content access method based on the grouping technique of the contents using the content access history and frequency. Then the evaluation of grouping efficiency (precision, recall) is performed by the simple simulation.

KEYWORDS

P2P network, distributed hash table, overlay network, content retrieval, grouping technique.

1. INTRODUCTION

As for the usage of the information, the attention of sharing of information using P2P network rises, and it is becoming to be used in many fields (“Skype”) (Zhang, R. et al, 2007) (Sripanidkulchai, K. et al, 2003). P2P networks directly connect computers in an equal relationship and the load is distributed by exchanging information and services among computers, and high fault tolerance is also one of the advantages. The most common type of structured P2P network is the DHT (Stoica, I. et al, 2001) (Maymounkov, P. et al, 2002), in which a variant of consistent hashing is used to assign ownership of each content to a particular peer. Therefore, DHT can search content with the number of very few messages; however, the perfect matching with the search keyword is demanded in the hash table. The DHT has the drawbacks that the partial matching or the group matching retrieval cannot be achieved. It is difficult to do a flexible retrieval. This becomes a problem that should be solved when the search engine is constructed.

In this paper, to improve the convenience and flexibility of the content retrieval, we propose an efficient content access method based on the grouping technique of the contents using the content access history and frequency in ubiquitous P2P network environment. We introduce grouping structure into DHT in P2P and enhance conventional keyword search capability (Necip et al, 2002) (Jin Y. et al, 2008). In DHT, it is a burden for the network manager to renew the hash table when a new group appears or an old group is deleted. Then, the automatic generation and reformation of group structure is examined. The group is updated in proportion to the retrieval frequency. Less frequently retrieved groups are deleted from the table. Thus, the user can find favorable data more easily and efficiently. Finally, we evaluate the proposed method by simulation to evaluate whether the automatic grouping on the DHT is carried out as we assumed in P2P network environment.

2. RELEVANCE INDEX FOR CONTENTS GROUPING

A P2P node works as a content holder and at the same time a user node who executes accesses to the contents on other nodes. When a user accesses plural contents in succession within a certain time period, e.g., a day,
we assume that those contents become to have high relevance indices with each other. Relevance index shows how much value the continuous access of two contents has for users. Relevance index is shown by equation (1). Here, $K_{s,i}$ shows the relevance index between contents $s$ and $i$. $V_n^{s,i}$ has value 1 if contents $s$ and $i$ are continuously accessed by a user at P2P node $n$, otherwise has value 0. In other words, Relevance index increases by value 1 when continuous access takes place by a user at some P2P node.

On the other hand, because time passes, in a contents group which has been formed, and was not used, relevance index decreases, and the grouping is broken off. The rate of reduction of the relevance index reflects the speed by that less frequently accessed group contents lose their relevance as a group.

$$K_{s,i} = \sum_{n \text{nodes in P2P}} V_n^{s,i}$$ (1)

3. SYSTEM CONFIGURATION

P2P network is constructed with the DHT that uses Kademlia algorithm (Maymounkov, P. et al, 2002). A P2P node usually participates to the network as a content holder (provider) and/or a user (consumer) of contents. In this system, we assume that one node holds a single content for simplicity.

3.1 P2P Node

In general, a node of P2P network holds hash table and the contents information. In this paper, a list of access history and a group list are hold at each node. Following four data structures are located in each node.

1) Contents information: Content files which this node holds.
2) Distributed hash table: Index information of the node, i.e., hash value and node ID.
3) A list of access history: We maintain a list of access history for a set of continuous access actions at a node. When there is a node which holds the object content that a user is accessing, the content ID is added to the list of access history, and it is erased when the acquisition of contents are completed.
4) A group list: We maintain a group list which consists of content ID of high relevance index with this hosting node, and it is sorted in descending order of the relevance index value. As time passes, the index values in the group list are maintained and decreases. When this hosting node is accessed from a user node, contents with high relevance indices are recommended back to the user node.

3.2 Contents Grouping by Content Access History and Frequency

The communication process between nodes for the grouping of the contents (or here, nodes which contain contents) is detailed. Below, we describe the communication process step by step.

1) A user at node A requests to access content at node B after some retrieval action.
2) Node A sends its own list of access history to node B. At node B, after receiving the list of access history, contents in the received list are merged to the group list of node B. If contents with the same IDs exist already in the B’s group list, only relevance indices are increased by one.
3) Node B returns the requested own content to node A.
4) Node B sends back information of contents, which are members of B’s group list and have relevance indices higher than a certain threshold set by the system parameter, to node A.
5) At node A, B’s content IDs are added to the access history. By using information of contents from node B, A might access contents in the B’s group list. This function is called the recommendation of contents in terms of the value of relevance index by node B to node A.

3.3 System Parameters for Contents Grouping

Table 1 shows system parameters that are main factors for determining characteristics of contents grouping. Number of participating nodes and number of nodes which hold contents determine the size of P2P network.
Access tendency index and access count are considered to affect how quickly contents groups are formed in the process. The rate of reduction of the relevance index reflects the fact that less frequently accessed group contents lose their relevance in a group.

Table 1. System parameters for contents grouping

<table>
<thead>
<tr>
<th>Name of parameter</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>Maximum number of participating nodes in P2P network.</td>
</tr>
<tr>
<td>Number of content nodes</td>
<td>Number of nodes which hold contents.</td>
</tr>
<tr>
<td>Max size of the list of the access history</td>
<td>Maximum number of content items the list of access history holds. The higher the value, the more it is possible to maintain the user's access history. It is set to hold the max number of accesses in one access cycle of an average user.</td>
</tr>
<tr>
<td>Max size of the group list</td>
<td>Maximum number of content items the group list hold. This parameter determines the size of the group that can be formed in the system. Number of highly relevant contents increases by increasing this size.</td>
</tr>
<tr>
<td>Reduction rate of the relevance index</td>
<td>The rate the relevance index is reduced in one day period. If the value is x% then current relevance index S becomes $S \times x/100$.</td>
</tr>
<tr>
<td>Threshold of relevance index</td>
<td>The threshold to tell whether contents are considered in the same group or not by the relevance among contents.</td>
</tr>
<tr>
<td>Access tendency index</td>
<td>Access tendency index indicates how frequently the user accesses contents which are candidates for grouping in a short period, e.g. a day. The tendency is higher the group could be formed more quickly.</td>
</tr>
<tr>
<td>Average access frequency</td>
<td>The average number of accesses a user executes a day on average.</td>
</tr>
<tr>
<td>Recommendation function (on/off)</td>
<td>This function is a recommendation function of contents in terms of the value of relevance index. If recommendation function is on, it could be considered to accelerate formation of contents grouping.</td>
</tr>
</tbody>
</table>

4. PERFORMANCE EVALUATION

We use the Distributed Environment Emulator of "Overlay Weaver" (“Overlay Weaver”), which is an overlay construction toolkit, and perform the experiment. We prepare the scenario file for the emulator which appoints the movement of the node such as retrieval /acquisition of contents. For the content node we pre-set temporary groups beforehand in this scenario. The access of contents by user node is carried out from the pre-determined candidate groups based on the access tendency index. At the end of the day, the relevance index is reduced by the reduction rate given by the parameter. In this experiment, we iterate this activity for thirty days. Table 2 shows parameter values used for the experiment.

Table 2. Parameters set for the experiment

<table>
<thead>
<tr>
<th>Name of parameter</th>
<th>Parameter values</th>
<th>Name of parameter</th>
<th>Parameter values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>10000 nodes</td>
<td>Threshold of relevance index</td>
<td>3</td>
</tr>
<tr>
<td>Number of content nodes</td>
<td>1000 nodes</td>
<td>Access tendency index</td>
<td>20%,50%,80%</td>
</tr>
<tr>
<td>Max size of the list of access history</td>
<td>5</td>
<td>Average access frequency</td>
<td>3</td>
</tr>
<tr>
<td>Max size of the group list</td>
<td>10</td>
<td>Recommendation function (on/off)</td>
<td>on, off</td>
</tr>
<tr>
<td>Reduction rate of the relevance index</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1 Performance Measures

The measures we use to express the performance of contents grouping are:

1) Precision: It is the fraction of the contents in the correctly (as intended) formed group relevant to contents in formed group. Precision is shown in equation (2), where, $P$: precision, $k$: number of contents in the correct groups, $n$: number of intended groups of contents, $m$: number of formed group, $l$: maximum size of group list $= 10$.

$$P = \frac{\sum_{i=1}^{n} k_i}{m}$$ (2)
2) Recall: It is the fraction of the contents in the correctly formed group relevant to all contents. Recall is shown in equation (3), where, \( R \): recall, \( k \): number of contents in the correct groups, \( n \): number of intended groups of contents, \( l \): maximum size of group list = 10.

\[
R = \frac{\sum_{i=1}^{k} \frac{k}{l}}{n}
\]  

(3)

4.2 Experimental Results and Consideration

In this section, we explain some of the experimental results and discuss the feature of the proposed contents grouping method. The experiment investigates the effect of the recommendation function. We evaluate it using parameter values of Case A through F as show in Table 3.

Table 3. Parameters set for the experiment

<table>
<thead>
<tr>
<th>Case</th>
<th>No. of nodes</th>
<th>Reduc. rate of the relev. index</th>
<th>Threshold of relev. index</th>
<th>Access tendency index</th>
<th>Ave. access freq.</th>
<th>Recomm. function</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>80%</td>
<td>3</td>
<td>off</td>
</tr>
<tr>
<td>B</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>80%</td>
<td>3</td>
<td>on</td>
</tr>
<tr>
<td>C</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>50%</td>
<td>3</td>
<td>off</td>
</tr>
<tr>
<td>D</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>50%</td>
<td>3</td>
<td>on</td>
</tr>
<tr>
<td>E</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>20%</td>
<td>3</td>
<td>off</td>
</tr>
<tr>
<td>F</td>
<td>10000</td>
<td>5%</td>
<td>3</td>
<td>20%</td>
<td>3</td>
<td>on</td>
</tr>
</tbody>
</table>

In this experiment, it is shown that the recommendation function works for certain range of access tendency index. We can confirm that the recommendation function accelerates the formation of groups in the daily retrieval and access activities of users on P2P network environment.

1) In case the access tendency index is 80% (Cases A, B)
   The result shows that under this circumstance, both precision and recall are very high (almost 100%) after day 5 irrelevant of the recommendation function. In this case, the grouping of contents is achieved very quickly without introducing recommendation function.

2) In case the access tendency index is 50% (Cases C, D)
   Figure 1 shows the outcome of experiments where the access tendency index is 50%. In Figure 1, keywords on/off indicate whether the recommendation function is used or not. In this experiment, both precision and recall are decreased drastically when the recommendation function is off. However, with the recommendation function on, both precision and recall reach 90% after day 15. This result shows that even if users access irrelevant contents 50% in one access cycle, it is possible to choose necessary contents and form a group with the help of recommendation function.

3) In case the access tendency index is 20% (Cases E, F)
   Figure 2 shows the outcome of experiments where the access tendency index is 20%. In this case, users access almost irrelevant contents and even with the recommendation function on, meaningful grouping is not performed and both precision and recall are very low.

In this experiment, we fixed the number of P2P nodes and verify the basic feature of the proposed access-based contents grouping function. And we confirmed the effect of the recommendation function under certain noisy access behavior of the user. Because this function forms a contents group dynamically depending on the content access request issued from users, it can be said that the flexible feature has been implemented that corresponds to dynamic changes of nodes, i.e., ad hoc join and secession, in P2P network. By providing a content group that an interested content belongs when a certain search request is made as we have shown in this paper, we assume that the efficiency of the contents retrieval action is improved to some extent.
5. CONCLUSION

In this paper, we proposed an access-based contents grouping method aiming at an information retrieval system targeting contents in the P2P network, and performed evaluation by simulation. We intend to implement and evaluate the full function of grouping of contents in DHT. Based on the evaluation, improving the function of grouping of contents, it is scheduled to go to the system design of the efficient ubiquitous data retrieval system. In addition, the verification of the usability is scheduled to be covered in consideration of a concrete application system, and experimenting with the realistic applications.

REFERENCES


Reflection Paper
ABSTRACT
Knowledge sharing is of vital importance to organizations, enabling them to develop skills and competences, increase value, and sustain their competitive advantage. Knowledge is a firm's most valuable resource because it embodies intangible assets, routines, and creative processes that are difficult to imitate. Thus in this study we have investigated the role of information technology as facilitator knowledge sharing in the organization. This study mainly probes IT as a tool which is able to manage, store, and transmit structural knowledge. It can support us in our efforts to make the knowledge stored in the human brain or in documents available to all employees of an organization. Also we present new technology for convenient knowledge sharing as a successful case study in the roads management center.

KEYWORDS
Knowledge sharing, knowledge management, information technology

1. INTRODUCTION

Due to the IT revolution and advancements of the Internet, the value of knowledge assets has been greatly enhanced. Many companies are building knowledge management system (KMS) in order to manage organizational learning and business know-how. The main purpose of such a policy is to help knowledge workers to create important business knowledge, to organize it, and to make it available whenever and wherever it is needed in the companies (O'Brien & Marakas, 2006).

The highest value of IT to KM is in allowing the expansion and universalization of the scope of knowledge and in increasing the speed of transferability. Additionally using IT, we are able to retrieve and store knowledge in individual or groups, which allows this knowledge to be shared with other divisions in the same organization or business partners in the world. Furthermore, IT contributes to the integration of knowledge or even to the stimulation of new knowledge (Davenport & Prusak, 1998).

However, many companies have faced various kinds of difficulties in implementing KMS. First, if knowledge is merely accumulated in workers' brains, there is no way of recording it systematically. Second, even though knowledge is recorded and recorded in documents, it is very complicated to search for, retrieve, or review it, a problem which erects barriers to the diffusion of knowledge. Thus, in past times, even though managers knew how important KM was, it was very difficult to implement it successfully (Bradley, Paul, & Seeman, 2006). Thus, the growth of KM has been closely tied to information and communication technology (Chumer, Hull, & Prichard, 2000). Therefore, it is found that IT plays a major role in the implementation of KMS (Hislop, 2002). Nevertheless, few studies explore the role and effect of information technologies in the KMS. Hence, the purpose of this study is to investigate the role and effect of IT in sharing knowledge in the KMS as a factor of success in knowledge management project, and introducing new and effective method for it. To deal with this issue more effectively, we focus on a key question:

How can information technology facilitate knowledge sharing in organization?
The research indicates an important issue of KM that IT is an indispensable enabler of KM. while IT-enabled knowledge management goes beyond mere automation to play an informing role in organizations by facilitating knowledge sharing.

2. LITERATURE REVIEW

2.1 Knowledge Sharing

Knowledge consists of truths and beliefs, perspectives and concepts, judgments and expectations, methodologies and know-how (Wiig, 1993). Organizational knowledge is the collective sum of human-centered assets, intellectual property assets, infrastructure assets, and market assets (Brooking, 1996). Sveiby (1998) defines knowledge management is the art of creating value from an organization's intangible assets. Moreover, he identifies two main tracks of knowledge management activities: one track focuses on knowledge management as the management of information and the other track as the management of people. Other researches show; linking the individual perspective of knowledge to the organizational level, organizational knowledge creation theory is concerned with the processes which make available individual knowledge to the organizational knowledge system (Nonaka and von Krogh, 2009).

In additional, according to researches; Knowledge sharing requires collaboration between the users of knowledge; namely the collaborators. This task cannot be accomplished simply by storing knowledge in the repository. It also requires a mechanism, which helps people find the collaborators with relevant knowledge. Collaboration over the Internet communities has characterized itself by heavily relying on interaction among the collaborators (Biström, 2005; Eikemeier and Lechner, 2003). Collaborators can be any virtual users who interact to achieve the goals of resources discovery, access, knowledge sharing, group communication and discussion. The collaboration for knowledge sharing should be enacted without spatial and temporal limitations. In addition, it should take place over medium such as the Internet and therefore beyond the geographical boundaries.

2.2 The Role of IT in Knowledge Sharing

The means by which knowledge is shared within organizations and the factors that facilitate knowledge sharing/transfer are core issues in knowledge management. advances in technology have facilitated the recent growth in systems designed for managing organizational knowledge, IT is comprehensively utilized by members in organization, IT is comprehensively constructed in organization, top management is capable of applying IT, members in organization apply IT to search and use current organizational knowledge, and members in an organization apply IT to create new knowledge (Peter J. Sher & Vivid C. Lee, 2003).

In the process of KM, the absorption, creation, arrangement, storage, transfer and diffusion of knowledge are all dependent on assistance provided by IT. Khandelwal and Gottschalk (2003) pointed out that the application of IT to the support of KM apparently influences the results of knowledge collaboration within the organization. There are some example of using information technology for implementing KMS and sharing knowledge in organization:

Hewlett-Packard (HP), a company competing in the market of computers, peripheral equipment and other electronic equipment developed CONNEX a People-Finder KMS (T Carrozza, phone interview and follow-up e-mail with developer of CONNEX at HP Labs, September 16, 1999). The goal of the project was to build a network of experts, available online, to provide a guide to human knowledge within HP. CONNEX consists of a centralized database of user knowledge profiles, with a Web browser interface that allows users to find profiles in multiple ways. User's profiles contain a summary of their knowledge and skills, affiliations, education and interests, as well as contact information. CONNEX users can easily find experts within HP by searching the database by any combination of profile fields or by browsing through the different areas of knowledge, geographies and/or names. To support a large user base with high volume of transactions, CONNEX was built using Sybase database and Verity's Topic search engine, on an HP platform. The National Security Agency (NSA) has also taken a step towards the implementation of a system to locate experts for using their knowledge in critical situations (A. Wright, W. Spencer, 1999).
3. CASE STUDY

The roads management center, in Ministry of Road & Transportation of Iran, is a medium organization with different branches across Iran. This center recently has implemented new knowledge management system by focusing on creating, gathering, and disseminating an organization's knowledge as opposed to 'information' or 'data'. The development of KMS, in this center, demands that knowledge be obtained, produced, shared, regulated and leveraged by a steady conglomeration of individuals, processes, information technology applications and a knowledge-sharing organizational culture.

For getting this target, they use new technology to make appropriate condition for fascinating knowledge sharing in organization because the Ministry of Road & Transportation had been used the technology for knowledge sharing in knowledge management project already and it was not proper for this goal. New technology is Wiki technology, as one of the advantages of web2, has many good points in sharing knowledge such as in customer/client collaboration, documentation, and developing an online community. The information is often added to wiki but not deleted when no longer relevant or accurate or updated when changed. Wiki offers an excellent way to manage documents and knowledge integrity. In wiki, foremost is the fact that documents are edited in a very visible way, which adds accountability and Members of organization have to justify the changes because everybody can see it. Also, each of members can edit or add new information to other knowledge or information that has been written by other members, previous technology did not have this feature and it was the main weakness of that. This faint was the reason of employee's discontent. A technical advantage of wikies over other document management tools is that there are plenty of good open source versions available at little or no cost. Plus, wiki is usually extensible, so organization can customize them to its needs and doesn't need an expert administrator or extra hardware resources. Despite wiki's benefits, the success of wiki in KMS depends on how dedicated the participants are in using the wiki and checking in regularly and wiki platforms have a bit of a learning curve. With training members in organization the usage of wiki reveals obviously. The advantages of wiki technology help roads management center for making proper condition between clients to share their knowledge more than before.

4. METHODOLOGY

For this research the qualitative research method make it better suited to be applied here. Therefore, there is a design phase involved, which possesses distinct methodology. The phase involved voluminous review of the literature and in-depth interviews with senior managers in roads management center, both of which were aimed at collecting data. Interviews are one of the most extensively used methods of data collection (Bryman & Burgess, 1999). The individual in-depth interviews conducted in this study are of a face-to face, which is one of the most common approaches in qualitative research. This type of interview involves asking a number of pre-determined questions and special topics. Under such circumstances, respondents are able to determine the direction and content of the interview within a broader framework provided by the interviewer. After the interview at each manager had been completed, the results were assembled, transcribed and e-mailed to the respondents for their review and approval in order to prevent any misinterpretations. This process is expected to provide this study with a richer and more holistic appreciation of the problems regarding. In our study, we have selected twenty top managers and about thirty employees of roads management center, in order to answer our research question we have asked the various questions.

5. RESULTS AND DISCUSSION

After analyses of all the replies we have found, most of the managers and employees believe that sharing knowledge is necessary for effectiveness and having efficiency in their organization and it can help reduce many cost the work again. They emphasized that new approach in implementing knowledge management with using wiki technology has help better and easier knowledge sharing in organization than previous approach and technology because new approach use story telling technic for sharing knowledge and wiki technology supports this method in best way. The result of this process will be added in wiki after that managers and employees in different sector of organization can link to wiki with intranet or internet and
according their knowledge and experience answer the question in their category. In addition the majority of participants in interview considered that wiki technology plays the role of reference in organization and internal knowledge needs of people encourage & motivate them for using this technology as a facilitating knowledge sharing in their organization. Also they were very satisfied with the features that wiki provides to the users, such as adding film, picture, voice or uploading files related to contents that user can share with the others. In wiki technology everyone can access to knowledge of other people and everybody can edit or add new entries in this system. Table1 reveals more clearly the advantages of this technology.

<table>
<thead>
<tr>
<th>No.</th>
<th>Most advantage of using wiki technology as facilitator of knowledge sharing in organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Collaborative authoring</td>
</tr>
<tr>
<td>2</td>
<td>Easy editing - knowledge can edited at any time by anyone</td>
</tr>
<tr>
<td>3</td>
<td>Easy citing and sourcing</td>
</tr>
<tr>
<td>4</td>
<td>Name of article is part of the hyperlink</td>
</tr>
<tr>
<td>5</td>
<td>Server storage of documents</td>
</tr>
<tr>
<td>6</td>
<td>Automatic Versioning and Difference Engines for documents</td>
</tr>
<tr>
<td>7</td>
<td>Search facilities</td>
</tr>
<tr>
<td>8</td>
<td>Automatic links to discussion pages</td>
</tr>
<tr>
<td>9</td>
<td>Most recent editions very visible; easy monitoring</td>
</tr>
<tr>
<td>10</td>
<td>Massively distributable collaboration</td>
</tr>
<tr>
<td>11</td>
<td>Groups and Categories</td>
</tr>
<tr>
<td>12</td>
<td>Easy links to multi-language documentation</td>
</tr>
</tbody>
</table>

6. CONCLUSION AND RECOMMENDATION

In this paper, we have analyzed the role of the information technology as facilitator of knowledge sharing in organization. We considered IT as a tool which is able to manage, store, and transmit structural knowledge is a critical solution for implementing impressive knowledge management. Also we realized the type of the IT’s tools are so important in quality of knowledge sharing. And we proposed new technology for better knowledge sharing that it is wiki technology. We understood wiki has more benefits than the rest of technologies that had been used for knowledge sharing in organization. The roads management center, as a successful case, with using the wiki technology can make suitable condition for their clients in knowledge sharing. Our recommendations for other organizations in implementing a successful knowledge management project is that before taking any action first understand their organization needs and then select an appropriate information technology as a tool in knowledge sharing. According to the results of our research, wiki is tested as proper technology amongst all other tools for sharing knowledge so we also recommend to other organization to use this technology in order to share knowledge within their organization.

REFERENCES


Posters
SOME KNOWLEDGE NORMALIZATION METHODS

Philippe Martin
ESIROI STIM (and adjunct researcher of Griffith University, Australia) - P.T.U. Bât. 2 - 2 rue Joseph Wetzell - 97490 Sainte Clotilde France

ABSTRACT
This article shows some high-level forms of lambda-abstractions that knowledge representation notations can allow to ease knowledge modeling and reduce information duplication. This is one structural method – or choice to make – to ease knowledge modeling. Other methods (structural, ontological, syntactical and lexical ones) are presented in an extended article at http://www.webkb.org/doc/bestPractices/articleOnBestPractices.pdf for improving the precision, inter-connection, comparison, retrieval and re-use of knowledge objects within and between ontologies on the Semantic Web. These choices can be enforced or seen as criteria to evaluate ontologies.

KEYWORDS
Knowledge Representation/Organization/Sharing/Normalization Methods

1. INTRODUCTION
Ontologies – or, more generally, knowledge bases (KBs) – within ontology repositories or the current Semantic Web, are most often small, single-authored, heterogeneous and poorly (or at least, not systematically) organized and inter-related. This is the way the Semantic Web is nowadays generally expected to be (Noy and Tudorache, 2008). There are at least four kinds of choices (resp. methods or best practices) that knowledge providers or knowledge engineers can make (resp. follow) for improving the precision and interconnection of knowledge objects within and between KBs, and therefore improve their re-usability. Such methods or choices are used in the shared knowledge server WebKB-2 and are presented in an extended article at http://www.webkb.org/doc/bestPractices/articleOnBestPractices.pdf. The next section presents one structural method: the use of high-level forms of lambda-abstractions in knowledge representation notations.

2. REDUCING ORGANIZATIONAL DUPLICATIONS AND NORMALIZING KNOWLEDGE VIA LAMBDA-ABSTRACTION LIKE SYSTEMS
Organizational duplication leads to cluttering and maintenance problems. To reduce it and to ease the task of knowledge providers, they should be proposed various simple ways to define a type without actually naming it (if it is named, it has to be manually or automatically inserted into the specialization hierarchy). The most general and common solution is to use a lambda-abstraction, i.e., a definition of an unnamed type within a statement, in place of a type, as with the expression in italics in the following belief in Formalized-English (FE), a notation used in WebKB-2 (Martin, 2002): pm# every `bird agent of a life with location wn#Sydney and period 2005 to 2007` can be agent of a flight` (informally: “according to pm, every bird living in ’Sydney’ (in the sense given to this word in WordNet) between 2005 and 2007 can fly”). However, classic lambda-abstractions are not always sufficient to ensure concise and normalized knowledge representations. Below are two derived but more automatic approaches that can be used in the notations designed for WebKB-2, such as FE. This illustrates the interest of using “high level notations”. Their constructs permit their users not to create types or complex expressions that people would either not be able to create or would often represent in ways that are not automatically comparable.
In FE, an expression such as `a French car´ is allowed and equivalent to `a car that is instance of French´ as well as (exists ((?c car)) (French ?c)) in KIF (Genesereth, 1988). In other words, this is a way to give multiple types to an unnamed object. However, the interpretation is different for the types that are subtypes of pm#attribute_or_quality_or_measure, e.g., wn#color, wn#red, own#red, own#redness (wn and own being abbreviations for WordNet 2.0 and OntoWordnet (Gangemi, Navigli and Velardi, 2003). Indeed, it would not be correct to represent the unnamed object as an instance of such types. own is a version of wn with a top-level ontology modified to be a specialization of DOLCE. Thence, some wn categories had to be re-interpreted to distinguish between "qualities" (e.g., size, color, redness, smell and duration) and “quale/regions” (categories of values for qualities, e.g., own#red and own#Greenwich_Mean_Time). The MSO integrates wn and own in a loss-less way, and hence i) has not modified the hierarchy of wn categories, and ii) use terms such as pm#attribute_or_quality_or_measure. Given that wn#red has for supertype wn#color and for subtypes own#redness and own#red which are subtypes of respectively dolce#quality and dolce#region, the expression `a wn#red car´ is interpreted as `a car that has for pm#attribute a wn#color that has for pm#measure a wn#red´. The term wn#red may have subtypes, and hence is a type and has to be quantified. Similarly, the expression ‘a car with wn#weight 923.5 wn#kg’ is allowed and interpreted as ‘a car that has for pm#attribute a wn#weight that has for pm#measure a wn#measure that has for pm#unit a wn#kg and for pm#value 923.5’.

In FE, an expression such as ‘a car with wn#color a wn#red´ is also allowed (and is equivalent to ‘a wn#red car´). Similarly, ‘John is wn#eating an apple´ is allowed and is a shortcut for ‘John is pm#agent of a wn#eating with pm#object an apple´. This means that, as in some other systems such as Ontoseek (Guarino and Welty, 2003), FE allows the use of concept types in place of relation types. However, to that end, FE, FL and FCG allow the association of a signature to the concept type (see Table 2 for the two allowed kinds of signature) and impose three conditions. For a concept type CT to be usable in a relation node R, as if it was a relation type, i) CT must be a subtype of pm#thing_that_can_be_seen_as_a_relation, ii) if CT has an “associated signature”, the source and destination of R must conform to that signature, and iii) if CT does not have a direct or inherited “associated signature”, the destination R is of type CT. In the MSO, pm#thing_that_can_be_seen_as_a_relation has for main direct subtypes wn#relation, pm#description_content/medium/container, pm#entity_playing_a_role and pm#attribute_or_quality_or_measure. This permits to re-use a good part of WordNet types as relation types, and avoids duplicating them in the relation type hierarchy.

Table 1. Declarations in FL of concept types with two different kinds of associated signature

<table>
<thead>
<tr>
<th>pm#process .(pm#agent: pm#causal_entity, pm#object: pm#entity)</th>
<th>pm#situation; /\ '^*'; pm#supertype</th>
</tr>
</thead>
<tbody>
<tr>
<td>father .(animal, male)</td>
<td>^&lt; parent;</td>
</tr>
</tbody>
</table>

The first kind of signature permits to define shortcuts such as the one illustrated above. The second kind is like a relation type signature. When classic knowledge representation languages (KRLs) are used for exporting knowledge, such signatures permit to automatically generate and use the corresponding relation types.

The constructs presented in this section ease knowledge representation, normalize it and reduce organizational duplication. Most ontologies and UML schema declare their own ad-hoc types for representing entities, attributes, measures and relations between them, thus preventing automatic comparisons (and hence retrieval or integrations) of the knowledge objects. The re-use of a general multi-source ontology such as the MSO and the use of high-level notations would strongly ease automatic comparisons. Expressions such as ‘a wn#red car´ and ‘a car with wn#weight 923.5 wn#kg’ only need to be expanded when they have to be compared with expressions not using such shortcuts. Such an expansion can be done in various ways to suit various ways of representing relations between entities, attributes and measures. In its “Semantic Web Best practices” (Swick, Schreiber and Wood, 2006) and tutorials, the W3C proposes or uses various ways of representing such relations. Some are ontology incorrect, e.g., when no type is used for kinds of values which, like Red, clearly can have sub-kinds, or when a type of datatype such as Integer is specialized by a type of quality such as Age (Carroll and Pan, 2006). Some lead to a strong manual organizational duplication, e.g., Approach 3 in (Uschold and Welty, 2005). Most essentially used ad-hoc types such as has_health_status and Good_health_value.
3. CONCLUSION

To ease knowledge representation and increase knowledge retrieval, comparison, understanding and re-use possibilities, instead of more or less independently creating RDF/XML+OWL files, the extended version of this article argues for the use of servers for cooperatively-built well-organized KBs (cbwoKB), federated into global virtual cbwoKBs, at-least-minimally-well-organized, integrating in a loss-less way many top-level ontologies and lexical ontologies, advocating best practices, having various complementary high-level expressive syntaxes that ease the following of some best practices, and having KB editing protocols that enforce the following of every best practice that a user has committed to follow. Then, when applications are designed, RDF/XML+OWL files can be generated since this is what this language is intended for. This distinction between application language and knowledge modeling means is a classic one (Schreiber et al., 1994) (Guizzardi et al., 2010).

To compare existing best practices and follow them while presenting them, these best practices should themselves be represented and organized by various relations (especially, specialization, subtask and argumentation ones) into a cbwoKB where they can be complemented by Web users and argued for or against in an organized, normalized, formal or semi-formal way. This is a work in progress, within the MSO (Multi Source Ontology), the default KB of WebKB-2: best practices have been represented and organized into its specialization hierarchy; various ones have been the subject of semi-formal structured discussions in the FL notation (especially those debated by the Standard Upper Ontology group, related to the use of RDF/XML and non-binary relations). The main concepts and tasks of the KADS methodology (Schreiber et al., 1994) have also been represented and organized. As previously noted, top-level ontologies are artifacts for following methodologies or best practices, and many top-level ontologies have already been integrated in the MSO.

Ontology design patterns (Presutti and Gangemi, 2008) are other kinds of best practices. Many current ones are for ontology alignment tasks. Most are currently described rather informally. Their collaborative representation and organization via a cbwoKB (e.g., via WebKB-2 in the MSO) would have the above cited advantages. More generally, one of the medium-term applications of WebKB-2 is the collaborative creation of a semi-formal semantically organized state-of-the-art in knowledge engineering by researchers in this domain.

A medium/long term next step for WebKB-2 is the representation of its search/comparison operators, knowledge evaluation procedures and display procedures in a declarative way, for the users to adapt. I proposed many steps made in that direction, e.g., a general ontology for notations exploitable to allow users to adapt their favorite input and output notations (FE, FL, KIF, RDF/XML, ...) to their needs or preferences (this would not change the underlying knowledge model).

REFERENCES

A DOMAIN SPECIFIC LANGUAGE TO REPRESENT STRUCTURED DATA FOR MULTIPLE PROGRAMMING LANGUAGES

Kazuaki Maeda

Department of Business Administration and Information Science - Chubu University, 1200 Matsumoto, Kasugai, Aichi 487-8501, Japan

ABSTRACT

This paper describes RugsOn, a new representation written in a text-based data format. The design principle of RugsOn is good readability and simplicity of structured data representation. One feature of RugsOn is a domain specific language to represent structured data. It is similar as JSON, but syntax of RugsOn is carefully chosen for multiple programming languages. A program generator was developed to create Ruby, Groovy, Scala and Java programs from RugsOn definitions. In the author's experience, productivity was improved in the design and implementation of programs that manipulate structured data.

KEYWORDS

Data Representation, Structured Data, Domain Specific Languages

1. INTRODUCTION

There are many applications (e.g. graph drawing and editing tools), which have their own representation for structured data. The representation has typically a textual format that is easy to read and write. XML is mostly used as a standard language for the data representation. If the structured data can be defined in a language, a validation tool using the definition is very useful to find human errors in entering data manually. XML Schema provides a means to define the representation in XML. The schema and the representation in XML are written in human readable text format, but they are composed of many redundant tags including start tags and end tags, so that there are some cases where it is difficult to read and understand them.

JSON is one of popular representations for structured data. The structured data in JSON is represented by object notation of JavaScript and it is easier to read and write than the representation in XML. A draft of JSON schema is now open for discussion. It is a JSON-based format to define the representation in JSON. According to the draft, however, we have to use link description objects to define graph-structured data in JSON. It is not easy task to define the graph-structured data containing hundreds of kinds of nodes.

Domain-specific language (DSL) is a computer language of limited expressiveness focused on a particular domain. Martin Fowler's book (Fowler 2011) describes external DSLs and internal DSLs. External DSLs have their own custom syntax. In the applications with the external DSLs, we need parsers using text-parsing techniques. XML syntax is completely different from general-purpose languages such as Java, C++ and Ruby, but XML is frequently chosen as an external DSL to describe structured data.

JSON is an internal DSL in developing JavaScript programs. It is easy to build programs to process the representation in JSON and it gives a simple way to program if we choose JavaScript as a host language. However, if we choose different program languages from JavaScript, such as Java and C++, the representation in JSON is parsed to a hierarchical tree for manipulation in the main memory. It is similar as an external DSL like XML.

From these considerations, the representation of structured data should be simple and easy to read. Moreover, structured data should be represented using an internal DSL and it should be available for multiple programming languages. This paper describes RugsOn (which stands for Ruby, Groovy and Scala becoming Object notation), a new representation of structured data in a text-based data format for multiple languages,
Ruby, Groovy and Scala. A common subset of Ruby, Groovy and Scala was carefully chosen for RugsOn design so that one representation in RugsOn is available for multiple programming languages.

2. REPRESENTATION AND DEFINITION OF STRUCTURED DATA

The representation in RugsOn is composed of several elements. For example, here is an element "www".nodeName which represents the value as "www" and the name of the element as nodeName. The element is not only a data representation, but also it is an executable method invocation without parentheses in programming languages (i.e. Ruby, Groovy and Scala). The method name is nodeName and the receiver is "www."

In RugsOn, a structure is represented using a block as an argument. For example, Figure 1 shows that node has four child elements: nodeName, kind, size and _id. In the case of the first node element, nodeName element's value is "www," the kind element's value is 1, the size element's value is 0.7, and _id element's value is "id0."

An element to represent a collection can have more than one element with the same name. In Figure 1, the nodes element has three child elements with the same name. The first node element has a nodeName element with a value "www," the second node element has a nodeName element with a value "mccsis," and the third node element has a nodeName element with a value "org." This represents a collection of node elements.

If we need to represent an element linked to another element across the structure, a unique identifier, called _id, is given to the element, and another element refers to the element using the identifier. In Figure 1, a _id element with the identifier "id0" is given to the first node element, and another _id element with the identifier "id1" is given to the second node element.

The first character ':' in the string constant, for example "':id0," means a reference to another element. The source element in the first link element has a reference to the identifier "id0," and the target element in the first link element has a reference to another identifier "id1." The representation shows links across the structure so that RugsOn supports graph-structured data.

![Figure 1. Graph-structured data in RugsOn and the diagram with the logical meanings](image-url)
Figure 2 shows the definitions of the representation in Figure 1. The definition language was designed using a subset of Ruby syntax so that it is an internal DSL in Ruby. The definitions are shown as follows:

- `cloneGraph` element has two child elements: `nodes` and `links`
- `nodes` element is a collection of `node` elements
- `links` element is a collection of `link` elements
- `node` element has three child elements: `nodeName`, `kind`, and `size`
- `kind` element has an integer value
- `size` element has a double value
- `link` element has four child elements: `source`, `target`, `kind`, and `visible`
- `source` element is a `node` type
- `target` element is a `node` type
- `visible` element has a boolean value

RugsGen generates Ruby programs, Groovy programs, Scala programs, Java programs, and a class diagram for the generated classes. In the right side of Figure 2, some keywords are used as follows:

- `java_root_class` specifies a base class of all other classes
- `java_prefix` specifies addition of a prefix “Grph” at the beginning of the Java class name
- `generate_java` specifies generation of Java programs with the package path “org.graph1”
- `generate_ruby` specifies generation of RugsOn-related Ruby programs under the directory “org.graph2”
- `generate_groovy` specifies generation of RugsOn-related Groovy programs with the package path “org.graph3”
- `generate_scala` specifies generation of RugsOn-related Scala programs with the package path “org.graph4”
- `generate_diagram` specifies generation of a class diagram for generated Java classes to the file name “GrpDiagram.dot”

3. CONCLUSION

This paper described RugsOn, a new representation written in a text-based data format for multiple programming languages. A program generator was developed to create Ruby, Groovy, Scala and Java programs from RugsOn definitions. RugsOn and its related tools are now being used in applications supporting the development of commercial products, such as a diagram editor and a compiler front-end. The development and results will be published in a future paper.

REFERENCES

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmad, N</td>
<td>99</td>
</tr>
<tr>
<td>Alhmiedat, T</td>
<td>75</td>
</tr>
<tr>
<td>Alinia, H</td>
<td>132</td>
</tr>
<tr>
<td>Arnold, H</td>
<td>107</td>
</tr>
<tr>
<td>Asif, M</td>
<td>99</td>
</tr>
<tr>
<td>Breuer, H</td>
<td>107</td>
</tr>
<tr>
<td>Bukhtoyarov, V</td>
<td>35</td>
</tr>
<tr>
<td>Caballero-Gil, C</td>
<td>122</td>
</tr>
<tr>
<td>Caballero-Gil, P</td>
<td>122</td>
</tr>
<tr>
<td>Cruickshank, H</td>
<td>99</td>
</tr>
<tr>
<td>Darmaros, T</td>
<td>3</td>
</tr>
<tr>
<td>Delavar, M</td>
<td>132</td>
</tr>
<tr>
<td>Detschew, V</td>
<td>59</td>
</tr>
<tr>
<td>Dong, T</td>
<td>91</td>
</tr>
<tr>
<td>Ekinci, O</td>
<td>59</td>
</tr>
<tr>
<td>Entrialgo, J</td>
<td>10</td>
</tr>
<tr>
<td>Frantis, P</td>
<td>127</td>
</tr>
<tr>
<td>García, D</td>
<td>10</td>
</tr>
<tr>
<td>Gattnar, E</td>
<td>59</td>
</tr>
<tr>
<td>Goede, R</td>
<td>51</td>
</tr>
<tr>
<td>Grabowski, H</td>
<td>107</td>
</tr>
<tr>
<td>Groh, G</td>
<td>137</td>
</tr>
<tr>
<td>Guy, C</td>
<td>143</td>
</tr>
<tr>
<td>Hasan, M</td>
<td>117</td>
</tr>
<tr>
<td>Hasoon, F</td>
<td>117</td>
</tr>
<tr>
<td>Hasoon, N</td>
<td>117</td>
</tr>
<tr>
<td>Hattori, S</td>
<td>67</td>
</tr>
<tr>
<td>He, D</td>
<td>83</td>
</tr>
<tr>
<td>Hodicky, J</td>
<td>127</td>
</tr>
<tr>
<td>Kekwaletswe, R</td>
<td>18</td>
</tr>
<tr>
<td>Khanjari, A</td>
<td>161</td>
</tr>
<tr>
<td>Kheirabadi, A</td>
<td>161</td>
</tr>
<tr>
<td>Latip, R</td>
<td>149</td>
</tr>
<tr>
<td>Lehmann, A</td>
<td>137</td>
</tr>
<tr>
<td>Li, S</td>
<td>83</td>
</tr>
<tr>
<td>Maeda, K</td>
<td>172</td>
</tr>
<tr>
<td>Martin, P</td>
<td>169</td>
</tr>
<tr>
<td>Medrano, R</td>
<td>10</td>
</tr>
<tr>
<td>Mohseni, A</td>
<td>132</td>
</tr>
<tr>
<td>Molina-Gil, J</td>
<td>122</td>
</tr>
<tr>
<td>Musangu, L</td>
<td>18</td>
</tr>
<tr>
<td>Nguyen, T</td>
<td>91</td>
</tr>
<tr>
<td>Petrova-Antonova, D</td>
<td>43</td>
</tr>
<tr>
<td>Rajabion, L</td>
<td>161</td>
</tr>
<tr>
<td>Rezaei, K</td>
<td>161</td>
</tr>
<tr>
<td>Roth, J</td>
<td>27</td>
</tr>
<tr>
<td>Sahar, A</td>
<td>149</td>
</tr>
<tr>
<td>Saini, D</td>
<td>117</td>
</tr>
<tr>
<td>Saka, O</td>
<td>143</td>
</tr>
<tr>
<td>Sasaki, T</td>
<td>154</td>
</tr>
<tr>
<td>Sawamoto, J</td>
<td>154</td>
</tr>
<tr>
<td>Sergienko, R</td>
<td>35</td>
</tr>
<tr>
<td>Souza, M</td>
<td>137</td>
</tr>
<tr>
<td>Sun, Z</td>
<td>83, 99</td>
</tr>
<tr>
<td>Vassilakopoulou, P</td>
<td>3</td>
</tr>
<tr>
<td>Wada, Y</td>
<td>154</td>
</tr>
<tr>
<td>Yang, S</td>
<td>75</td>
</tr>
<tr>
<td>Zare, M</td>
<td>132</td>
</tr>
</tbody>
</table>