

Simon Fong (Ed.)

Communications in Computer and Information Science

136

# Networked Digital Technologies

Third International Conference, NDT 2011  
Macau, China, July 2011  
Proceedings



Springer



Simon Fong (Ed.)

# Networked Digital Technologies

Third International Conference, NDT 2011  
Macau, China, July 11-13, 2011  
Proceedings

Volume Editor

Simon Fong  
University of Macau  
Faculty of Science and Technology, Building N, Room N410  
Av. Padre Tomas Pereira, Taipa, Macau, PR China  
E-mail: ccfong@umac.mo

ISSN 1865-0929  
ISBN 978-3-642-22184-2  
DOI 10.1007/978-3-642-22185-9  
Springer Heidelberg Dordrecht London New York

e-ISSN 1865-0937  
e-ISBN 978-3-642-22185-9

Library of Congress Control Number: 2011930275

CR Subject Classification (1998): I.2, H.3, H.4, C.2, H.5, J.1

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Message from the Chairs

We are honored to present the final papers of the prestigious conference the Third International Conference on Networked Digital Technologies (NDT 2011).

The NDT conference has gained significance in the last couple of years and we hope to make it a standardized conference in computer and information sciences. NDT 2011 was co-sponsored (published) by Springer like the previous editions. The University of Macau was happy to organize this scholarly conference in July 2011.

NDT 2011 aimed to provide support for concerted efforts in building federated digital technologies that will enable the formation of a network of digital technologies.

NDT 2011 consisted of invited talks, papers and presentations. We accepted 41 papers out of 127 submissions. We believe that the conference stimulated interesting discussions.

We are grateful to the University of Macau for hosting this conference. We use this occasion to express our thanks to the Technical Committee and to all the external reviewers. We are grateful to Springer for co-sponsoring the event. Finally, we would like to thank all the participants and sponsors.

April 2011

Editors  
Simon Fong  
Pit Pichappan

# Table of Contents

## Information Security

Comparison between PKI (RSA-AES) and AEAD (AES-EAX PSK) Cryptography Systems for Use in SMS-Based Secure Transmissions . . . . .	1
<i>Hao Wang and William Emmanuel Yu</i>	
Authentication and Authorization in Web Services . . . . .	13
<i>Khalil Challita, Hikmat Farhat, and Joseph Zalaket</i>	
Integrating Access Control Mechanism with EXEL Labeling Scheme for XML Document Updating . . . . .	24
<i>Meghdad Mirabi, Hamidah Ibrahim, Ali Mamat, and Nur Izura Udzir</i>	
An Access Control Model for Supporting XML Document Updating . . . . .	37
<i>Meghdad Mirabi, Hamidah Ibrahim, Leila Fathi, Nur Izura Udzir, and Ali Mamat</i>	
A Secure Proxy Blind Signature Scheme Using ECC . . . . .	47
<i>Daniyal M. Alghazzawi, Trigui Mohamed Salim, and Syed Hamid Hasan</i>	
Accelerated Particle Swarm Optimization and Support Vector Machine for Business Optimization and Applications . . . . .	53
<i>Xin-She Yang, Suash Deb, and Simon Fong</i>	

## Networks

A Study on the Reliability of Data Transmission of an over the Top Network Protocol on SMS versus UDP/ GPRS (Fragmented) . . . . .	67
<i>Alyssa Marie Dykimching, Jan Aaron Angelo Lee, and William Emmanuel Yu</i>	
QuickFlood: An Efficient Search Algorithm for Unstructured Peer-to-Peer Networks . . . . .	82
<i>Hassan Barjini, Mohamed Othman, and Hamidah Ibrahim</i>	

Improved-XY: A High Performance Wormhole-Switched Routing Algorithm for Irregular 2-D Mesh NoC .....	93
<i>Ladan Momeni, Arshin Rezazadeh, and Davood Abednejad</i>	
XRD Metadata to Make Digital Identity Less Visible and Foster Trusted Collaborations across Networked Computing Ecosystems .....	105
<i>Ghazi Ben Ayed and Solange Ghernaouti-Hélie</i>	
An Overview of Performance Comparison of Different TCP Variants in IP and MPLS Networks.....	120
<i>Madiha Kazmi, Muhammad Younas Javed, and Muhammad Khalil Afzal</i>	
Routing in Mobile Ad-Hoc Networks as a Reinforcement Learning Task .....	128
<i>Saloua Chettibi and Salim Chikhi</i>	
<b>Information Management</b>	
Identifying Usability Issues in Personal Calendar Tools .....	136
<i>Dezhi Wu</i>	
Adaptive Query Processing for Semantic Interoperable Information Systems .....	147
<i>Benharzallah Saber, Kazar Okba, and Guy Caplat</i>	
Statistical Character-Based Syntax Similarity Measurement for Detecting Biomedical Syntax Variations through Named Entity Recognition .....	164
<i>Hossein Tohidi, Hamidah Ibrahim, and Masrah Azrifan Azmi</i>	
Correction of Invalid XML Documents with Respect to Single Type Tree Grammars .....	179
<i>Martin Svoboda and Irena Mlýnková</i>	
Identification of Scholarly Papers and Authors .....	195
<i>Kensuke Baba, Masao Mori, and Eisuke Ito</i>	
A Social Network Model for Academic Collaboration .....	203
<i>Sreedhar Bhukya</i>	
Performance Evaluation of Preference Evaluation Techniques .....	212
<i>Alwan A. Ali, Ibrahim Hamidah, Tan Chik Yip, Sidi Fatimah, and Udzir Nur Izura</i>	
Collective Action Theory Meets the Blogosphere: A New Methodology .....	224
<i>Nitin Agarwal, Merlyna Lim, and Rolf T. Wigand</i>	

## Multimedia

Evaluating <i>K</i> -Best Site Query on Spatial Objects . . . . .	240
<i>Yuan-Ko Huang and Lien-Fa Lin</i>	
Performance of Annotation-Based Image Retrieval . . . . .	251
<i>Phani Kidambi, Mary Fendley, and S. Narayanan</i>	
Multimedia Streams Retrieval in Distributed Systems Using Learning Automata . . . . .	269
<i>Safiye Ghasemi and Amir Masoud Rahmani</i>	
Real-Time Detection of Parked Vehicles from Multiple Image Streams . . . . .	280
<i>Kok-Leong Ong and Vincent C.S. Lee</i>	

## E Learning and E Government

An Adaptive Framework for Personalized E-Learning . . . . .	292
<i>Ronnie Cheung and Hassan B. Kazemian</i>	
Gov 2.0 and Beyond: Using Social Media for Transparency, Participation and Collaboration . . . . .	307
<i>F. Dianne Lux Wigand</i>	

## Web Services/Semantics

MATAWS: A Multimodal Approach for Automatic WS Semantic Annotation . . . . .	319
<i>Cihan Aksoy, Vincent Labatut, Chantal Cherifi, and Jean-François Santucci</i>	
Meshing Semantic Web and Web2.0 Technologies to Construct Profiles: Case Study of Academia Europea Members . . . . .	334
<i>Petra Korica-Pehserl and Atif Latif</i>	
Towards Ontology-Based Collaboration Framework Based on Messaging System . . . . .	345
<i>Gridaphat Sriharee</i>	
A QoS and Consumer Personality Considered Services Discovery . . . . .	357
<i>Xiuqin Ma, Norrozila Sulaiman, and Hongwu Qin</i>	
User Centric Homogeneity-Based Clustering Approach for Intelligence Computation . . . . .	364
<i>Yun Wei Zhao, Chi-Hung Chi, and Chen Ding</i>	



## Data Mining

Mining Temporal Association Rules with Incremental Standing for Segment Progressive Filter . . . . .	373
<i>Mohsin Naqvi, Kashif Hussain, Sohail Asghar, and Simon Fong</i>	
Multi-way Association Clustering Analysis on Adaptive Real-Time Multicast Data . . . . .	383
<i>Sheneela Naz, Sohail Asghar, Simon Fong, and Amir Qayyum</i>	
Multi Level Mining of Warehouse Schema . . . . .	395
<i>Muhammad Usman and Russel Pears</i>	
On Inserting Bulk Data for Linear Hash Files. . . . .	409
<i>Satoshi Narata and Takao Miura</i>	

## Cloud Computing

Cloud Data Storage with Group Collaboration Supports . . . . .	423
<i>Jyh-Shyan Lin</i>	
A Negotiation Mechanism That Facilitates the Price-Timeslot-QoS Negotiation for Establishing SLAs of Cloud Service Reservation . . . . .	432
<i>Seokho Son and Kwang Mong Sim</i>	
<b>Author Index</b> . . . . .	447

# Comparison between PKI (RSA-AES) and AEAD (AES-EAX PSK) Cryptography Systems for Use in SMS-Based Secure Transmissions\*

Hao Wang and William Emmanuel Yu

Ateneo de Manila University

**Abstract.** In today's mobile communication systems, security offered by the network operator is often limited to the wireless link. This means that data delivered through mobile networks are not sufficiently protected. In the particular growing field of interest of machine-to-machine (M2M) communications, these applications typically require a mobile, secure and reliable means of data communication. This paper compared two (2) cryptographic mechanisms, the RSA-AES and the AES-EAX PSK which provide end-to-end security for SMS-based transmission. We implemented these two (2) mechanisms assuming the constraints of standard SMS network and measured their performance in terms of transaction time. Our study indicated that in terms of processing time, the Authenticated Encryption and Associate Data (AEAD) modes represented by EAX performed better even when the digital signature of the Public Key Infrastructure (PKI) mode represented by RSA was not included.

**Keywords:** Cryptography, Encryption, RSA, EAX, GSM, SMS.

## 1 Introduction

The Global System for Mobile Communications (GSM) is a common standard issued by the European Telecommunications Standards Institute (ETSI). Phase I of the GSM specification was published in 1990 and is currently the most widely used mobile phone system in the world. The Short Message Service (SMS) standard was first discussed in the early 1980s but the world's first commercial SMS service was not introduced until 1992. SMS was created as part of Phase I of the GSM standard. SMS is widely adopted with approximately one (1) billion SMS messages sent every day only in the Philippines<sup>[1]</sup>.

Recently, a survey carried by the Internet Data Center (IDC) shows that more than 90% of mobile users prefer SMS as their main communication tool<sup>[3]</sup>. The report has concluded that with the statistic of 65% of the mobile users sending text messages every day, SMS will continue to play an important role as the most popular mobile data application for a few more years. This also goes to show that network operators have invested significantly in ensuring the optimal performance of their SMS networks.

---

\* This work was supported by the Department of Information Systems and Computer Science of the Ateneo de Manila University.

With the rise of mobile communications and commerce and the increasingly wide use of machine-to machine (M2M) communication applications, such as the fields of Automatic Teller Machine (ATM) banking, telemetry and telematics, navigation, smart metering and many others, a mobile, secure and reliable means of data communication is a primary necessity. Currently, the SMS M2M networks have become a popular means of transmitting the sensitive information necessary for these applications. However, SMS security needs to be improved.

## 2 Statement of the Context

ABI Research estimates that the total number of cumulative global M2M connections rose from 46.78 million connections in 2007 to 71.09 million cumulative connections in 2009, and this number is still growing[21]. M2M market boosted by thriving technologies, and is currently being applied widely; some of the use cases involve financial, telemetry and telematics, navigation, logistics and voting systems. The most widely available data service in GSM networks today is SMS. This is why we focus on SMS for this study. However, the current GSM data transmission in some cases cannot provide a secure and stable environment. So its security has become an increasingly important issue. In particular, some specific M2M applications (such as ATM banking, POS machines, voting systems) need a higher level of security than currently provided by mobile networks.

When sensitive information is exchanged using SMS, it is crucial to protect the content from eavesdroppers. By default, SMS content is sent over the Global System for Mobile communications (GSM) network in clear text form, or in a predictable format[20]. The message sent from the mobile device will store at the message centre of associate network provider. The message will travel across different base station in unprotected manner. This means there is an opportunity to allow the middle man attack on those confidential messages. Moreover, this allows an attacker with the right equipment to eavesdrop on the information that is being sent. Another problem with SMS is that the originating address (OA) field in the SMS header can be forged, thus allowing masquerading and replay attacks. Therefore SMS is not totally secure and cannot always be trusted. For example, there has been at least one case in the UK where SMS information has been abused by the operator employees[20].

In some cases, SMS messages are encrypted using a family of cryptography algorithms collectively called A5. A5/1 is the “standard” encryption algorithm, which was used by about 130 million customers in Europe. While A5/2 is the “export” (weakened) algorithm, which was used by another 100 million customers in other markets. A5/3 is a new algorithm based on the UMTS/WCDMA algorithm Kasumi[13].

However, a number of attacks on A5 have been published[14][12][22]. Some require an expensive pre-processing stage after which the cipher can be attacked in minutes or seconds. Until 2000, the weaknesses have been passive attacks using the known plaintext assumption. In 2003, more serious weaknesses were identified which can be exploited in the ciphertext-only scenario, or by an active

attacker. In 2006, Elad Barkan, Eli Biham and Nathan Keller demonstrated attacks against A5/1, A5/3, or even GPRS that allow attackers to tap GSM mobile phone conversations and decrypt them either in real-time, or at any later time. It follows that the current GSM network does not provide end-to-end security services even with A5 [17]. This requires system to provide external privacy guarantees.

### 3 Statement of the Objectives

The objective of this study is to implement and compare the performance of a PKI and an AEAD encryption system for securing SMS-based transmission networks [26] in terms of transaction time. We first introduce a PKI-based mechanism on the Rivest, Shamir and Adleman (RSA) algorithm [23]. Followed by describing an Authenticated Encryption and Associate Data (AEAD) mechanism called EAX (AES-EAX PSK) [10]. Both these systems are used to provide privacy/confidentiality, integrity and authenticity as security guarantees. Then, we describe the implementation of both mechanisms. Finally, we evaluate and compare the performance in terms of transaction time between these two mechanisms.

### 4 Scope and Limitation of the Study

Cryptography does not “solve” computer security. Security is always relative. It’s hard to say that any cryptography algorithm is always safe. With the hardware and network development, or there is a probability that the current encryption algorithms used are likely to be cracked sooner or later, then we have to use a longer key or more advanced algorithms to ensure data security. These cryptography algorithms, therefore, still need to be constantly developed and improved, providing greater strength and speed.

In this study, the computing system used and platforms are controlled, the payload for both was also controlled. Key sizes used were based on equivalent strength provided. In order for these security mechanisms to be used in the SMS-based transmission network, the final payload must be broken into fragments 140 bytes which is the maximum amount of data an SMS can carry [6].

### 5 Security and Mobile Networks

In this section, we present an overview of the required security guarantees and current state of mobile network security.

#### 5.1 Security Guarantees

In the current scheme of information security practice, there are some specific security guarantees that we require to consider a service secure. The ISO 17799 [2] names the following guarantees:

1. Authentication: The process of guaranteeing the identity of the message sender.

2. Privacy/confidentiality: Ensuring that no one can read the message except the intended receiver.
3. Integrity: Assuring the receiver that the received message has not been altered in any way from the original.

For this study, we used these three (3) security guarantees as the baseline for designing our security mechanisms.

## 5.2 State of Mobile Network Security

SMS is a highly suitable bearer given its pervasiveness, reliability and integrity. The payload is small, with only 160 ASCII characters or 140 bytes for binary-encoded messages, which results in a highly efficient means for transmitting short bursts of data [6]. SMS is globally available, and requires no further external protocols or provisioning since it is a complete, two-way delivery system native to the GSM protocol. SMS is delivered to a GSM network that will further the message to the necessary recipient or service [15].

However, the most pervasively deployed GSM encryption algorithm, A5, is now considered ineffective. Some solutions exist that only require an expensive pre-processing stage after which the cipher can be attacked in minutes or seconds [13] [16] [8]. There are a number of studies that cover the safety aspect of SMS transmission. In discussing the weaknesses of the SMS, Lo et al [19] suggested a PKI-based approach to overcome SMS communications security problems. This is an over-the-top approach, the principles of which can be implemented in other network packages or mechanisms. To ensure the secure transport of keys, PKI, particularly RSA, is employed as the key exchange mechanism. In addition, AES in CTR mode and HMAC with SHA256 are used for integrity and privacy, which, according to Bellare et al [11], is an example of a non-composite scheme.

In this study however, key exchange was not considered; the use of pre-shared keys was assumed; and a composite authenticated encryption scheme was employed.

A good overview of the built-in security infrastructure of today's mobile networks is given by Schmidt [24] who faults current security mechanisms such as A5 and A3 as potentially weak and untrustworthy. Existing crypto-system, he notes, does not provide some security guarantees such as non-repudiation. However, this can be supplemented by an additional security infrastructure such as TLS/SSL, he recommended, which is a PKI-based approach.

Another research by Abidrahman et al [7] compared the Secure Hash Algorithm (SHA) family and provided estimates of the amount of increase in energy, area and time consumption. After reviewed the standard SHA family members' designs, the results and the compatibility of the SHA algorithms for Wireless Sensor Networks (WSNs) were implemented on hardwares. The author indicated the feasibility of SHA-2 family algorithms as a replacement of the broken SHA-1 and Message-Digest 5 (MD-5) algorithms for WSNs. SHA-256 is shown as the better energy consumption per block.

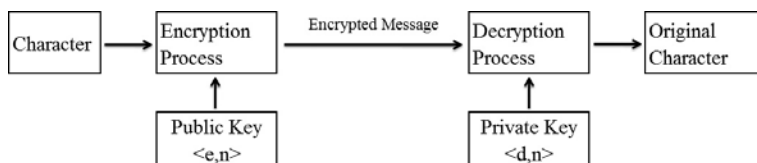
## 6 Framework

The study compared the performance of a PKI and AEAD cryptography systems for use in SMS-based secure transmissions in terms of transaction time. AES algorithm with 128-bit keys was used to serve as the baseline block cipher. A comparison between these two mechanisms was attempted in terms of security guarantees and performance.

### 6.1 RSA-AES Mechanism

The most popular PKI algorithm used is RSA. The algorithm generates two keys, a public key and a private key, by manipulating two prime numbers with a series of computations. The public key distributed publicly and the private key can be kept secretly by the user. This is to ensure that the secure message, which was encrypted using the recipient's public key, will be read by the targeted person, with the private key to decrypt the encryption. Furthermore, the public key can be used to verify the digital signature which is signed with the sender's private key. The RSA scheme is a block cipher in which the original non-ciphered text and cipher text are integers between 0 and  $n-1$  for some 'n'. That is, the block size of RSA is determined by the bit length of the integer 'n' and regarded as the key size of the RSA scheme [23].

Decryption: For a given cipher text  $C$ , the original non-ciphered text is computed by  $M = C d \text{ mod } n$ .



**Fig. 1.** The RSA procedure for sending an encrypted short message

As RSA requires very large prime numbers, it is impractical to use it to encrypt the entire payload. So RSA cryptography is used to encrypt the keys. In this study, 128 bit AES cryptography is used first to encrypt the data, specifically the Cipher Block Chaining (CBC); then RSA is used to signature and encrypt the transaction key of AES. This mechanism is similar to the one used by Transport Layer Security and Secure Sockets Layer (TLS/SSL) [18].

### 6.2 AES-EAX PSK Mechanism

EAX is an  $n$ -bit mode of operation. This allows the mode to operate on AES with 128 bits; or Secure Hash Algorithm (SHACAL-2) with its 256 bit block size. EAX is online, meaning the data does not need to be known in advance; it can be streamed into the object though there are some practical implementation

constraints. The AES-EAX PSK scheme is an Authenticated Encryption with Associated Data (AEAD) algorithm designed to simultaneously provide both authentication and privacy of the message (Authenticated encryption) with a two-pass scheme, one pass for achieving privacy and one for authenticity for each block[10]. EAX is also shown to provide all three (3) required security guarantees: privacy, integrity and authentication[9].

### 6.3 Overview Of Cryptography Mechanisms

In the RSA-AES mechanism, RSA is used to encrypt the transaction keys. AES in CBC mode is then used to encrypt the data with the transaction key for transmission. Finally, RSA is then used to sign the payload. In the AES-EAX PSK mechanism, EAX is used with the AES block cipher using a pre-shared key (PSK).

Table 1 shows the comparison between RSA-AES and AES-EAX PSK according to protection guarantees they provide.

**Table 1.** Comparison between RSA-AES and AES-EAX PSK according to protection guarantees they provide

	RSA-AES	AES-EAX PSK
Authentication	RSA	AES-EAX
Privacy/confidentiality	AES-CBC	AES-EAX
Integrity	RSA	AES-EAX
Non-repudiation	RSA	AES-EAX
Key exchanging support	RSA	PSK
Key exchanging size	1024 bit	128 bit
Payload key size	128 bit	128 bit

## 7 Methodology

### 7.1 Tools

Python is an object-oriented, literal-style computer programming language, and has a history of more than ten years of development, maturity and stability[5]. Python has a very large library, they can be quickly adopted by most common tasks, such as: string processing (regular expressions, Unicode, calculating differences between files), Internet protocols (HTTP, FTP, SMTP, XML-RPC, POP, IMAP, CGI programming), software engineering (unit testing, logging, profiling, parsing Python code), and operating system interfaces (system calls, file systems, TCP/IP sockets).

Botan is a BSD-licensed cryptographic library written in C++ and with Python bindings[4]. It is one of the few libraries which can provide complete and functional AES-CBC, AES-EAX and RSA cryptographic algorithms. In this paper, we have used a development build of Botan with Python bindings and the RSA-PrivateKey fix for Fedora 13[25].

## 7.2 Methodology

First of all, the composite encryption modes were used in this study, and the process for data splitting, combination, encryption and decryption was coded by Python and the Botan cryptography libraries. The target environment is described in Table 2.

**Table 2.** Experiment environment

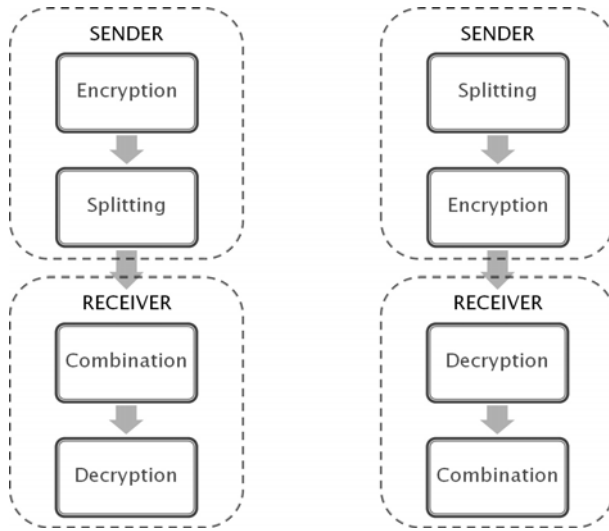
Operating System	Linux Fedora 13
Program Language	Python 2.6
Cryptography Library	Botan 1.9.3 with RSA-Private Key Fix
Computer CPU	Intel Core Duo T7300 2.00GHz
Computer Memory	1024MB

Second, there are two different modes of transmitter flows. Figure 2 shows the two different kinds of transmitter flow.

Based on these two modes, four (4) different schemes were compared as follows:

1. For the sender, encrypt the entire data using AES-EAX first, then split it into block size; In the receiver's side, combine the encrypted block files together, then do the decryption. In the following experiments, "eax-es" can be used to refer to this set of data;
2. For the sender, split the data into block size first, then encrypt each block file using AES-EAX; In the receiver's side, decrypt each encrypted block file, then combine the decrypted block files together. In the following experiments, "eax-se" can be used to refer to this set of data;
3. For the sender, encrypt the entire data using RSA-AES first, then split it into block size; In the receiver's side, combine the encrypted block files together, then do the decryption. In the following experiments, "rsa-es" can be used to refer to this set of data;
4. For the sender, split the data into block size first, then encrypt each block file using RSA-AES; In the receiver's side, decrypt each encrypted block file, then combine the decrypted block files together. In the following experiments, "rsa-se" can be used to refer to this set of data.





**Fig. 2.** Two Kinds of Transmitter Flow

Measurements were taken according to the transaction time consumed for each step. The measure of the performance was the transaction time of encryption and decryption for each scheme. Splitting and combining times were recorded as well. All the experiments conducted were under the same environment to ensure a fair comparison. In the RSA mechanism, the digital signature was not included in the computation as it would significantly skew the results if each part was computed for a digital signature.

The experiments were carried out on four different sizes of files, which are 1KB, 10KB, 100KB and 1MB. The block size for each experiment was the same, 140 bytes, which is the maximum SMS payload. Each experiment mode was repeated 100 times in order to guarantee the accuracy of data as possible.

## 8 Results

Through the discussion of the previous chapter, we have developed the system in Python. There are four (4) sets of data for each experiment, which is “eax-es”, “eax-se”, “rsa-es”, “rsa-se”. The format of each output result is shown in Table 3.

**Table 3.** Preliminary results format

Protocol	Action	File Name	Creation Time	File Size (bytes)	Block Size (bytes)	Part	Repeat Number	Time (milliseconds)
----------	--------	-----------	---------------	-------------------	--------------------	------	---------------	---------------------

The “Action” field can be “Splitting”, “Combination”, “Encryption” or “Decryption”. “File Name” refers to the name of the file which is going to be

**Table 4.** Samples of preliminary results

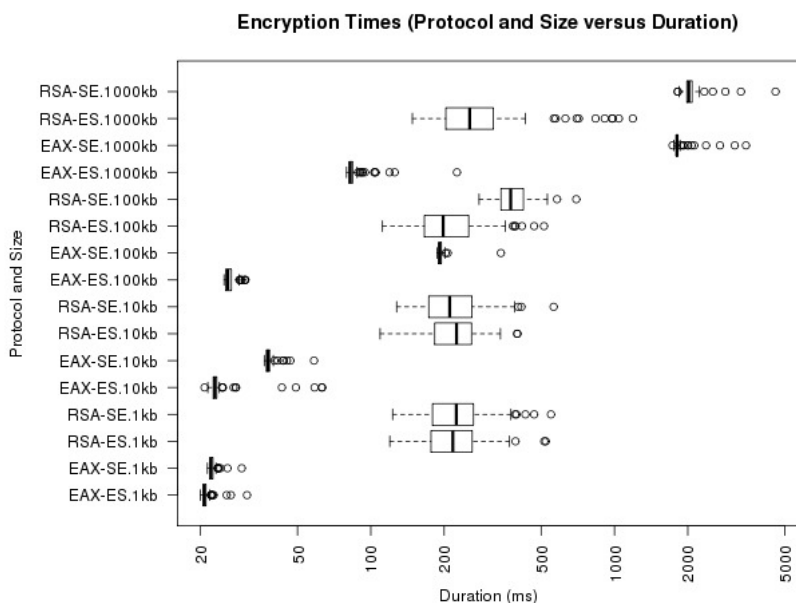
Action	File Name	Creation Time	File Size (bytes)	Block Size (bytes)	Part	Repeat Number	Time (milliseconds)
Encryption	test10kb.bak	2010/9/30 8:33	10240	140	74	8	22.87197113
Splitting	test10kb.bak	2010/9/30 8:33	10240	140	74	8	10.96510887

encrypted. “Creation Time” can be accurate to the second. Both “File Size” and “Block Size” are in bytes. The values of the “Part” represent the number of files which original document can be divided into. The “Repeat Number” refers to the current number of repetitions. “Time” is in milliseconds.

In table 4, this output refers to the 8<sup>th</sup> repetition, encrypt the “test10kb.bak” file first, then split the encrypted file into 140 bytes, which becomes 74 parts, and the encryption time is 22.87197113 ms, the splitting time is 10.96510887 ms. A box plot is used to reflect the results of the experiment.

The following is the box plot for a preliminary experimental data.

Obviously, in the 1kb file encryption, both eax-es and eax-se, were much faster than rsa-es and rsa-se. But in the decryption, their time difference was not large, this is because the original file was only 1kb, divided into 140 bytes, the number of block files was very small. Let us move forward to 10kb file. In the 10kb file encryption, eax-es and eax-se was still shown very obvious advantages, and eax-es seems more faster. In the decryption, the “se” and the “es” began to have

**Fig. 3.** Box plot for encryption

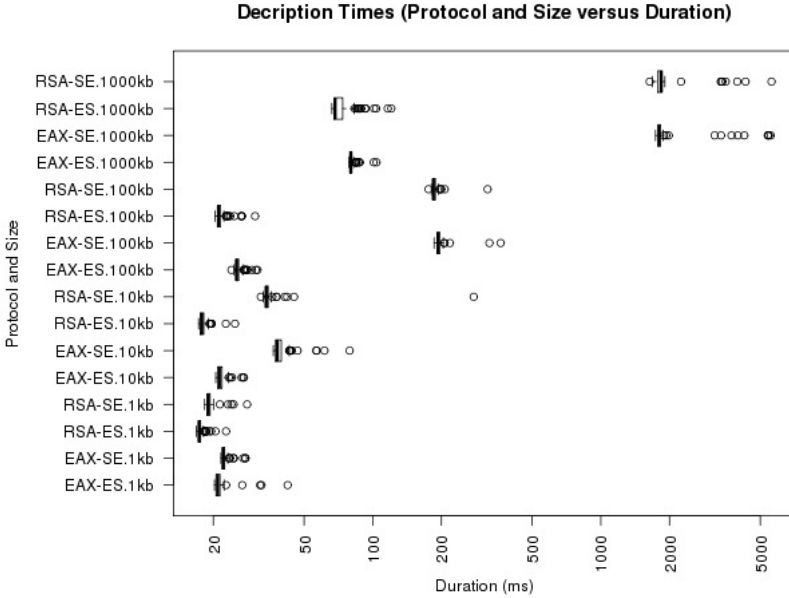


Fig. 4. Box plot for decryption

some time gaps. The advantage of “eax-es” was clearly reflected in 100kb file, both in encryption and decryption.

In the experiments of 1mb file, which was split 1MB file to 140 Bytes, it generated more than 10,000 files. In this case, the speed of EAX and RSA encryption and decryption were not the main factor; the limitation of the hardware became an issue. Obviously, encrypting the entire data then splitting into block size, was much faster than splitting the data into small pieces then encrypting every small block data. As a next step, the researchers will try to re-implement the system without using files to avoid hitting this I/O bottleneck.

## 9 Conclusion

In summary, this paper shows an comparison of RSA-AES and AES-EAX PSK cryptography mode of operation for use in SMS-based secure transmissions. Considering the security guarantees and the speed of encryption and decryption, under the same hardware and software platform, which are suitable for some M2M applications. This study pointed out the AES-EAX PSK mechanism and the RSA-AES mechanism have the same security guarantees but AES-EAX PSK mechanism performed better, and was more instantaneous. It was noted that despite the non-inclusion of the digital signatures on the RSA-AES mechanism, the AEAD mechanism still performed better. This paper also indicated that applying encryption before splitting the data is better than encryption after.

## References

1. SMS (Short Message Service), <http://www.gsmworld.com/technology/sms>
2. Information Technology-Security Techniques-Code of Practice for Information Security Management, geneva (2005)
3. SMS is top service for Asian mobile phone users (March 2006)
4. Botan cryptography library (2010), <http://botan.randombit.net/>
5. Python Programming Language (2010), <http://www.python.org/>
6. 3rd Generation Partnership Project: GSM 03.40: Digital cellular telecommunications system (Phase 2+). Technical Realization of the Short Message Service, SMS (2001)
7. Moh'd, A., Nauman Aslam, H.M.L.T.: Hardware Implementations of Secure Hashing Functions on FPGAs for WSNs. *Journal of Networking Technology* 1(1), 1–43 (2010)
8. Barkan, E., Eli, B.: Conditional Estimators: An Effective Attack on A5/1, pp. 1–19 (2005)
9. Bellare, M., Namprempre, C.: Authenticated encryption: Relations among notions and analysis of the generic composition paradigm. *Journal of Cryptology* 21(4), 469–491 (2008)
10. Bellare, M., Rogaway, P., Wagner, D.: The EAX mode of operation. In: Roy, B., Meier, W. (eds.) *FSE 2004*. LNCS, vol. 3017, pp. 389–407. Springer, Heidelberg (2004)
11. Bellare, M.N.C.: Authenticated encryption: Relations among notions and analysis of the generic composition paradigm. *Journal of Cryptology* 21(4), 469–491 (2008)
12. Biham, E., Orr, D.: Cryptanalysis of the A5/1 GSM Stream Cipher, 43–51 (2000)
13. Biham, E., Orr, D.: Cryptanalysis of the A5/1 GSM Stream Cipher. *Indocrypt* (2000)
14. Biryukov, A., Adi, S., Wagner, D.: Real Time Cryptanalysis of A5/1 on a PC. *Encryption-FSE*, 1–18 (2000)
15. Dye, M.S.: End-to-End M2M (Sample/Excerpts Copy only - Not Full Report).
16. Ekdahl, P., Thomas, J.: Another attack on A5/1. *IEEE Transactions On Information Theory* 49(1), 284–289 (2003)
17. Elad, B., Biham, E., Keller, N.: Instant Ciphertext-Only Cryptanalysis of GSM Encrypted Communication by Barkan and Biham of Technion, Full Version (2006)
18. Elgamal, T., Hickman, K.: Secure socket layer application program apparatus and method. US Patent 5, 390–657 (1997)
19. Lo, J., Binshop, J., Eloff, J.: SMSec: an end-to-end protocol for secure SMS. *Computers and Security* 27(5–6), 154–167 (2008)
20. LORD, S.: Trouble at Telco: When GSM Goes Bad. 1, 10–12 (2003)
21. Lucero, S.: Maximizing Mobile Operator Opportunities in M2M (2010)
22. Patrik, E., Johansson, T.: Another attack on A5/1. *IEEE Transactions on Information Theory* 49(1), 284–289 (2003) doi:10.1109/TIT.2002.806129
23. Rivest, R., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM* 21 (1978)
24. Schmidt, M.: Consistent M-Commerce Security on Top of GSM-based Data Protocols-A security Analysis (2001)

25. Yu, W.: New Botan, the C++ Crypto Library, built for Fedora 13 with Python Bindings Enabled and the RSA-PrivateKey fix (August 2010), <http://hip2b2.yutivo.org/2010/08/23/botan-patch/>
26. Yu, W., Tagle, P.: Development of an Over-the-Top Network Protocol for Pervasive, Secure and Reliable Data Transmission over GSM Short Messaging Service. In: To be presented at the 2010 International Conference on Computer and Software Modeling (ICCSM 2010), IACSIT (2010)

# Authentication and Authorization in Web Services

Khalil Challita<sup>1</sup>, Hikmat Farhat<sup>1</sup>, and Joseph Zalaket<sup>2</sup>

<sup>1</sup> Notre-Dame University  
Computer Science Department, Lebanon  
kchallita,hfarhat@ndu.edu.lb

<sup>2</sup> Holy Spirit University of Kaslik  
Computer Science Department, Lebanon  
josephzalaket@usek.edu.lb

**Abstract.** In this paper, we give the design of a security model that encapsulates the basic modules needed for securing the access to a web service, which are authentication and authorization. Our model relies on WS-Security standards and another application layer technology, namely the "Lightweight Directory Access Protocol". We also implement and test the model, and provide several test case scenarios. Moreover, an evaluation in terms of performance is done in order to reduce the concerns about security bottleneck and overheads. Finally, we highlight some of our model's advantages and drawbacks.

**Keywords:** Web Services, SOAP, WS-Security, WSS4J, LDAP.

## 1 Introduction

Today's companies are more intrigued to publish web services as a scheme to generate additional revenues, as they are selling their existing functionalities over the Internet and making use of the low cost communication protocol *Simple Object Access Protocol* (SOAP). As mentioned in [5], the fast increasing number of web services is transforming the web from a data oriented repository to a service oriented repository.

It is important to say that given the reality of today's open networks, it is just impossible to conduct business transactions in a networked environment without full security capabilities. Although web services are a boon to e-commerce [19], they come at a high cost of shaky security. Current methods and technology cannot fully ensure trustworthiness in loosely coupled web services. Current interests remain focused on implementing web services where the lack of security is not a bottleneck to industry's adoption of this Internet model. So far, significant progress has been made towards making the web service technology a suitable solution for areas such as e-business and e-government and e-commerce [18]. However, there are still some issues blocking the progress of the wide scale deployment of web services, one of those main issues is the security of web services.

Our aim in this paper is to provide a combined security model for web services that ensures both authentication and authorization, in order to allow

a client and a service to communicate securely, and be protected from potential problems such as unidentified client requests or unauthorized access to resources. Our model combines components that belong to different perspectives of security technologies such as *Web Service Security* (WS-Security) standards, application level protocol, and application layer processing. Our model is inspired from the one provided by Garcia and Toledo [8], where only the issues related to the security of a message, namely confidentiality and integrity, are dealt with. We complement the work done in [8] by providing a combined security model that covers both authentication and authorization based on predefined web service security standards such as WS-Security that defines standardized syntax dictating security goals that can be carried out within the messages exchanged between a client and a service. Our proposed model tends to provide a security mechanism that, once applied, will indeed achieve the issues of identity verification and access control management.

The rest of this paper is divided as follows. We introduce in Section 2 the major concepts and technologies in web services' security, in addition to widely known projects built to support web services' security. The related work is presented in Section 3. We describe in Section 4 our model that ensures both authentication and authorization. The implementation of our model is given in Section 5. Finally, a benchmark discussion is given in Section 6.

## 2 Web Services' Security

Lafon [12] defines web services as programmatic interfaces made available over the World Wide Web to enable application-to-application communications. They enable software applications to communicate and conduct transactions by adopting the *Extensible Markup Language* (XML) as a data exchange format and industry standard delivery protocols, regardless of the programming languages they are developed with and the platforms they are deployed on. Yu et al. [18] state that web services support direct interaction with other software agents using XML-based messages exchanged via Internet based protocol. Examples of web services include online reservation, stock trading, auction, etc. Web services are currently being widely adopted as a platform independent middleware. However, web services were not that interesting until a few years ago. Thanks to the major IT development the last few years, most people and companies have broadband connection and are using the web increasingly.

### 2.1 The Need of a Web Service Security Standard

Security is an important factor for deploying web services, as Yu et al. [18] mention. Web services need to be concerned with the security aspects of authentication, authorization, confidentiality, and integrity. According to Tang et al. [16], SOAP, which is a messaging protocol that allows applications to exchange information, does not provide security for messages, since it brings threats to both sender and receiver of the message. That is why the web service security

specification was developed. Nandanlin et al. [13] define web service security (WS-security) as a set of communications protocols set to address security concerns.

The WS-Security specification describes enhancements for the SOAP messaging to achieve message integrity, confidentiality, and authentication. Developed by a committee in Oasis-Open, it specifies mechanisms that can be used to accommodate a wide variety of security models and encryption technologies. It is flexible and can be used within a variety of security models like Secure Socket Layers (SSL) and Kerberos.

The web services' security challenge specified by Hondo et al. [10] is to understand and assess the risk involved based on an existing security technology, and at the same time follow the standards and understand how they will be used to offset the risk in new web services. Most of the current security discussions address identity authentication and message exchange privacy. Additional security measure at the application level could be of use, targeted at preventing authorized visitors from performing unauthorized actions. In this paper, our main concern is to address the following two fundamental security issues: authentication and authorization.

## 2.2 WS-Security Specifications

The WS-Security specification described in [13] provides mechanisms to address the three security requirements: authentication, integrity and confidentiality. With WS-Security, we can selectively employ one or more mechanism to implement a specific security requirement.

The specification given by Zhang [19] provides a mechanism to associate security tokens with message contents. It is designed to be extensible and supports multiple security token formats. The mechanisms provided by this specification allow to send security tokens (that are embedded within the message itself) in order to achieve message integrity and message confidentiality. Note that the WS-Security standard given in [13] uses the XML Encryption standard to allow encryption of any portion of the SOAP message.

## 2.3 Some WS-Security Projects

We next present several web services' security projects on which we rely to implement our model in Section 5. Then we describe the LDAP protocol and the Active Directory, which constitute a main component of our model, because they are at the basis of the authentication process we aim to achieve.

**WSS4J.** Apache WSS4J is an implementation of the WS-security [2]. It is an open source java library that can be used to sign and verify SOAP messages. It also provides resources for building interoperable, trusted web services using the WS-security standard. The libraries can be deployed to provide protocol support for both client and server applications.

We are interested in using WSS4J as a library within our model because it is an open source project, and because it is interoperable with "Java API for XML based Remote Procedure Calls" and .NET server/clients.



**SOAP.** Box et al. [5] define SOAP as a simple XML-based protocol to let applications exchange information over an Application Layer protocols like SMTP, HTTP, or HTTPS. The SOAP specification is currently maintained by the XML Protocol Working Group of the World Wide Web Consortium. SOAP can encapsulate WS-security information. As described by O'Neill [14], security data in SOAP include security tokens to indicate the identity of the sender, and a digital signature to ensure that the SOAP message has not been modified since its original signing. SOAP is used to send data from one application to another. It is supported by all Internet browsers and servers, and at the same time allows applications to communicate when running on different operating systems with different technologies and programming languages.

**Axis.** The Apache organization [1] created Axis2 to be a core engine for web services. Axis2 not only provides the capability to add web services interfaces to web applications, but can also function as a stand-alone server application.

We used Axis2 to create an application based on our model (given in Section 5), and where both the client and the server were developed as stand-alone applications, without having to create a separate web application to act as a service.

**Rampart.** Apache Rampart [3] is a project that implements the WS-Security standard for the Axis2 web services engine created by the Apache Software Foundation. It provides some security features to web services.

**Lightweight Directory Access Protocol.** According to Koutsonikola et al. [11], Lightweight Directory Access Protocol (LDAP) is an application layer protocol for querying and modifying directory services running over TCP/IP. We use LDAP to authenticate users requesting a service against a Service provider's active directory in order to restrict accesses to known and specified users only. Moreover, we chose LDAP because it is widely supported, very general and includes basic security, and it can support many types of applications.

**Active Directory.** Active Directory (AD) [6] is a technology created by Microsoft providing multiple network services such as LDAP directory services, Kerberos based authentication and DNS base naming and network information. It is a type of databases that can be searched to provide useful network information. We chose Active directory in conjunction with LDAP because it allows users and applications to make use of published information over a network without requiring any knowledge about this network. Moreover, an AD is an optimized database for querying which makes information retrieval easier and faster.

We next give an overview of some researchers' work done in the field of web services security, and summarize their main contributions to the field.

### 3 Review of Related Work

Yamaguchi et al. [17] proposed an application programming model called "Web Service Security Application Programming Interfaces" to simplify the programming for end users who are not very familiar with WS-Security. Their model was based on the Service Oriented Architecture (SOA), the WS-security requirements, and on the existing APIs proposed by Microsoft. It consisted of six APIs that tend to achieve confidentiality and integrity through signatures and encryption.

Bhargavan et al. [4] addressed the problem of securing sequences of SOAP messages exchanged between web services providers and clients. Their work confirmed the inefficiency of using WS-Security alone for each message and that the integrity of a whole session as well as each message should be secured. They relied on WS-Secure conversation, which goal is to secure sessions between two parties, and on the WS-Trust, which describes how security contexts are obtained.

Gutierrez et al. [9] intended to describe a security model for web services in order to facilitate the development phase. Their model is based on web service security requirement specification, web service security architecture design and web service security standard selection. The research focused mainly on the web services security architecture.

Rahaman et al. [15] describe web services security architectures in a simplified way using WS standards, in addition to addressing the issue of attacking a SOAP message from XML rewriting attack. The research focused on message level security and discussed two different message flows that use (or do not use) SOAP message structure information.

Felix et al. [7] addressed the scalability and flexibility limitations of the WS authentication model where the acquirement of identity claims requires online interactions with security token services, thus introducing communication overhead and creating performance bottlenecks. They presented a new model where they addressed these limitations through two concepts: credentials for claim inference and claim-based issuer references. They showed how credentials are used both to increase the scalability and to reduce the number of online token requests. They also showed how the simultaneous usage of security tokens and credentials results in several advantages when compared to credentials used in trust management models.

Zhang [19] enumerated the main challenges that threaten a web service and make it "untrustworthy". He proposed a solution to address these challenges by adding an additional layer (i.e. WS-Trustworthy layer) on top of the WS-Security layer.

Garcia and Toledo [8] proposed a security model for web services that is based on semantic security policies. Their main goal was to ensure confidentiality and integrity of a SOAP message. The main components of the security model they designed are equivalent to some XML elements of the WS-Security, XML encryption and XML Signature standards. Note that they only addressed the issues of confidentiality and integrity.

To our knowledge, no single model addressed both authentication and authorization at the same time. We next give the description of our model that addresses these two security issues.

## 4 Authentication and Authorization-Based Model

Our model, given in Figure 1, is inspired by Garcia and Toledo's one [8], and can be considered as a complement to their model since they focused only on the security and integrity of a message, regardless of the identity of the message issuer and her access rights. Our main goal is given in the main module, namely "Securing Web Services", which in turn is composed of two submodules: Authentication and Authorization.

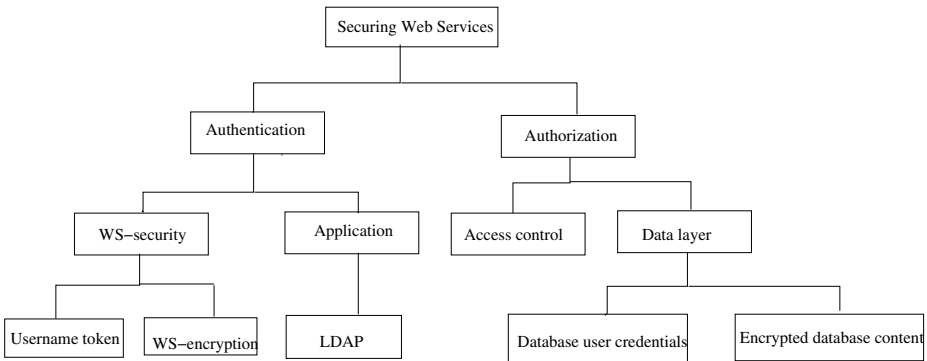


Fig. 1. Authentication and Authorization Diagram

### 4.1 Authentication

Authentication is achieved through the use of WS-Security standard and, more specifically, the "UsernameToken" that enforces the use of a username and a password. Note that both the username and the password will be protected inside the SOAP message through encryption. In addition, an application-level protocol (i.e. LDAP) is integrated in the authentication mechanism in order to ensure that access is only allowed to known and identified users.

### 4.2 Authorization

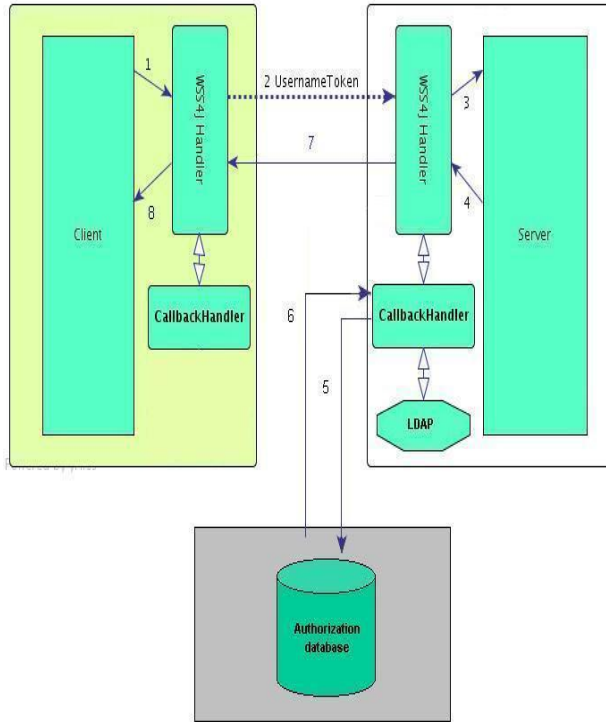
Authorization is achieved through two layers:

1. The "Access Control layer", whose purpose is to identify a requestor and to check her access rights;
2. The "Data Layer", which is related to a Database system containing the logical mapping of users to the requested services for access control purposes. The data layer is composed of two submodules: the "Database user credentials" that are used to retrieve the required information from the Database; and the "Encrypted database content" that stores encrypted data.

## 5 Implementation of Our Model

In this section we describe the application that we built on top of our proposed model in order to test its feasibility, efficiency and performance.

The diagram in Figure 2 shows the engine structure of the created application.



**Fig. 2.** Application Engine Diagram

The application engine we created is divided into two main components: the client and the server. The client is the part that requests a service providing the service the necessary credentials, the server is the service provider, its goal is to authenticate the requestor and validate its access rights before sending back a response. Our engine is based on WSS4J, which is an Apache project implementing WS-Security specification. It also uses Rampart engine, which is another Apache project based on WSS4J to secure SOAP messages.

We next explain the steps shown in Figure 2.

1. The client issues a service request. This request is intercepted by the WSS4J handler, which then invokes the client CallbackHandler class where the user password is provided.
2. The client's request now holds the UsernameToken with the corresponding credentials and passes it on to the server.

3. At server side, the WSS4J handler intercepts the message containing the token.
4. The server then invokes its corresponding Callbackhandler class where the credentials are retrieved in order to be processed.
5. The credentials are passed through LDAP to the corresponding server's Active Directory where they are validated for authentication.
6. In case of successful authentication, the user information is sent to the database for authorization. The role of the user is fetched as well as its corresponding services. Once the allowed services are retrieved they are compared to the incoming service name call and the validation result will be communicated in the next step back to the client.
7. This is the result (of success or failure) that is communicated back to the client through its CallbackHandler.
8. The response is sent to the client application either in an exception mode when the authorization failed, or as a service result in case the authorization succeeded.

## 6 Benchmark

The aim here is to examine the performance level of our model. The next scenarios were performed using the benchmark tool Mercury LoadRunner 8.1.

### 6.1 Scenarios

#### Scenario 1: Web Services with UsernameToken and LDAP

In this scenario our application uses a UsernameToken and performs an LDAP search in the active directory. The results collected reflect how the combination of both WS-security and application level protocol affect the overall performance and response time of the system. The results in Figure 3 show the response time in milliseconds by number of simultaneous users access.

Number of users	Average transaction response time
10	235.46
25	760.78
50	1914.75
100	3853.74
150	5238.35
200	9351.27

**Fig. 3.** Application Engine Diagram

As we notice, the average response time per user when having a full load of 200 simultaneous users is almost 46.7 ms per user.

## Scenario 2: Web Services with SSL

For SSL, and as we can see from Figure 4, the average response time when having a full load of 200 simultaneous users is around 21.1 ms per user. Even though the results show a clear difference in response time when using SSL against WS-Security and LDAP, however, there are many reasons that weaken the value of SSL as we explain in the next subsection.

Number of users	Average transaction response time
10	98.89
25	323.33
50	825.25
100	1676.38
150	2310.11
200	4226.68

Fig. 4. Web services with SSL benchmark

## 6.2 Advantages and Limitations

The main advantage of our authentication process is that it is a direct authentication process, which means that the client and the web service trust each other to securely handle the authentication credentials.

The WS-Security standard and the use of LDAP play an important role in our model because they have some advantages over other authentication techniques. For example, the main limitations of SSL are:

1. SSL is only good for POINT-TO-POINT communication.
2. Authentication becomes difficult when using SSL.
3. SSL is bound to HTTP protocol.
4. Encrypting or signing part of a message is not possible with SSL.

LDAP is widely supported, very general and includes basic security, and it can support many types of applications. It also allows the reuse of domain credentials so one does not necessarily have to create a separate database to list the users who are entitled to call services.

As for the authorization process, the main advantage is that it is an application level process and thus it could be easily customized according to each service provider. It is a scalable module that offers easy integration of new rules and new entities that make it ready for any change of service policy.

## 7 Conclusion

In this research we addressed and proposed a solution to the problem of authentication and authorization in web services security. For that purpose we created a model that combines security technologies from multiple ends. First, we used the WS-Security standards, mainly the security token "UsernameToken" in the authentication process; next we combined this standard with the use of an

application level protocol, which is the LDAP for credentials validation. Then in the authorization process, we relied on an application level manipulation, where we created a logical and physical model to underline the mappings between users and their corresponding business roles, as well as between the roles with their related services. These mappings are at the basis of the authorization control management since they decide the access rights for a user requesting a service. Moreover we implemented our model and conducted different test case scenarios to put under examination its efficiency. We also conducted benchmark scenarios to evaluate our model's performance in terms of response time and compared it to other security techniques.

Even though our model deals with issues of authentication and authorization, it needs to be extended in order to give a wider coverage of security, namely confidentiality and integrity.

Fortunately, our model is flexible enough to encapsulate new security measures whether at the level of WS-security (since we can add new tokens or even use different standards like WS-policy or WS-Trust), or at the application layer by integrating new entities to the data model.

## References

1. Apache axis2/java next generation web services (2008), <http://ws.apache.org/axis2/>
2. The apache software foundation (2008), <http://ws.apache.org/wss4j/>
3. Rampart: Ws-security module foraxis2 (2008), [http://ws.apache.org/axis2/modules/rampart/1\\_3/security-module.html](http://ws.apache.org/axis2/modules/rampart/1_3/security-module.html)
4. Bhargavan, Corin, Fournet, Gordon: Secure sessions for web services. *ACM Transactions on Information and System Security* (2007)
5. Box, Enhebuske, Kakivaya: Simple object access protocol (2000), <http://www.w3.org/TR/2000/NOTE-SOAP-20000508/>
6. Dias, J.: A guide to microsoft active directory (ad) design. Department of Energy Lawrence Livermore National Laboratory (2002)
7. Felix, Pedro, Ribeiro: A scalable and flexible web services authentication model. In: *Proceedings of the 2007 ACM workshop on Secure web services*, pp. 62–72 (2007)
8. Garcia, D., de Toledo, F.: Web service security management using semantic web techniques. In: *CM Symposium On Applied Computing*, pp. 2256–2260 (2008)
9. Gutierrez, Fernandez-Medina, Piattini: Web services enterprise security architecture: a case study. In: *Proceedings of the 2005 Workshop On Secure Web Services*, pp. 10–19 (2005)
10. Hondo, Nagaratnam, Nadalin: New developments in web services and e-commerce, securing web services. *IBM Systems Journal* 41 (2002)
11. Koutsonikola, V., Vakali, A.: Ldap: Framework, practices, and trends. *IEEE Internet Computing* 8(5), 66–72 (2004)
12. Lafon, Y.: Web services activity statement (2008), <http://www.w3.org/2002/ws/Activity>
13. Nadalin, A.: Web services security: Soap message security 1.1 (2006), <http://docs.oasis-open.org/wss/v1.1>
14. O'Neill, M.: *Web services security*. McGraw-Hill, New York (2005)

15. Rahaman, Schaad, Rits: Towards secure soap message exchange in a soa. In: Proceedings of the 3rd ACM Workshop On Secure Web Services, pp. 77–84 (2006)
16. Tang, Chen, Levy, Zic, Yan: A performance evaluation of web services security. IEEE Computer Society, 67–74 (2006)
17. Yamaguchi, Chung, Teraguchi, Uramoto: Easy-to-use programming model for web services security. In: Proceedings of the The 2nd IEEE Asia-Pacific Service Computing Conference, pp. 275–282 (2007)
18. Yu, Liu, Bouguettaya, Medjahed: Deploying and managing web services: issues, solutions, and directions. In: VLDB, pp. 537–572 (2008)
19. Zhang, J.: Trustworthy web services: actions for now. In: IT professional, pp. 32–36 (2005)



# Integrating Access Control Mechanism with EXEL Labeling Scheme for XML Document Updating

Meghdad Mirabi, Hamidah Ibrahim, Ali Mamat, and Nur Izura Udzir

Department of Computer Science, Faculty of Computer Science and Information Technology,  
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia  
meghdad.mirabi@gmail.com, {hamidah, ali, izura}@fsktm.upm.edu.my

**Abstract.** Recently, an urgent need for XML access control mechanism over World Wide Web has been felt. Moreover, an efficient dynamic labeling scheme is required in order to eliminate the re-labeling process of existing XML nodes during XML document updating. However, the previous research on access control mechanisms for XML documents has not addressed the issue of integrating access control with a dynamic labeling scheme. In this paper, we propose an XML access control mechanism integrated with EXEL encoding and labeling scheme to eliminate the re-labeling process for updating the well-formed XML documents. The key idea is to regard an authorization as a query condition to be satisfied. Therefore, the benefit of speeding up searching and querying processes is obtained by employing such a labeling scheme in our proposed access control mechanism.

**Keywords:** Labeling Scheme, Node Filtering, Tree-Awareness Metadata, XML Updating.

## 1 Introduction

Recently, the eXtensible Markup Language (XML) [1] as a de facto standard for sharing and exchanging information over Internet and Intranet has been suggested. Thus, the need of managing XML documents over World Wide Web arises. Considering some operations such as querying and updating XML document, these kinds of operations should be quick to be carried out and more importantly safe from unauthorized access.

In order to quickly retrieve some parts of the XML document, several XML query languages such as XPath [2] and XQuery [3] have been proposed. In general, path expressions are used in these query languages for searching structural relationships between the XML nodes. These structural relationships should be evaluated efficiently. In order to determine the structural relationships between the XML nodes, they are labeled in such a way that the structural relationship between two arbitrary XML nodes can be computed efficiently. Many labeling schemes such as the region numbering scheme [4, 5] and the prefix labeling scheme [6] have been proposed. However, these labeling schemes have high update cost; they cannot completely eliminate re-labeling process for the existing XML nodes when the XML document is

updated. For instance, insertion of an XML node may change the XML tree structure, and the labels of XML nodes may need to be changed. Thus, many works have been studied to offer an efficient labeling scheme for the process of XML document updating such as [7-9].

Also, several researches have implemented the XML access control mechanisms which include studies by [10-27]. In general, XML access control mechanisms are classified into two groups: node filtering [10-15, 25] and query rewriting mechanisms [16-24, 26, 27]. In node filtering mechanism, access authorizations are determined by labeling the tree nodes with a permission (+), or a denial (-) and then pruning the tree based on associated signs. In query rewriting technique, access authorizations are employed to rewrite probable unsafe user queries into safe ones which should be evaluated against the original XML dataset. A safe query is a query which its result does not violate any access authorizations.

In this paper, we propose an XML access control mechanism tightly integrated with EXEL (Efficient XML Encoding and Labeling) [7-9] encoding and labeling scheme for XML document. The key idea is to regard an authorization as a query condition to be satisfied. Therefore, the benefit of speeding up searching and querying processes is obtained by employing such a labeling scheme in our proposed access control mechanism.

The rest of the paper is organized as follows: in Section 2, existing XML access control mechanisms are investigated. In Section 3, the access control model employed by our proposed mechanism is presented. Our proposed access control mechanism is presented in Section 4. Finally, the paper is concluded in Section 5.

## 2 Related Works

The first process in the traditional node filtering mechanism is to parse the XML document and generate its DOM tree then label the DOM tree based on access authorizations defined by a security administrator and finally prune unnecessary parts of the XML tree according to its labeling and show the result to the user [11, 12]. Due to traversing the DOM tree, access control mechanism proposed by [11, 12] is not scalable. If the DOM tree is very large it needs to have large memory space. Besides, to answer a user query, the whole DOM tree should be traversed which need long time to process. In order to resolve the problem, a fine grain access control mechanism which stores the XML documents as tables in relational database is presented in [25]. It is scalable with efficient response time and storage cost compared to [11, 12].

Another possible solution to overcome the shortages of the traditional node filtering mechanism is to separate the DOM and the SAX (Simple API for XML) when parsing the XML documents [13, 14]. If a user request has permission to read the XML document, it is processed by the SAX otherwise by the DOM.

In order to determine accessibility of XML elements, top-down and bottom-up strategies [10] traverse the paths between the root of XML tree and the element. In top-down strategy, authorization checking starts from the authorizations specified at the root while in bottom-up strategy, authorization checking starts from the ones specified at the most specific granularity level and going up through XML tree to find

appropriate authorization. In the worst case, both strategies require traversing the whole XML tree which can be time consuming.

The main idea of static analysis proposed by [16], as a pre-processing access control mechanism, is to make automata for XML queries, XML access control policies, and XML schemas and then compare them. As a result, it does not examine real XML documents and runtime checking is not needed. Runtime checking is only required when real XML documents are required to check. The static analysis is not intended to entirely eliminate runtime checking, but rather intended to complement it. When static analysis cannot provide determinate answer, runtime checking is needed. This method classifies an XML query at compile time into three categories: entirely authorized, entirely prohibited, or partially authorized. Entirely authorized or entirely prohibited queries can be executed without access control. However, the static analysis cannot obtain any benefits when a query is classified as a partially authorized one. QFilter [22] as an external pre-processing XML access control system checks XPath queries against access control policies and rewrite queries according to access control policies before passing the revised queries to XML engine to process. Static analysis method [16] needs a runtime checking to filter out the unauthorized data while QFilter [22] solves this problem by rewriting XPath queries to filter out unauthorized part of input queries before passing them to XML query engine. Thus, QFilter has much better performance than static analysis method. However, if there are many access control policies for each role, NFA (Non deterministic Finite Automata) based approach in QFilter may have unacceptable overhead. Moreover, QFilter rewrites some kind of queries incorrectly according to examples derived in [24]. On the contrary, a DFA (Deterministic Finite Automata) based enforcement mechanism is devised in [18, 21] which decreases the complexity of query rewriting and always check whether the user has the right to consult the nodes that occur within the predicates.

A view based access control mechanism is proposed by [17] which generates not only a view called security view, but also a DTD view in which the security view conforms. The DTD view is generated to improve the efficiency of query rewriting and optimization. In contrast to [17], [19, 20, 26, 27] consider general XML DTDs defined in terms of regulations rather than normalized DTDs. Furthermore, [19, 20, 26, 27] do not permit dummy element types in the definition of security views.

The XML access control mechanism proposed by [23, 24] is based on access control abstraction and query rewriting. Access control abstraction is an efficient mechanism to check only the necessary access control rules based on user query instead of checking all of the access control rules. Also, user queries are rewritten by extending or eliminating XML tree nodes of DTD and operators such as union, intersection, and except are supported to transfer user queries into safe and correct queries which maintain the user's access control policies.

An efficient XML access control mechanism which integrates with query processing using DP (Dynamic Predicate) is devised in [15]. Accessibility of elements is checked during query execution using the DP. The key idea for integrating access control with XML query processing is to discover a set of elements which have the same accessibility. To effectively search the authorization, the mechanism proposed by [15] uses authorization index and nearest ancestor search technique.

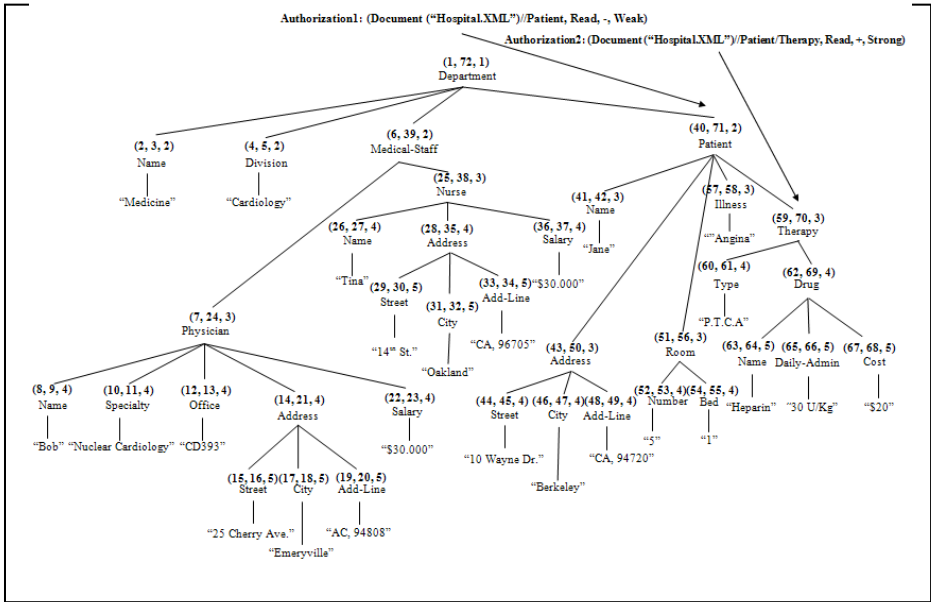
### 3 XML Access Control Model

An access control policy includes a set of access authorizations. In general, an authorization can be defined as 5-tuple <subject, object, action, permission, propagation> where subject is the user or role concerned by the authorization. In this study, we assume that the subject is fixed; therefore the authorization is formed as 4-tuple <object, action, permission, propagation>. Object is presented by XPath expression [2] which contains element(s) of the XML document. Action is an executable action which can be Read, Insert, Delete, Update, and Rename. Permission represents the acceptance (+) or denial (-) of rights. Therefore, we refer to the access authorizations that grant access to an object as positive and those that deny access as negative. Details of the executable actions in the model are described below:

- If a user holds a Read privilege on a node  $u$ , s/he is permitted to read the content of node  $u$  and its descendants.
- If a user holds an Insert privilege on a node  $u$ , s/he is permitted to insert a new node as a sibling, child, or parent node of the node  $u$ .
- If a user holds a Delete privilege on a node  $u$ , s/he is permitted to delete the node  $u$  and its sub-trees.
- If a user holds an Update privilege on a node  $u$ , s/he is permitted to update the content of node  $u$ .
- If a user holds a Rename privilege on a node  $u$ , s/he is permitted to rename the node  $u$ .

An authorization can be explicit or implicit with the aim of limiting the number of authorizations which must be defined. An explicit authorization is explicitly defined on an XML element while an implicit authorization is implied by an explicit authorization which is defined on the nearest ancestor of XML element. Also, authorizations can be strong or weak with the purpose of providing more flexibility to the model. A strong authorization does not permit an implicit authorization to be overridden while a weak authorization permits an explicit authorization overrides an implicit authorization. The propagation policy in the model is *most specific override takes precedence*. It means an explicit authorization on an element overrides any weak authorizations defined on the ancestors of the element. Also, the *closed* policy is employed when there is not any authorization for an element. It means if there is not any authorization for an element or its ancestors explicitly or implicitly, the element is inaccessible. Besides, *denials take precedence* policy is as the conflict resolution policy. It means if both positive explicit authorization and negative explicit authorization for the same action are defined, negative authorization overrides positive ones.

An example of XML document and authorization for Read action is illustrated in Fig. 1. Authorization1 does not permit the user to read the element "Patient" and its descendants. However, the Authorization2 overrides Authorization1 and permits the user to read the element "Therapy" and its descendants.



**Fig. 1.** An example of XML document and authorizations

According to the authorization actions supported by the model, the query operations are as follows:

- Read (target)
- InsertChild (source, target)
- InsertBefore | After (source, target)
- InsertParent (source, target)
- Delete (target)
- Update (source, target)
- Rename (source, target)

Read is a read operation, in which target can be an element or an attribute. InsertChild is an insert operation, in which source can be a PCDATA, an element or an attribute. InsertChild inserts source as the child of element denoted by target. If the XML document contains a sequence of information, InsertBefore and InsertAfter are employed in the user query. InsertBefore inserts source before element denoted by target, and InsertAfter does after element denoted by target. In addition, InsertParent insert source as the parent of element denoted by target. Delete is a delete operation, in which target can be PCDATA, an element or an attribute. Update is an update operation, in which target can be an element or an attribute, and source can be a PCDATA. Rename is a rename operation, in which target can be an element or an attribute, and source is a new name.

## 4 XML Access Control Mechanism

In our proposed XML access control mechanism, relational database is employed with the purpose of storing the tree-awareness metadata of the XML document. Tree-awareness metadata contains information related to EXEL encoding and labeling scheme of the XML document tree which is required to eliminate the re-labeling process for existing XML nodes. Besides, access authorizations are stored in RDBMS with the aim of accelerating the process of access control over the XML document. Instead of traversing the whole XML tree to find the proper authorization, our proposed mechanism is capable to select only the necessary authorizations for processing a user query.

In the following, first EXEL encoding and labeling scheme proposed by [7, 8] is described and then our proposed access control mechanism integrated with the EXEL labeling scheme is explained.

### 4.1 EXEL Labeling and Encoding Scheme

EXEL [7, 8] encoding and labeling scheme is capable to remove the need of re-labeling as well as to compute the structural relationship between XML nodes effectively. Bit string is employed in the EXEL to encode the XML nodes. This bit string is ordinal as well as insert friendly. The definition of Lexicographical order ( $<$ ) of bit string is defined as follows:

#### Lexicographical Order ( $<$ ):

1. 0 is smaller than 1 ( $0 < 1$ ) lexicographically.
2. Bit string  $a$  is equal to bit string  $b$  lexicographically, if  $a$  and  $b$  are the same ( $a = b$ ).
3. For bit strings  $a_1, a_2, b_1,$  and  $b_2, a_1b_1 < a_2b_2$ , iff ( $a_1 < a_2$ ) or ( $a_1 = a_2$  and  $b_1 < b_2$ ) or ( $a_1 = a_2$  and  $b_1$  is null (empty string)), where  $\text{length}(a_1) = \text{length}(a_2)$ .

According to the above definition, for each bit string  $s$  which ends with '0', the largest bit string among bit strings which are smaller than  $s$  lexicographically is the  $s$ 's longest prefix  $p$  (i.e.,  $s = p0$ ). However, we cannot generate any bit string which is greater than the prefix  $p$  and smaller than  $s$ . For example, there is not any bit string which can be inserted between '1110' and its longest prefix bit string '111'. Thus, if the last bit of any two consecutive bit strings is '1', we can insert a new one between the bit strings without any changes on them. The key idea to remove re-labeling during updating process of a node in the XML tree is *property 1*.

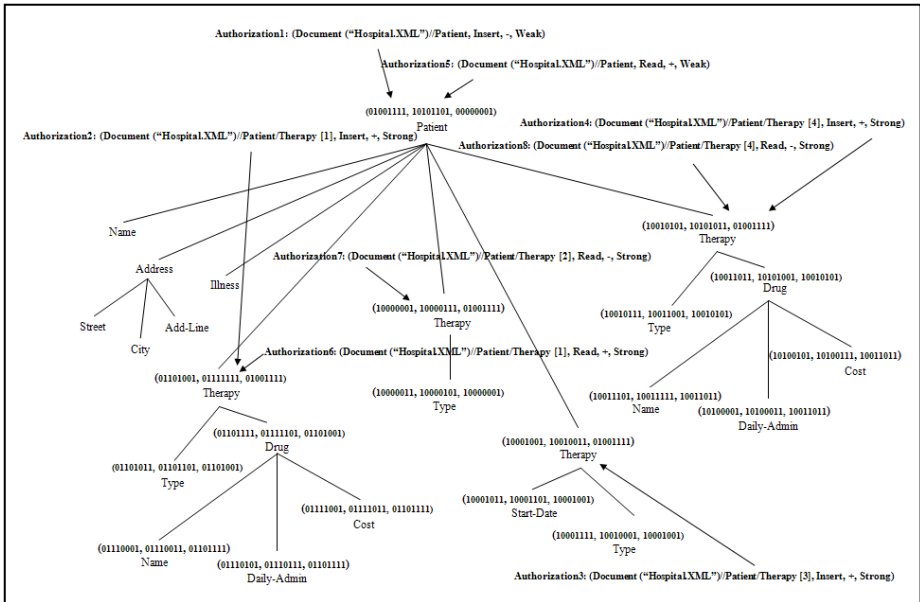
**Property 1.** For two bit strings  $a1$  and  $b1$ , if  $a1 < b1$  lexicographically, then  $a < b$  lexicographically.

The algorithm of generating the bit string for nodes is illustrated in Fig. 2 which is the enhanced binary encoding algorithm in [7, 8]. This algorithm obeys the *property 1*.

Let  $N$  be the total number of nodes of XML tree,  
 For first bit string  $b(1)$ ,  $b(1) = (0^{\lceil \log_2^2 N \rceil})1$ .  
 For  $(i+1)$ th bit string  $b(i+1)$ ,  $b(i+1) = b(i) + 10$ .

**Fig. 2.** Algorithm of bit string generation

EXEL uses bit string generation algorithm with region labeling approach to solve the problem of re-labeling the nodes in the updating process. In region labeling approach [4, 5], each XML node of the XML tree is assigned to a region which contains a pair of start and end values which are determined by the position of the start tag and end tag of the node respectively. Additionally, to identify the Parent-Child (P-C) relationship between nodes efficiently, the level of nodes is added to this approach. An example of the XML tree labeled with  $\langle \text{start, end, level} \rangle$  is illustrated in Fig. 1. In order to determine the P-C relationship between nodes, the region numbering approach uses the level information but during the update process of XML data, the level information is sensitive information. For example, when a new node is inserted as an ancestor, the level information of a lot of nodes should be changed. To solve this problem, EXEL uses the parent information instead of the level information. An example of XML tree encoded and labeled by EXEL is shown in Fig. 3. For more information about EXEL labeling scheme, refer to [7-9].



**Fig. 3.** An example of XML document and authorizations

## 4.2 The Proposed Access Control Mechanism

In order to explain our proposed mechanism which is tightly integrated with EXEL encoding and labeling scheme, we consider a part of well-formed XML document with access authorizations in Fig. 3.

Each element of the XML document illustrated in Fig. 3 is labeled with <Start, End, ParentStart> based on EXEL encoding and labeling scheme. Such information is stored in EXEL-METADATA relation. The schema of EXEL-METADATA is as follows:

### EXEL-METADATA (NodeLabel, Start, End, ParentStart)

An instance of EXEL-METADATA relation of the XML document illustrated in Fig. 3 is shown in Table 1. ParentStart attribute is the parent's start value of a node.

**Table 1.** An instance of EXEL-METADATA relation

NodeLabel	Start	End	ParentStart
Patient	01001111	10101101	00000001
...	...	...	...
Therapy	01101001	01111111	01001111
...	...	...	...
Therapy	10000001	10000111	01001111
...	...	...	...
Therapy	10001001	10010011	01001111
...	...	...	...
Therapy	10010101	10101011	01001111
...	...	...	...

Also, authorizations specified by a security administrator for the XML document are stored in AUTHORIZATION relation. The schema of this relation is as follows:

### AUTHORIZATION (ID, Object, Start, End, Action, Permission, Type)

An instance of AUTHORIZATION relation for the XML document illustrated in Fig. 3 is shown in Table 2.

**Table 2.** An instance of AUTHORIZATION relation

ID	Object	Start	End	Action	Permission	Type
1	//Patient	01001111	10101101	Insert	-	Weak
2	//Patient/Therapy[1]	01101001	01111111	Insert	+	Strong
3	//Patient/Therapy[3]	10001001	10010011	Insert	+	Strong
4	//Patient/Therapy[4]	10010101	10101011	Insert	+	Strong
5	//Patient	01001111	10101101	Read	+	Weak
6	//Patient/Therapy[1]	01101001	01111111	Read	+	Strong
7	//Patient/Therapy[2]	10000001	10000111	Read	-	Strong
8	//Patient/Therapy[4]	10010101	10101011	Read	-	Strong



Given a well-formed XML document, a set of access authorizations and a user query, our proposed access control mechanism checks authorizations according to the access authorizations defined by a security administrator for the user query and execute the user query action if the user is authorized to carry out. Therefore, our proposed XML access control mechanism contains the following steps:

1. Extract the target node of the user query.
2. Retrieve EXEL metadata of the target node. The retrieved information contains EXEL metadata for a set of candidate nodes.
3. Find the nearest positive ancestor authorization for each candidate node.
  - 3.1. If the candidate node has such an authorization, the candidate node with EXEL metadata will be forwarded to step 4.
  - 3.2. Else the query will be rejected.
4. Execute the user query action for all candidate nodes forwarded from the step 3 and update metadata stored in relational database based on the user query action.

As mentioned in Section 3, authorization objects deal with XPath expression [2] which contains element(s) of the XML document. In order to extract the target node of a user query, it is desirable to define the target node.

**Definition of Target Node:** the last node of each XPath expression in a user query is a target node. For instance, the target node of “//Department/Patient” is the “Patient” node.

According to the above definition, the proposed mechanism is able to extract the target node of a user query in the first step. The following SQL query is constructed in the second step with the purpose of retrieving EXEL metadata of the target node. The result of the SQL query contains a set of the EXEL metadata of candidate nodes.

```
SELECT *
FROM EXEL-METADATA
WHERE NodeLabel = Target-Node
```

Due to propagation policy *most specific override takes precedence* employed in our model, the accessibility of an element can be determined by finding authorizations specified on the nearest ancestor of the element. We borrow the definition of nearest ancestor authorization from [15] as defined below.

**Definition of Nearest Ancestor Authorization:** An authorization is called the nearest ancestor authorization  $auth_{naa}$  of element  $e$  if it satisfies the following two conditions:

1.  $auth_{naa}$  is an explicit authorization granted on the element  $e$  or one of its ancestor elements regardless of its authorization action;
2. No explicit authorization exists on element in the depth between the element  $e$  and the element on which  $auth_{naa}$  is granted.

Note: if a strong authorization satisfies the first condition, it automatically satisfies the second condition by the definition of the strong authorization.

Due to supporting different actions in our proposed mechanism, such a mechanism must be able to find the nearest ancestor authorization which its action is the same as the user query action. The algorithm of finding the nearest ancestor authorizations is shown in Fig. 4. Instead of traversing all nodes to find the proper authorization, the process of determining the accessibility of an element is accelerated with this algorithm in the step 3 of our proposed mechanism.

```

QueryAction ← user query action;
For each candidate node n do {
  // AA is a set of Ancestor Authorizations;
  AA ← (Select * from AUTHORIZATION
        Where (Start <= (Start value of n))
              AND
              (End >= (End value of n))
              AND
              (Action = QueryAction));
  MinLength ← abs ((Start value of n) – (Start value of AA[0]));
  NAA ← AA[0]; //NAA is Nearest Ancestor Authorization;
  For (i=1; i<length(AA); i++){
    temp ← abs((Start value of n) – (Start value of AA[i]));
    If (temp < MinLength) then
      MinLength ← temp;
      NAA ← AA[i];
    else if (temp == MinLength) then
      NAA ← negative authorization between AA[i] and NAA
            is selected; //based on conflict resolution policy;
  }}

```

**Fig. 4.** The algorithm of finding nearest ancestor authorization

Due to positive and negative authorizations, the candidate nodes which the nearest ancestor authorizations have positive permission are passed to the forth step. Therefore, in the forth step, our proposed mechanism executes the user query action for all candidate nodes passed from the third step using the algorithm illustrated in Fig. 5.

The reason of using EXEL as an encoding and labeling scheme in this study is its behavior in XML updating process. Re-labeling the existing nodes of the XML document after some operations such as renaming the node tag, updating the value of leaf node, or deleting a leaf node or whole sub-tree is not needed. The problem of updating the XML document is in insertion operations. In XML document tree, three types of insertion can occur depending on the position of node to be inserted: insertion of a node as a child of a leaf node, insertion of a node as a sibling node, and insertion of a node as a parent node.

Before inserting a node, an algorithm is needed to generate new bit string between two preexisting bit strings. We use the MakeNewBitString algorithm proposed by [7, 8] to generate new bit string. Also, InsertSiblingAfter algorithm proposed by [7, 8] is

used to insert a node as a sibling node after the node denoted by *cur*. The behavior of inserting a new sibling node before a node is similar to that of inserting a node after. In addition, InsertChildOf algorithm proposed by [7, 8] is used to insert a node as the child of node denoted by *cur* and InsertParentOf algorithm proposed by [7, 8] is used to insert a node as the parent of node denoted by *cur*. Refer to [7-9] for more information.

```

QueryAction ← user query action;
Switch (QueryAction) {
  Case "Read":
    Show the candidate nodes and its descendants;
    Break;
  Case "Update":
    Update content of the candidate nodes with source parameter in the user query;
    //re-labeling the existing nodes is not needed;
    Break;
  Case "Delete":
    Delete the candidate nodes with their sub-trees;
    //re-labeling the existing nodes is not needed;
    Delete EXEL metadata of the candidate nodes and their descendants from
    EXEL-METADATA relation;
    Break;
  Case "Rename":
    Rename tag of the candidate nodes with source parameter of the user query;
    //re-labeling the existing nodes is not needed;
    Break;
  Case "InsertBefore":
    For each candidate node n do
      Insert source parameter as a sibling node before node n using
      InsertSiblingBefore Algorithm;
      Insert EXEL metadata of inserted node into EXEL-METADATA relation;
    Break;
  Case "InsertAfter":
    For each candidate node n do
      Insert source parameter as a sibling node after node n using
      InsertSiblingAfter Algorithm;
      Insert EXEL metadata of inserted node into EXEL-METADATA relation;
    Break;
  Case "InsertChild":
    For each candidate node n do
      Insert source parameter as the child of node n using InsertChildOf Algorithm;
      Insert EXEL metadata of inserted node into EXEL-METADATA relation;
    Break;
  Case "InsertParent":
    For each candidate node n do
      Insert source parameter as the parent of node n using InsertParentOf Algorithm;
      Insert EXEL metadata of inserted node into EXEL-METADATA relation;
    Break;
}

```

**Fig. 5.** User query execution algorithm

## 5 Conclusion and Future Works

In this study, we propose an XML access control mechanism integrated well with EXEL encoding and labeling scheme. Consequently, the process of re-labeling the existing node is not required when the XML document is updated. Also, another benefit of integrating access control mechanism with EXEL encoding and labeling scheme for XML document is to accelerate the query response time depending on access authorizations.

As a future study, we intend to compare our proposed mechanism with the traditional node filtering mechanism. Also, we intend to extend our mechanism to support value based predicates for user queries as well as access authorizations.

## References

1. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible Markup Language (XML) 1.0. W3C Recommendation, 10th edn. (2008), <http://www.w3.org/TR/REC-xml/>
2. Clark, J., DeRose, S.: XML Path Language (XPath) Version 1.0 (1999), <http://www.w3.org/TR/xpath/>
3. Boag, S., Chamberlin, D., Fernández, M.F., Florescu, D., Robie, J., Siméon, J.: XQuery 1.0: An XML Query Language (2007), <http://www.w3.org/TR/xquery/>
4. Li, Q., Moon, B.: Indexing and Querying XML Data for Regular Path Expressions. In: Proceedings of the 27th International Conference on Very Large Data Bases, pp. 361–370. Morgan Kaufmann, Roma (2001)
5. Zhang, C., Naughton, J., DeWitt, D., Luo, Q., Lohman, G.: On Supporting Containment Queries in Relational Database Management Systems. *ACM SIGMOD Record Journal* 30(2), 425–436 (2001)
6. Tatarinov, I., Viglas, S.D., Beyer, K., Shanmugasundaram, J., Shekita, E., Zhang, C.: Storing and Querying Ordered XML Using a Relational Database System. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 204–215. ACM Press, Madison (2002)
7. Min, J.-K., Lee, J., Chung, C.-W.: An Efficient XML Encoding and Labeling Method for Query Processing and Updating on Dynamic XML Data. *Journal of Systems and Software* 82(3), 503–515 (2009)
8. Min, J.-K., Lee, J.-H., Chung, C.-W.: An Efficient Encoding and Labeling for Dynamic XML Data. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 715–726. Springer, Heidelberg (2007)
9. Mirabi, M., Ibrahim, H., Mamat, A., Udzir, N.I., Fathi, L.: Controlling Label Size Increment of Efficient XML Encoding and Labeling Scheme in Dynamic XML Update. *Journal of Computer Science* 6(12), 1529–1534 (2010)
10. Bertino, E., Castano, S., Ferrari, E., Mesiti, M.: Specifying and Enforcing Access Control Policies for XML Document Sources. *Journal of World Wide Web* 3(3), 139–151 (2000)
11. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: Securing XML Documents. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) EDBT 2000. LNCS, vol. 1777, pp. 121–135. Springer, Heidelberg (2000)
12. Damiani, E., di Vimercati, S.D.C., Paraboschi, S., Samarati, P.: A Fine-Grained Access Control System for XML Documents. *Journal of ACM Transactions on Information and System Security (TISSEC)* 5(2), 169–202 (2002)

13. Jo, S.-M., Kim, K.-T., Kouh, H.-J., Yoo, W.-H.: Access Authorization Policy for XML Document Security. In: Chen, G., Pan, Y., Guo, M., Lu, J. (eds.) ISPA-WS 2005. LNCS, vol. 3759, pp. 589–598. Springer, Heidelberg (2005)
14. Jo, S.-M., Yang, C.-M., Yoo, W.-H.: XML Access Control for Security and Memory Management. In: Alford, M.W., Hommel, G., Schneider, F.B., Ansart, J.P., Lamport, L., Mullery, G.P., Zhou, T.H. (eds.) Distributed Systems. LNCS, vol. 190, pp. 179–189. Springer, Heidelberg (1985)
15. Lee, J.-G., Whang, K.-Y., Han, W.-S., Song, I.-Y.: The Dynamic Predicate: Integrating Access Control with Query Processing in XML Databases. *The VLDB Journal* 16(3), 371–387 (2007)
16. Murata, M., Tozawa, A., Kudo, M., Hada, S.: XML Access Control Using Static Analysis. *Journal of ACM Transactions on Information and System Security (TISSEC)* 9(3), 292–324 (2006)
17. Fan, W., Chan, C.-Y., Garofalakis, M.: Secure XML Querying with Security Views. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 587–598. ACM Press, Paris (2004)
18. Damiani, E., Fansi, M., Gabillon, A., Marrara, S.: A General Approach to Securely Querying XML. *Journal of Computer Standards & Interfaces* 30(6), 379–389 (2008)
19. Rassadko, N.: Query Rewriting Algorithm Evaluation for XML Security Views. In: Jonker, W., Petković, M. (eds.) SDM 2007. LNCS, vol. 4721, pp. 64–80. Springer, Heidelberg (2007)
20. Rassadko, N.: Policy Classes and Query Rewriting Algorithm for XML Security Views. In: Damiani, E., Liu, P. (eds.) Data and Applications Security 2006. LNCS, vol. 4127, pp. 104–118. Springer, Heidelberg (2006)
21. Damiani, E., Fansi, M., Gabillon, A., Marrara, S.: Securely Updating XML. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 1098–1106. Springer, Heidelberg (2007)
22. Luo, B., Lee, D., Lee, W.-C., Liu, P.: QFilter: Fine-Grained Run-Time XML Access Control via NFA-based Query Rewriting. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 543–552. ACM Press, Washington, D.C., USA (2004)
23. Byun, C., Park, S.: Two Phase Filtering for XML Access Control. In: Jonker, W., Petković, M. (eds.) SDM 2006. LNCS, vol. 4165, pp. 115–130. Springer, Heidelberg (2006)
24. Byun, C., Park, S.: An Efficient Yet Secure XML Access Control Enforcement by Safe and Correct Query Modification. In: Bressan, S., Küng, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 276–285. Springer, Heidelberg (2006)
25. Tan, K.-L., Lee, M.L., Wang, Y.: Access Control of XML Documents in Relational Database Systems. In: Proceedings of the International Conference on Internet Computing, pp. 185–191. CSREA Press, Las Vegas (2001)
26. Kuper, G., Massacci, F., Rassadko, N.: Generalized XML security views. *International Journal of Information Security* 8(3), 173–203 (2009)
27. Kuper, G., Massacci, F., Rassadko, N.: Generalized XML security views. In: Proceedings of the 10th ACM Symposium on Access Control Models and Technologies, pp. 77–84. ACM Press, Stockholm (2005)

# An Access Control Model for Supporting XML Document Updating

Meghdad Mirabi, Hamidah Ibrahim, Leila Fathi, Nur Izura Udzir, and Ali Mamat

Department of Computer Science, Faculty of Computer Science and Information Technology,  
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia  
meghdad.mirabi@gmail.com, hamidah@fsktm.upm.edu.my, fathi\_leila67@yahoo.com,  
izura@fsktm.upm.edu.my, ali@fsktm.upm.edu.my

**Abstract.** Recently, specifying access control for XML documents over World Wide Web have been proposed. Most of the proposed models for specifying XML access control only support read privilege. Therefore, offering an access control model considering update operations is a challenging issue in XML access control specification. This paper proposes an XML access control model to specify access authorizations which deal with updating the well-formed XML documents. Also, a formal representation of update operations supported by our proposed model is defined in terms of the different policies which are taken for each update operation.

**Keywords:** Access Authorization, XML Access Control Model, XML Updating.

## 1 Introduction

Recently, XML [1] as a de facto standard for sharing and exchanging information over Internet has been suggested. A vital topic is to provide access control over XML document to make sure that only authorized users have permission to access to the portion of XML document they are permitted to.

Access control is the process of determining whether the user requests to the system resources should be granted or denied. The requirements of developing an access control system are first to define access control rules according to which accesses should be controlled and second to implement access control functions executable by the computer system [2]. However, the focus of this paper is the first requirement which is defining a set of XML access control rules. We study how we can specify access control rules for the well-formed XML documents.

The basic concepts of access control are *access control policy*, *access control model*, and *access control mechanism*. Based on the access for which adjustments should be made, the access control rules should be defined by the access control policy. A formal representation of access control policy and its operation can be defined through access control model. Those controls that are enforced by the policy and the formal statement in the model are implemented by the functions of the access control mechanism [2].

Generally, access control policy includes a set of access control rules expressed in natural language. Therefore, it is needed to translate these access control rules to a formal representation through an access control model.

There has been a number of researches with different aspects to XML access control: to propose languages to specify access control policies [3-6], to devise mechanisms to enforce access policies [7-14], and to suggest special data structure to optimize query processing [15-18]. Most of the previous researches only support read privileges. Therefore, one of the challenging issues is to specify XML access control when access policies support update rights.

This paper proposes an access control model for the well-formed XML document which is able to handle update operations such as insertion, deletion, and modification. Our proposed model transforms the access control rules expressed in natural language to a formal representation of access authorizations for updating the well-formed XML document.

The paper proceeds as follows: related works are analyzed in Section 2; the proposed XML access control model is introduced in Section 3; and, finally, conclusions are drawn in Section 4.

## 2 Related Works

The idea of an XML security view is originally proposed in [8]. The security view only contains the information which the users are authorized to access. The access control model proposed by [8] provides an XML view for each user group as well as a DTD view which the XML view conforms to. In contrast to [8], the proposed model in [7, 9] consider general XML DTDs defined in terms of regulations rather than normalized DTDs. Furthermore, [7, 9] do not permit dummy element types in the definition of security views. However, the access control models proposed by [7-9] only support read privilege and specifying XML access control when access policies support update rights is untouched.

A declarative access constraint specification (DACS) language is proposed by [6] to help DBAs in order to specify access constraints to be assigned on the XML documents. This language is able to reveal or hide XML nodes and block the structural relationship between a node and its sub-tree. Also, an algebraic security view specification language SSX is devised in [3] to hide or recognize XML nodes or XML sub-trees. The approach proposed by [14] simplifies the task of DBAs in specifying access constraints on an XML document. The formal algorithm proposed in [14] uses DACS language to generate possible XML views. Therefore, DBA uses DACS language to specify access constraints and then DACS primitives are converted to SSX. Then, all the possible security views are generated based on SSX primitives.

The main idea of static analysis proposed by [10] as a pre-processing access control system is to make automata for XML queries, XML access control policies, and XML schemas and then compare them. As a result, it does not examine real XML documents and run time checking is not necessary. The static analysis cannot totally eliminate run time checking. This method classifies an XML query at compile time

into three categories: entirely authorized, entirely prohibited, or partially authorized ones. Entirely authorized or entirely prohibited queries can be executed without access control. However, the static analysis cannot gain any benefits when a query is classified as a partially authorized. QFilter [11] as an external pre-processing XML access control system checks XPath queries against access control policies and rewrite queries based on access control policies. Static analysis method [10] needs to run time checking to filter out the unauthorized data while QFilter [11] solves this problem by rewriting XPath queries to filter out unauthorized part of input queries before passing them to XML query engine. Static analysis [10] and QFilter [11] only support read privileges similar to [7-9].

In [12], a language similar to XPath [19] is employed to specify the *objects*. The write actions supported by the model proposed in [12] are: *append* and *write*. The *append* privilege allows a *subject* to modify the content of the XML element while the *write* privilege allows to modify the content of the XML element by deleting the node. In addition, the proposed model in [12] defines *propagation options* in order to determine whether an access control rule should be applied to an XML element or to an XML element and all of its descendants.

In [13], XPath expression [19] is employed in order to specify the *objects*. Access authorizations are specified at two levels: DTD level and document level and they can be local or recursive. A local authorization is propagated to an XML element and its attributes while a recursive authorization is propagated to an XML element and all of its descendants. The write actions supported by the model proposed in [13] are: *insert*, *delete*, and *update*. The accessibility of each XML element for write actions is based on a labeling algorithm. Labeling is the process of signing each XML element with “+” if it is accessible or “-“ if it is not accessible.

### 3 The Proposed Access Control Model

Here, we first explain a motivating example regarding the XML document updating and then we present our proposed XML access control model.

Consider the well-formed XML document in Fig. 1 and its access control rules in Fig. 2. The access control rules are expressed in natural language.

Assume that a nurse wishes to modify the name of a patient. According to the rule #2, such a request is denied. Now assume that a doctor wishes to modify his/her name. Such a request does not have any access control rules. This is one of the problems in specifying access authorizations. Another problem is when a request has two conflicting rules. For instance, according to rule #3, a medical-staff cannot modify his/her salary while rule #4 says that a medical-staff can modify his/her private information. Note that Salary element is a descendant of Medical-Staff element and by default each rule which explicitly defines for Medical-Staff element is implicitly defined for its descendants.

In the case of supporting update privileges by an XML access control model, subjects can have permission to insert a new node as a child of leaf nodes as well as a sibling or a parent node. Also, they can have privilege to delete leaf nodes as well as rename and update them.



```

<?xml version="1.0" ?>
<!DOCTYPE Hospital Department SYSTEM "dept.dtd">
<Department Name="Medicine">
  <Division> Cardiology </Division>
  <Medical-Staff>
    <Physician>
      <Name> Bob </Name>
      <Specialty> Nuclear Cardiology </Specialty>
      <Office> CD393 </Office>
      <Address>
        <Street> 25 Cherry Ave. </Street>
        <City> Emeryville </City>
        <Add-Line> CA, 94808 </Add-Line>
      </Address>
      <Salary> $ 30.000 </Salary>
    </Physician>
    <Nurse>
      <Name> Tina </Name>
      <Address>
        <Street> 14th St. </Street>
        <City> Oakland </City>
        <Add-Line> CA, 94705 </Add-Line>
      </Address>
      <Salary> $ 20.000 </Salary>
    </Nurse>
  </Medical-Staff>
  <Patient>
    <Name> Jane </Name>
    <Address>
      <Street> 10 Wayne Dr. </Street>
      <City> Berkeley </City>
      <Add-Line> CA, 94720 </Add-Line>
    </Address>
    <Room>
      <Number> 5 </Number>
      <Bed> 1 </Bed>
    </Room>
    <Illness> Angina </Illness>
    <Therapy>
      <Type> P.T.C.A. </Type>
      <Drug>
        <Name> heparin </Name>
        <Daily-Admin> 30 U/Kg </Daily-Admin>
        <Cost> $ 20 </Cost>
      </Drug>
    </Therapy>
  </Patient>
</Department>

```

**Fig. 1.** The well-formed XML document

1. a doctor can insert and delete the drug information;
2. a nurse cannot modify the patient information;
3. a medical-staff cannot modify his/her salary;
4. a medical-staff can modify his/her private information;

**Fig. 2.** Access control rules

The update operations supported by our proposed XML access control model are as follows:

- InsertChild (source, target)
- InsertBefore | After (source, target)
- InsertParent (source, target)
- Delete (target)
- Update (source, target)
- Rename (source, target)

InsertChild is an insert operation, in which source can be a PCDATA, an element or an attribute. InsertChild inserts source as the child of element denoted by target. If the XML document contains a sequence of information, InsertBefore and InsertAfter are employed. InsertBefore inserts source before element denoted by target, and InsertAfter does after element denoted by target. In addition, InsertParent inserts source as the parent node of element denoted by target. Delete is a delete operation, in which target can be a PCDATA, an element or an attribute. Update is an update operation, in which target can be an element or an attribute, and source can be a PCDATA. Rename is a rename operation, in which target can be an element or an attribute, and source is a new name.

An access control policy includes a set of access authorizations. In general, an authorization can be defined as 4-tuple  $\langle \text{subject}, \text{object}, \text{action}, \text{permission} \rangle$  where subject is the user or role concerned by the authorization; Object is presented by XPath expression [19] which contains the element(s) of the XML document; Action is an executable action which can be InsertChild, InsertBefore, InsertAfter, InsertParent, Delete, Update, and Rename; Permission represents the acceptance (+) or denial (-) of rights.

Now, the semantic of an access authorization  $\langle \text{subject}, \text{object}, \text{action}, \text{permission} \rangle$  is explained informally for each authorization action supported by our proposed model. We consider only positive access authorizations for simplicity and negative access authorizations are left.

Let  $T$  be the well-formed XML document and  $S$  be the set of XML nodes returned from the evaluation of XPath expression of object on  $T$ .

- $\langle \text{subject}, \text{object}, \text{InsertChild}, + \rangle$ : subject can insert a new node as a child node of the nodes in  $S$ .
- $\langle \text{subject}, \text{object}, \text{InsertBefore/InsertAfter}, + \rangle$ : subject can insert a new node as a preceding/following sibling node of the nodes in  $S$ .
- $\langle \text{subject}, \text{object}, \text{InsertParent}, + \rangle$ : subject can insert a new node as a parent node of the nodes in  $S$ .
- $\langle \text{subject}, \text{object}, \text{Delete}, + \rangle$ : subject can delete the nodes in  $S$ .
- $\langle \text{subject}, \text{object}, \text{Update}, + \rangle$ : subject can update the content of nodes in  $S$ .
- $\langle \text{subject}, \text{object}, \text{Rename}, + \rangle$ : subject can rename the nodes in  $S$ .

Let us consider an example. According to rule #1 in Fig. 2, a doctor can insert and delete the drug information. The rule expressed in natural language can be written as follows:

(Doctor, //Therapy/Drug, InsertChild, +)  
 (Doctor, //Therapy/Drug//\*, Delete, +)

Given an access control policy  $ACP$  which is a set of access authorizations and a well-formed XML document  $T$ , the semantics of  $ACP$  determines the XML nodes of  $T$  to which a subject can apply a certain update operation.

Now, we consider when an XML node is accessible for a specific update action. The following two questions must be answered in order to determine the accessibility of an XML node.

1. What happens if there is no defined access control rule for an XML node?
2. What happens if there are both positive and negative access control rules for an XML node?

We can answer these two questions by defining a *default policy* and an *override policy*. The *default policy* says that if a rule is not defined explicitly for a specific XML node, it can either by default allowed, or by default forbidden while the *override policy* says that either a positive rule overrides a conflicting negative one, or a negative rule overrides a conflicting positive one. Several approaches have been suggested to solve the problem of conflicts between access control rules such as *deny overrides* and *grant overrides* [2]. In *deny overrides*, negative access authorizations have priority over positive access authorizations while in *grant overrides*, positive access authorizations have priority over negative access authorizations.

Fig. 3 shows the combination of *default policy* with *denies override* as *the override policy* while Fig. 4 shows the combination of *default policy* with *grant overrides* as *the override policy*. Here,  $P$  denotes the truth value of *has positive permission to carry out the update operation*,  $N$  denotes the truth value of *has negative permission to carry out the update operation*,  $A$  denotes *accessible*, and  $NA$  denotes *inaccessible*. For instance, Fig. 3 says that an XML node is accessible if and only if “ $P$ , not  $N$ ” for *deny* as the *default policy*.

$[P, N] \rightarrow NA/NA$ $[\text{not } P, \text{not } N] \rightarrow A/NA$ $[\text{not } P, N] \rightarrow NA/NA$ $[P, \text{not } N] \rightarrow A/A$
---

**Fig. 3.** Allow/deny as the default policy if deny is the override policy

$[P, N] \rightarrow A/A$ $[\text{not } P, \text{not } N] \rightarrow A/NA$ $[\text{not } P, N] \rightarrow NA/NA$ $[P, \text{not } N] \rightarrow A/A$
---

**Fig. 4.** Allow/deny as the default policy if grant is the override policy

Now, we consider when an XML node is accessible for a specific update operation. Assume  $S$  is a set of XML nodes obtained by evaluating the XPath expression of target parameter in update operation on a well-formed XML document  $T$  and  $\vee$  denotes OR and  $\wedge$  denotes AND. Also, we assume  $P_a/N_a$  denotes a set of positive/negative permissions for a specific action  $a$ .

**a) Allow as the default policy and deny overrides as the override policy**

- InsertChild (source, target): “insert source as a child node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is not in the scope of a negative InsertChild. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{InsertChild}}} \text{not self} :: f]$$

- InsertBefore/After (source, target): “insert source as a preceding/following node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is not in the scope of a negative InsertBefore/After. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{InsertBefore}}} \text{not self} :: f], [\wedge_{f \in N_{\text{InsertAfter}}} \text{not self} :: f]$$

- InsertParent(source, target): “insert source as a parent node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is not in the scope of a negative InsertParent. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{InsertParent}}} \text{not self} :: f]$$

- Delete (target): “delete node  $n$ ” is granted if:  $n \in S$ ;  $n$  is not in the scope of a negative Delete. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{Delete}}} \text{not self} :: f]$$

- Update (source, target): “update the content of the target node with source” is granted to node  $n$  if:  $n \in S$ ;  $n$  is not in the scope of a negative Update. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{Update}}} \text{not self} :: f]$$

- Rename (source, target): “rename the target node with source” is granted to node  $n$  if:  $n \in S$ ;  $n$  is not in the scope of a negative Rename. Formally the above conditions can be expressed as follows:

$$[\wedge_{f \in N_{\text{Rename}}} \text{not self} :: f]$$

**b) Deny as the default policy and deny overrides as the override policy**

- Insert Child (source, target): “insert source as a child node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is in the scope of a positive InsertChild;  $n$  is not in the scope of a negative InsertChild. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{InsertChild}}} \text{self} :: p \wedge_{f \in N_{\text{InsertChild}}} \text{not self} :: f]$$

- InsertBefore|After (source, target): “insert source as a preceding/following node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is in the scope of a positive InsertBefore|After;  $n$  is not in the scope of a negative InsertBefore|After. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{InsertBefore}}} \text{self} :: p \wedge_{f \in N_{\text{InsertBefore}}} \text{not self} :: f], [\vee_{p \in P_{\text{InsertAfter}}} \text{self} :: p \wedge_{f \in N_{\text{InsertAfter}}} \text{not self} :: f]$$

- InsertParent(source, target): “insert source as a parent node of target node” is granted to node  $n$  if:  $n \in S$ ;  $n$  is in the scope of a positive InsertParent;  $n$  is not in the scope of a negative InsertParent. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{InsertParent}}} \text{self} :: p \wedge_{f \in N_{\text{InsertParent}}} \text{not self} :: f]$$

- Delete (target): “delete node  $n$ ” is granted if:  $n \in S$ ;  $n$  is in the scope of a positive Delete;  $n$  is not in the scope of a negative Delete. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{Delete}}} \text{self} :: p \wedge_{f \in N_{\text{Delete}}} \text{not self} :: f]$$

- Update (source, target): “update the content of the target node with source” is granted to node  $n$  if:  $n \in S$ ;  $n$  is in the scope of a positive Update;  $n$  is not in the scope of a negative Update. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{Update}}} \text{self} :: p \wedge_{f \in N_{\text{Update}}} \text{not self} :: f]$$

- Rename (source, target): “rename the target node with source” is granted to node  $n$  if:  $n \in S$ ;  $n$  is in the scope of a positive Rename;  $n$  is not in the scope of a negative Rename. Formally the above conditions can be expressed as follows:

$$[\vee_{p \in P_{\text{Rename}}} \text{self} :: p \wedge_{f \in N_{\text{Rename}}} \text{not self} :: f]$$

In this way, we can also express formally the conditions in the case of *allow* as *the default policy* and *grant overrides* as *the override policy* and *deny* as *the default policy* and *grant overrides* as *the override policy*.

## 4 Conclusion and Future Works

In this paper, we propose an access control model to specify the access authorizations for the well-formed XML documents considering update operations. Also, a formal representation of update operations when a node is accessible is presented in terms of the different policies which are taken for each update operation.

As future work, we plan to extend our proposed XML access control model to support value based predicates as extra conditions.

## References

1. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: ExtensibleMarkup Language (XML) 1.0 (5th Edition). W3C Recommendation (2008), <http://www.w3.org/TR/REC-xml/>,
2. Samarati, P., di Vimercati, S.D.C.: Access control: Policies, models, and mechanisms. In: Focardi, R., Gorrieri, R. (eds.) FOSAD 2000. LNCS, vol. 2171, pp. 137–196. Springer, Heidelberg (2001)
3. Mohan, S., Klinginsmith, J., Sengupta, A., Wu, Y.: ACXESS - Access Control forXML with Enhanced Security Specifications. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006), pp. 171–171. IEEE Computer Society Press, New York (2006)
4. Moses, T.: eXtensible Access Control Markup Language (XACML) Version 2.0.OASIS Standard (2005), [http://docs.oasis-open.org/xacml/2.0/access\\_control-xacml-2.0-core-spec-os.pdf](http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf),
5. Hada, S., Kudo, M.X.: Access Control Language: Provisional Authorization forXML Documents (2000), <http://www.tr1.ibm.com/projects/xml/xacl/xacl-spec.html>
6. Mohan, S., Wu, Y.: IPAC: An Interactive Approach to Access Control for Semi-Structured Data. In: Proceedings of the 32nd International Conference on Very LargeData Bases,VLDB Endowment, Korea, pp. 1147–1150 (2006)
7. Kuper, G., Massacci, F., Rassadko, N.: Generalized XML Security Views. In: Proceedings of the 10th ACM Symposium on Access Control Models and Technologies, pp. 77–84. ACM Press, New York (2005)
8. Fan, W., Chan, C.-Y., Garofalakis, M.: Secure XML Querying with Security Views. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. ACM Press, New York (2004)
9. Kuper, G., Massacci, F., Rassadko, N.: Generalized XML Security Views. International Journal of Information Security 8(3), 173–203 (2009)
10. Murata, M., Tozawa, A., Kudo, M., Hada, S.: XML Access Control Using StaticAnalysis. Journal of ACM Transactions on Information and System Security(TISSEC) 9(3), 292–324 (2006)
11. Luo, B., Lee, D., Lee, W.-C., Liu, P.: QFilter: Fine-Grained Run-Time XML AccessControl via NFA-based Query Rewriting. In: Proceedings of the Thirteenth ACMInternational Conference on Information and Knowledge Management, pp. 543–552. ACM Press, New York (2004)

12. Bertino, E., Castano, S., Ferrari, E., Mesiti, M.: Specifying and Enforcing AccessControl Policies for XML Document Sources. *Journal of World Wide Web* 3(3), 139–151 (2000)
13. Damiani, E., Vimercati, S.D.C.D., Paraboschi, S., Samarati, P.: A Fine-GrainedAccess Control System for XML Documents. *Journal of ACM Transactions on Information and System Security (TISSEC)* 5(2), 169–202 (2002)
14. Tripathi, A., Gore, M.M.: Hasslefree: Simplified Access Control Management for XML Documents. In: Janowski, T., Mohanty, H. (eds.) *ICDCIT 2007*. LNCS, vol. 4882, pp. 116–128. Springer, Heidelberg (2007)
15. Cho, S., Amer-Yahia, S., Lakshmanan, L.V.S., Srivastava, D.: Optimizing the SecureEvaluation of Twig Queries. In: *Proceedings of the 28th International Conference on Very Large Data Bases, VLDB Endowment, China*, pp. 490–501 (2002)
16. Yu, T., Srivastava, D., Lakshmanan, L.V.S., Jagadish, H.V.: A CompressedAccessibility Map for XML. *Journal of ACM Transactions on Database Systems(TODS)* 29(2), 363–402 (2004)
17. Jiang, M., Fu, A.W.-C.: Integration and Efficient Lookup of Compressed XMLAccessibility Maps. *IEEE Transactions on Knowledge and Data Engineering* 17(7), 939–953 (2005)
18. Zhang, H., Zhang, N., Salem, K., Zhuo, D.: Compact Access Control Labeling forEfficient Secure XML Query Evaluation. *Journal of Data & Knowledge Engineering* 60(2), 326–344 (2007)
19. Clark, J., DeRose, S.X.: *Path Language (XPath) Version 1.0* (1999), <http://www.w3.org/TR/xpath/>,

# A Secure Proxy Blind Signature Scheme Using ECC

Daniyal M. Alghazzawi, Trigui Mohamed Salim, and Syed Hamid Hasan\*

Department of Information Systems,  
King Abdul Aziz University, Kingdom of Saudi Arabia  
shh786@hotmail.com

**Abstract.** This paper describes an efficient simple proxy blind signature scheme. The security of the scheme is based on Elliptic Curve Discrete Logarithm Problem (ECDLP). This can be implemented in low power and small processor mobile devices such as smart card, PDA etc. A proxy blind signature scheme is a special form of blind signature which allows a designated person called proxy signer to sign on behalf of two or more original signers without knowing the content of the message or document. It combines the advantages of proxy signature and blind signature scheme and satisfies the security properties of both proxy and blind signature scheme.

**Keywords:** ECDLP, blind signature, proxy signature.

Blind signature scheme was first introduced by Chaum [3]. It is a protocol for obtaining a signature from a signer on any message, without revealing any information about the message or its signature. In 1996, Mamo et al proposed the concept of proxy signature [1]. In proxy signature scheme, the original signer delegates his signing capacity to a proxy signer who can sign a message submitted on behalf of the original signer. A verifier can validate its correctness and can distinguish between a normal signature and a proxy signature. In multi-proxy signature scheme, an original signer is allowed to authorize a group of proxy members to generate the multi signature on behalf of the original signer. In 2000, Hwang et al. proposed the first multi-proxy signature scheme [4]. A proxy blind signature scheme is a digital signature scheme that ensures the properties of proxy signature and blind signature. In a proxy blind signature, an original signer delegates his signing capacity to proxy signer.

## 1 Background

In this section we brief overview of prime field, Elliptic Curve over that field and Elliptic Curve Discrete Logarithm Problem.

---

\* Corresponding author. jayaprakashkar@yahoo.com



### 1.1 The Finite Field $F_p$

Let  $p$  be a prime number. The finite field  $F_p$  is comprised of the set of integers  $0, 1, 2, \dots, p-1$  with the following arithmetic operations [5] [6] [7]:

- Addition: If  $a, b \in F_p$ , then  $a + b = r$ , where  $r$  is the remainder when  $a + b$  is divided by  $p$  and  $0 \leq r \leq p-1$ . This is known as addition modulo  $p$ .
- Multiplication: If  $a, b \in F_p$ , then  $a.b = s$ , where  $s$  is the remainder when  $a.b$  is divided by  $p$  and  $0 \leq s \leq p-1$ . This is known as multiplication modulo  $p$ .
- Inversion: If  $a$  is a non-zero element in  $F_p$ , the inverse of  $a$  modulo  $p$ , denoted  $a^{-1}$ , is the unique integer  $c \in F_p$  for which  $a.c = 1$ .

### 1.2 Elliptic Curve over $F_p$

Let  $p \geq 3$  be a prime number. Let  $a, b \in F_p$  be such that  $4a^3 + 27b^2 \neq 0$  in  $F_p$ . An elliptic curve  $E$  over  $F_p$  defined by the parameters  $a$  and  $b$  is the set of all solutions  $(x, y)$ ,  $x, y \in F_p$ , to the equation  $y^2 = x^3 + ax + b$ , together with an extra point  $O$ , the point at infinity. The set of points  $E(F_p)$  forms an Abelian group with the following addition rules [9]:

1. Identity :  $P + O = O + P = P$ , for all  $P \in E(F_p)$
2. Negative : if  $P(x, y) \in E(F_p)$  then  $(x, y) + (x, -y) = O$ , The point  $(x, -y)$  is denoted as  $-P$  called negative of  $P$ .
3. Point addition: Let  $P((x_1, y_1), Q(x_2, y_2) \in E(F_p)$ , then  $P + Q = R \in E(F_p)$  and coordinate  $(x_3, y_3)$  of  $R$  is given by  $x_3 = \lambda^2 - x_1 - x_2$  and  $y_3 = \lambda(x_1 - x_3) - y_1$  where  $\lambda = \frac{y_2 - y_1}{x_2 - x_1}$
4. Point doubling : Let  $P(x_1, y_1) \in E(F_p)$  where  $P \neq -P$  then  $2P = (x_3, y_3)$  where  $x_3 = (\frac{3x_1^2 + a}{2y_1})^2 - 2x_1$  and  $y_3 = (\frac{3x_1^2 + a}{2y_1})(x_1 - x_3) - y_1$ .

### 1.3 Elliptic Curve Discrete Logarithm Problem (ECDLP)

Given an elliptic curve  $E$  defined over a finite field  $F_p$ , a point  $P \in E(F_p)$  of order  $n$ , and a point  $Q \in \langle P \rangle$ , find the integer  $l \in [0, n-1]$  such that  $Q = lP$ . The integer  $l$  is called discrete logarithm of  $Q$  to base  $P$ , denoted  $l = \log_P Q$  [9].

## 2 Preliminaries

### 2.1 Notations

Common notations used in this paper as follows.

- $p$  : the order of underlying finite field.
- $F_p$  : the underlying finite field of order  $p$
- $E$  : elliptic curve defined on finite field  $F_p$  with large order.
- $G$  : the group of elliptic curve points on  $E$ .
- $P$  : a point in  $E(F_p)$  with order  $n$ , where  $n$  is a large prime number.
- $\mathcal{H}(\cdot)$  : a secure one-way hash function.
- $d$  : the secret key of the original signer  $S$  to be chosen randomly from  $[1, n-1]$ .
- $Q$  is the public key of the original signer  $S$ , where  $Q = d \cdot P$ .
- $\parallel$  : Concatenation operation between two bit strings.

### 3 Proxy Signature and Proxy Blind Signature

A proxy blind signature is a digital signature scheme that ensures the properties of proxy signature and blind signature schemes. Proxy blind signature scheme is an extension of proxy blind signature, which allows a single designated proxy signer to generate a blind signature on behalf of group of original signers. A proxy blind signature scheme consists of the following three phases:

- Proxy key generation
- Proxy blind signature scheme
- Signature verification

### 4 Security Properties

The security properties described in [2] for a secure blind signature scheme are as follows

- **Distinguishability:** The proxy blind signature must be distinguishable from the ordinary signature.
- **Strong unforgeability:** Only the designated proxy signer can create the proxy blind signature for the original signer.
- **Non-repudiation:** The proxy signer can not claim that the proxy signer is disputed or illegally signed by the original signer.
- **Verifiability:** The proxy blind signature can be verified by everyone. After verification, the verifier can be convinced of the original signer's agreement on the signed message.
- **Strong undeniability:** Due to fact that the delegation information is signed by the original signer and the proxy signature are generated by the proxy signer's secret key. Both the signer can not deny their behavior.
- **Unlinkability:** When the signer is revealed, the proxy signer can not identify the association between the message and the blind signature he generated.
- **Secret key dependencies:** Proxy key or delegation pair can be computed only by the original signer's secret key.
- **Prevention of misuse:** The proxy signer cannot use the proxy secret key for purposes other than generating valid proxy signatures. In case of misuse, the responsibility of the proxy signer should be determined explicitly.

### 5 Proposed Protocol

The protocol involves three entities : Original signer  $S$ , Proxy signer  $P_s$  and verifier  $V$ . It is described as follows.

#### 5.1 Proxy Phase

- **Proxy generation:** The original signer  $S$  selects random integer  $k$  in the interval  $[1, n - 1]$ . Computes  $R = k \cdot P = (x_1, y_1)$  and  $r = x_1 \bmod n$ . Where  $x_1$  is regarded as an integer between 0 and  $q - 1$ . Then computes  $s = (d + k \cdot r) \bmod n$  and computes  $Q_p = s \cdot P$ .

- **Proxy delivery:** The original signer  $S$  sends  $(s, r)$  to the proxy signer  $P_s$  and make  $Q_p$  public.
- **Proxy Verification:** After receiving the secret key pairs  $(s, r)$ , the proxy signer  $P_s$  checks the validity of the secret key pairs  $(s, r)$  with the following equation.

$$Q_p = s \cdot P = Q + r \cdot R \quad (1)$$

## 5.2 Signing Phase

- The Proxy signer  $S_p$  chooses random integer  $t \in [1, n - 1]$  and computes  $U = t \cdot P$  and sends it to the verifier  $V$ .
- After receiving the verifier chooses randomly  $\alpha, \beta \in [1, n - 1]$  and computes the following

$$\tilde{R} = U + \alpha \cdot P - \beta \cdot Q_p \quad (2)$$

$$\tilde{e} = \mathcal{H}(\tilde{R} \| M) \quad (3)$$

$$e = (\tilde{e} + \beta) \bmod n \quad (4)$$

and verifier  $V$  sends  $e$  to the proxy signer  $S_p$

- After receiving  $e$ ,  $S_p$  computes the following

$$\tilde{s} = (t - s \cdot e) \bmod n \quad (5)$$

and sends it to  $V$ .

- Now  $V$  computes

$$s_p = (\tilde{s} + \alpha) \bmod n \quad (6)$$

The tuples  $(M, s_p, \tilde{e})$  is the proxy blind signature.

## 5.3 Verification Phase

The verifier  $V$  computes the following equation.

$$\gamma = \mathcal{H}((s_p \cdot P + \tilde{e} \cdot Q_p) \| M) \quad (7)$$

and verifies the validity of proxy blind signature  $(M, s_p, \tilde{e})$  with the equality  $\gamma = \tilde{e}$ .

## 6 Security Analysis

**Theorem 1.** *It is infeasible for adversary  $\mathcal{A}$  to derive signer's private key from all available public information.*

Proof: Assume that the adversary  $\mathcal{A}$  wants to derive signer's private key  $d$  from his public key  $Q$ , he has to solve ECDLP problem which is computationally infeasible. Similarly, the adversary will encounter the same difficulty as she/he tries to obtain proxy signer's private key.

**Theorem 2.** *Proxy signature is distinguishable from original signer's normal signature.*

Proof: Since proxy key is different from original signer's private key and proxy keys created by different proxy signers are different from each other, any proxy signature is distinguishable from original signer's normal signature and different proxy signer's signature are distinguishable.

**Theorem 3.** *The scheme satisfies Unlinkability security requirement*

Proof: In verification stage, the signer checks only whether  $\gamma = \mathcal{H}((s_p \cdot P + \tilde{e} \cdot Q_p) \| M)$  holds. He does not know the original signer's private key and proxy signer's private key. Thus the signer knows neither the message nor the signature associated with the signature scheme.

## 7 Correctness

**Theorem 4.** *The proxy blind signature  $(M, s_p, \tilde{e})$  is universally verifiable by using the system public parameters.*

Proof: The of correctness of the signature is verified as follows

We have to prove that

$$\mathcal{H}((s_p \cdot P + \tilde{e} \cdot Q_p) \| M) = \mathcal{H}(\tilde{R} \| M)$$

$$\begin{aligned} \text{i.e to show } s_p \cdot P + \tilde{e} \cdot Q_p &= \tilde{R} \\ &= (\tilde{s} + \alpha) \cdot P + \tilde{e} \cdot Q_p \\ &= \tilde{s} \cdot P + \alpha \cdot P + \tilde{e} \cdot Q_p \\ &= (t - s \cdot e) \cdot P + \alpha P + \tilde{e} \cdot Q_p \\ &= t \cdot P - e \cdot Q_p + \alpha \cdot P + \tilde{e} \cdot Q_p \\ &= t \cdot P - (\tilde{e} + \beta) \cdot Q_p + \alpha P + \tilde{e} \cdot Q_p \\ &= t \cdot P - \tilde{e} \cdot Q_p - \beta \cdot Q_p + \alpha \cdot P + \tilde{e} \cdot Q_p \\ &= t \cdot P - \beta \cdot Q_p + \alpha \cdot P \\ &= U - \beta \cdot Q_p + \alpha \cdot P \\ &= \tilde{R} \end{aligned}$$

## 8 Conclusion

The security of the scheme is hardness of solving ECDLP. The primary reason for the attractiveness of ECC over systems such as RSA and DSA is that the best algorithm known for solving the underlying mathematical problem namely, the ECDLP takes fully exponential time. In contrast, sub-exponential time algorithms are known for underlying mathematical problems on which RSA and DSA are based, namely the integer factorization (IFP) and the discrete logarithm (DLP) problems. This means that the algorithms for solving the ECDLP become infeasible much more rapidly as the problem size increases more than those algorithms for the IFP and DLP. For this reason, ECC offers security equivalent to RSA and DSA while using far smaller key sizes [2]. The benefits of this higher-strength per-bit include higher speeds, lower power consumption, bandwidth

savings, storage efficiencies, and smaller certificates. This can be implemented in low power and small processor mobile devices such as smart card, PDA etc. In this proposed scheme it is infeasible for adversary to derive signer's private key from all available public information. This protocol also achieves the security like requirements distinguishability, strong unforgeability, non-repudiation, strong undeniability and unlinkability.

## References

1. Mambo, M., Usda, K., Okamoto, E.: Proxy signature: Delegation of power to sign messages. IEICE Transaction on Fundamentals E79-A, 1338–1353 (1996)
2. J.P.Kar Proxy Blind multi-signature scheme using ECC for handheld devices, ePrint Archive: Report 2011/043, <http://eprint.iacr.org/2011/43>
3. Chaum, D.: Blind Signature for Untraceable Payments. In: Crypto 82, pp. 199–203. Plenum Press, New York (1983)
4. Hwang, S.J., Shi, C.H.: A Simple multi-signature scheme. In: Proceeding of 10th National Conference On Information Security, Taiwan (2000)
5. Koblitz, N.: A course in Number Theory and Cryptography, 2nd edn. Springer, Heidelberg (1994)
6. Rosen, K.H.: Elementary Number Theory in Science and Communication, 2nd edn. Springer-Verlag, Berlin (1986)
7. Menezes, A., Van Oorschot, P.C., Vanstone, S.A.: Handbook of applied cryptography. CRC Press, New York (1997)
8. Hankerson, D., Menezes, A., Vanstone, S.: Guide to Elliptic Curve Cryptography. Springer Verlag, Berlin (2004)
9. Certicom, E.C.C.: Challenge and The Elliptic Curve Cryptosystem, <http://www.certicom.com/index.php>.
10. Dwork, C., Naor, M., Sahai, A.: Concurrent zero-knowledge. In: Proceedings of 30th ACM STOC1998, pp. 409–418 (1998)
11. Abdalla, M., Bellare, M., Rogaway, P.: The Oracle Diffie-Hellman Assumptions and an Analysis of DHIES. In: Naccache, D. (ed.) CT-RSA 2001. LNCS, vol. 2020, pp. 143–158. Springer, Heidelberg (2001)
12. Aumann, Y., Rabin, M.O.: Authentication, Enhanced Security and Error Correcting Codes. In: Krawczyk, H. (ed.) CRYPTO 1998. LNCS, vol. 1462, pp. 299–303. Springer, Heidelberg (1998)
13. Diffie, W., Hellman, M.E.: Directions in cryptography. IEEE Transactions on Information Theory 22, 644–654 (1976)
14. Shi, Y., Li, J.: Identity-based deniable authentication protocol. Electronics Letters 41, 241–242 (2005)
15. Shoup, V.: Sequences of games: a tool for taming complexity in security proofs, in Cryptology ePrint Archive: Report 2004/332, <http://eprint.iacr.org/2004/332>

# Accelerated Particle Swarm Optimization and Support Vector Machine for Business Optimization and Applications

Xin-She Yang<sup>1</sup>, Suash Deb<sup>2</sup>, and Simon Fong<sup>3</sup>

<sup>1</sup> Department of Engineering, University of Cambridge, Trumpinton Street, Cambridge CB2 1PZ, UK

`xy227@cam.ac.uk`

<sup>2</sup> Department of Computer Science & Engineering, C.V. Raman College of Engineering, Bidyannagar, Mahura, Janla, Bhubaneswar 752054, India

`suashdeb@gmail.com`

<sup>3</sup> Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Taipa, Macau

`ccfong@umac.mo`

**Abstract.** Business optimization is becoming increasingly important because all business activities aim to maximize the profit and performance of products and services, under limited resources and appropriate constraints. Recent developments in support vector machine and metaheuristics show many advantages of these techniques. In particular, particle swarm optimization is now widely used in solving tough optimization problems. In this paper, we use a combination of a recently developed Accelerated PSO and a nonlinear support vector machine to form a framework for solving business optimization problems. We first apply the proposed APSO-SVM to production optimization, and then use it for income prediction and project scheduling. We also carry out some parametric studies and discuss the advantages of the proposed metaheuristic SVM.

**Keywords:** Accelerated PSO, business optimization, metaheuristics, PSO, support vector machine, project scheduling.

## 1 Introduction

Many business activities often have to deal with large, complex databases. This is partly driven by information technology, especially the Internet, and partly driven by the need to extract meaningful knowledge by data mining. To extract useful information among a huge amount of data requires efficient tools for processing vast data sets. This is equivalent to trying to find an optimal solution to a highly nonlinear problem with multiple, complex constraints, which is a challenging task. Various techniques for such data mining and optimization

have been developed over the past few decades. Among these techniques, support vector machine is one of the best techniques for regression, classification and data mining [5,9,16,19,20,24].

On the other hand, metaheuristic algorithms also become powerful for solving tough nonlinear optimization problems [1,7,8,27,32]. Modern metaheuristic algorithms have been developed with an aim to carry out global search, typical examples are genetic algorithms [6], particle swarm optimisation (PSO) [7], and Cuckoo Search [29,30]. The efficiency of metaheuristic algorithms can be attributed to the fact that they imitate the best features in nature, especially the selection of the fittest in biological systems which have evolved by natural selection over millions of years. Since most data have noise or associated randomness, most these algorithms cannot be used directly. In this case, some form of averaging or reformulation of the problem often helps. Even so, most algorithms become difficult to implement for such type of optimization.

In addition to the above challenges, business optimization often concerns with a large amount but often incomplete data, evolving dynamically over time. Certain tasks cannot start before other required tasks are completed, such complex scheduling is often NP-hard and no universally efficient tool exists. Recent trends indicate that metaheuristics can be very promising, in combination with other tools such as neural networks and support vector machines [5,9,15,21].

In this paper, we intend to present a simple framework of business optimization using a combination of support vector machine with accelerated PSO. The paper is organized as follows: We first will briefly review particle swarm optimization and accelerated PSO, and then introduce the basics of support vector machines (SVM). We then use three case studies to test the proposed framework. Finally, we discuss its implications and possible extension for further research.

## 2 Accelerated Particle Swarm Optimization

### 2.1 PSO

Particle swarm optimization (PSO) was developed by Kennedy and Eberhart in 1995 [7,8], based on the swarm behaviour such as fish and bird schooling in nature. Since then, PSO has generated much wider interests, and forms an exciting, ever-expanding research subject, called swarm intelligence. PSO has been applied to almost every area in optimization, computational intelligence, and design/scheduling applications. There are at least two dozens of PSO variants, and hybrid algorithms by combining PSO with other existing algorithms are also increasingly popular.

PSO searches the space of an objective function by adjusting the trajectories of individual agents, called particles, as the piecewise paths formed by positional vectors in a quasi-stochastic manner. The movement of a swarming particle consists of two major components: a stochastic component and a deterministic component. Each particle is attracted toward the position of the current global

best  $\mathbf{g}^*$  and its own best location  $\mathbf{x}_i^*$  in history, while at the same time it has a tendency to move randomly.

Let  $\mathbf{x}_i$  and  $\mathbf{v}_i$  be the position vector and velocity for particle  $i$ , respectively. The new velocity vector is determined by the following formula

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \alpha\epsilon_1[\mathbf{g}^* - \mathbf{x}_i^t] + \beta\epsilon_2[\mathbf{x}_i^* - \mathbf{x}_i^t]. \quad (1)$$

where  $\epsilon_1$  and  $\epsilon_2$  are two random vectors, and each entry taking the values between 0 and 1. The parameters  $\alpha$  and  $\beta$  are the learning parameters or acceleration constants, which can typically be taken as, say,  $\alpha \approx \beta \approx 2$ .

There are many variants which extend the standard PSO algorithm, and the most noticeable improvement is probably to use an inertia function  $\theta(t)$  so that  $\mathbf{v}_i^t$  is replaced by  $\theta(t)\mathbf{v}_i^t$

$$\mathbf{v}_i^{t+1} = \theta\mathbf{v}_i^t + \alpha\epsilon_1[\mathbf{g}^* - \mathbf{x}_i^t] + \beta\epsilon_2[\mathbf{x}_i^* - \mathbf{x}_i^t], \quad (2)$$

where  $\theta \in (0, 1)$  [23]. In the simplest case, the inertia function can be taken as a constant, typically  $\theta \approx 0.5 \sim 0.9$ . This is equivalent to introducing a virtual mass to stabilize the motion of the particles, and thus the algorithm is expected to converge more quickly.

## 2.2 Accelerated PSO

The standard particle swarm optimization uses both the current global best  $\mathbf{g}^*$  and the individual best  $\mathbf{x}_i^*$ . The reason of using the individual best is primarily to increase the diversity in the quality solutions, however, this diversity can be simulated using some randomness. Subsequently, there is no compelling reason for using the individual best, unless the optimization problem of interest is highly nonlinear and multimodal.

A simplified version which could accelerate the convergence of the algorithm is to use the global best only. Thus, in the accelerated particle swarm optimization (APSO) [27,32], the velocity vector is generated by a simpler formula

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \alpha\epsilon_n + \beta(\mathbf{g}^* - \mathbf{x}_i^t), \quad (3)$$

where  $\epsilon_n$  is drawn from  $N(0, 1)$  to replace the second term. The update of the position is simply

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1}. \quad (4)$$

In order to increase the convergence even further, we can also write the update of the location in a single step

$$\mathbf{x}_i^{t+1} = (1 - \beta)\mathbf{x}_i^t + \beta\mathbf{g}^* + \alpha\epsilon_n. \quad (5)$$

This simpler version will give the same order of convergence. Typically,  $\alpha = 0.1L \sim 0.5L$  where  $L$  is the scale of each variable, while  $\beta = 0.1 \sim 0.7$  is sufficient for most applications. It is worth pointing out that velocity does not appear in equation (5), and there is no need to deal with initialization of velocity



vectors. Therefore, APSO is much simpler. Comparing with many PSO variants, APSO uses only two parameters, and the mechanism is simple to understand.

A further improvement to the accelerated PSO is to reduce the randomness as iterations proceed. This means that we can use a monotonically decreasing function such as

$$\alpha = \alpha_0 e^{-\gamma t}, \quad (6)$$

or

$$\alpha = \alpha_0 \gamma^t, \quad (0 < \gamma < 1), \quad (7)$$

where  $\alpha_0 \approx 0.5 \sim 1$  is the initial value of the randomness parameter. Here  $t$  is the number of iterations or time steps.  $0 < \gamma < 1$  is a control parameter [32]. For example, in our implementation, we will use

$$\alpha = 0.7^t, \quad (8)$$

where  $t \in [0, t_{\max}]$  and  $t_{\max}$  is the maximum of iterations.

### 3 Support Vector Machine

Support vector machine (SVM) is an efficient tool for data mining and classification [25,26]. Due to the vast volumes of data in business, especially e-commerce, efficient use of data mining techniques becomes a necessity. In fact, SVM can also be considered as an optimization tool, as its objective is to maximize the separation margins between data sets. The proper combination of SVM with metaheuristics could be advantageous.

#### 3.1 Support Vector Machine

A support vector machine essentially transforms a set of data into a significantly higher-dimensional space by nonlinear transformations so that regression and data fitting can be carried out in this high-dimensional space. This methodology can be used for data classification, pattern recognition, and regression, and its theory was based on statistical machine learning theory [21,24,25].

For classifications with the learning examples or data  $(\mathbf{x}_i, y_i)$  where  $i = 1, 2, \dots, n$  and  $y_i \in \{-1, +1\}$ , the aim of the learning is to find a function  $\phi_\alpha(\mathbf{x})$  from allowable functions  $\{\phi_\alpha : \alpha \in \Omega\}$  such that  $\phi_\alpha(\mathbf{x}_i) \mapsto y_i$  for  $(i = 1, 2, \dots, n)$  and that the expected risk  $E(\alpha)$  is minimal. That is the minimization of the risk

$$E(\alpha) = \frac{1}{2} \int |\phi_\alpha(x) - y| dQ(\mathbf{x}, y), \quad (9)$$

where  $Q(\mathbf{x}, y)$  is an unknown probability distribution, which makes it impossible to calculate  $E(\alpha)$  directly. A simple approach is to use the so-called empirical risk

$$E_p(\alpha) \approx \frac{1}{2n} \sum_{i=1}^n |\phi_\alpha(\mathbf{x}_i) - y_i|. \quad (10)$$

However, the main flaw of this approach is that a small risk or error on the training set does not necessarily guarantee a small error on prediction if the number  $n$  of training data is small [26].

For a given probability of at least  $1 - p$ , the Vapnik bound for the errors can be written as

$$E(\alpha) \leq R_p(\alpha) + \Psi\left(\frac{h}{n}, \frac{\log(p)}{n}\right), \tag{11}$$

where

$$\Psi\left(\frac{h}{n}, \frac{\log(p)}{n}\right) = \sqrt{\frac{1}{n} \left[ h \left( \log \frac{2n}{h} + 1 \right) - \log\left(\frac{p}{4}\right) \right]}. \tag{12}$$

Here  $h$  is a parameter, often referred to as the Vapnik-Chervonenskis dimension or simply VC-dimension [24], which describes the capacity for prediction of the function set  $\phi_\alpha$ .

In essence, a linear support vector machine is to construct two hyperplanes as far away as possible and no samples should be between these two planes. Mathematically, this is equivalent to two equations

$$\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = \pm 1, \tag{13}$$

and a main objective of constructing these two hyperplanes is to maximize the distance (between the two planes)

$$d = \frac{2}{\|\mathbf{w}\|}. \tag{14}$$

Such maximization of  $d$  is equivalent to the minimization of  $\|w\|$  or more conveniently  $\|w\|^2$ . From the optimization point of view, the maximization of margins can be written as

$$\text{minimize } \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} (\mathbf{w} \cdot \mathbf{w}). \tag{15}$$

This essentially becomes an optimization problem

$$\text{minimize } \Psi = \frac{1}{2} \|\mathbf{w}\|^2 + \lambda \sum_{i=1}^n \eta_i, \tag{16}$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) \geq 1 - \eta_i, \tag{17}$$

$$\eta_i \geq 0, \quad (i = 1, 2, \dots, n), \tag{18}$$

where  $\lambda > 0$  is a parameter to be chosen appropriately. Here, the term  $\sum_{i=1}^n \eta_i$  is essentially a measure of the upper bound of the number of misclassifications on the training data.

### 3.2 Nonlinear SVM and Kernel Tricks

The so-called kernel trick is an important technique, transforming data dimensions while simplifying computation. By using Lagrange multipliers  $\alpha_i \geq 0$ , we

can rewrite the above constrained optimization into an unconstrained version, and we have

$$L = \frac{1}{2} \|\mathbf{w}\|^2 + \lambda \sum_{i=1}^n \eta_i - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) - (1 - \eta_i)]. \quad (19)$$

From this, we can write the Karush-Kuhn-Tucker conditions

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = 0, \quad (20)$$

$$\frac{\partial L}{\partial \mathbf{b}} = - \sum_{i=1}^n \alpha_i y_i = 0, \quad (21)$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) - (1 - \eta_i) \geq 0, \quad (22)$$

$$\alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) - (1 - \eta_i)] = 0, \quad (i = 1, 2, \dots, n), \quad (23)$$

$$\alpha_i \geq 0, \quad \eta_i \geq 0, \quad (i = 1, 2, \dots, n). \quad (24)$$

From the first KKT condition, we get

$$\mathbf{w} = \sum_{i=1}^n y_i \alpha_i \mathbf{x}_i. \quad (25)$$

It is worth pointing out here that only the nonzero  $\alpha_i$  contribute to overall solution. This comes from the KKT condition (23), which implies that when  $\alpha_i \neq 0$ , the inequality (17) must be satisfied exactly, while  $\alpha_i = 0$  means the inequality is automatically met. In this latter case,  $\eta_i = 0$ . Therefore, only the corresponding training data  $(\mathbf{x}_i, y_i)$  with  $\alpha_i > 0$  can contribute to the solution, and thus such  $\mathbf{x}_i$  form the support vectors (hence, the name support vector machine). All the other data with  $\alpha_i = 0$  become irrelevant.

It has been shown that the solution for  $\alpha_i$  can be found by solving the following quadratic programming [24,26]

$$\text{maximize } \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j), \quad (26)$$

subject to

$$\sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \lambda, \quad (i = 1, 2, \dots, n). \quad (27)$$

From the coefficients  $\alpha_i$ , we can write the final classification or decision function as

$$f(\mathbf{x}) = \text{sgn} \left[ \sum_{i=1}^n \alpha_i y_i (\mathbf{x} \cdot \mathbf{x}_i) + \mathbf{b} \right], \quad (28)$$

where  $\text{sgn}$  is the classic sign function.

As most problems are nonlinear in business applications, and the above linear SVM cannot be used. Ideally, we should find some nonlinear transformation  $\phi$  so that the data can be mapped onto a high-dimensional space where the classification becomes linear. The transformation should be chosen in a certain way so that their dot product leads to a kernel-style function  $K(\mathbf{x}, \mathbf{x}_i) = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i)$ . In fact, we do not need to know such transformations, we can directly use the kernel functions  $K(\mathbf{x}, \mathbf{x}_i)$  to complete this task. This is the so-called kernel function trick. Now the main task is to choose a suitable kernel function for a given, specific problem.

For most problems in nonlinear support vector machines, we can use  $K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i)^d$  for polynomial classifiers,  $K(\mathbf{x}, \mathbf{x}_i) = \tanh[k(\mathbf{x} \cdot \mathbf{x}_i) + \Theta]$  for neural networks, and by far the most widely used kernel is the Gaussian radial basis function (RBF)

$$K(\mathbf{x}, \mathbf{x}_i) = \exp \left[ -\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{(2\sigma^2)} \right] = \exp \left[ -\gamma \|\mathbf{x} - \mathbf{x}_i\|^2 \right], \quad (29)$$

for the nonlinear classifiers. This kernel can easily be extended to any high dimensions. Here  $\sigma^2$  is the variance and  $\gamma = 1/2\sigma^2$  is a constant. In general, a simple bound of  $0 < \gamma \leq C$  is used, and here  $C$  is a constant.

Following the similar procedure as discussed earlier for linear SVM, we can obtain the coefficients  $\alpha_i$  by solving the following optimization problem

$$\text{maximize } \sum_{i=1}^n \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad (30)$$

It is worth pointing out under Mercer's conditions for the kernel function, the matrix  $\mathbf{A} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$  is a symmetric positive definite matrix [26], which implies that the above maximization is a quadratic programming problem, and can thus be solved efficiently by standard QP techniques [21].

## 4 Metaheuristic Support Vector Machine with APSO

### 4.1 Metaheuristics

There are many metaheuristic algorithms for optimization and most these algorithms are inspired by nature [27]. Metaheuristic algorithms such as genetic algorithms and simulated annealing are widely used, almost routinely, in many applications, while relatively new algorithms such as particle swarm optimization [7], firefly algorithm and cuckoo search are becoming more and more popular [27,32]. Hybridization of these algorithms with existing algorithms are also emerging.

The advantage of such a combination is to use a balanced tradeoff between global search which is often slow and a fast local search. Such a balance is important, as highlighted by the analysis by Blum and Roli [1]. Another advantage of this method is that we can use any algorithms we like at different stages of

the search or even at different stage of iterations. This makes it easy to combine the advantages of various algorithms so as to produce better results.

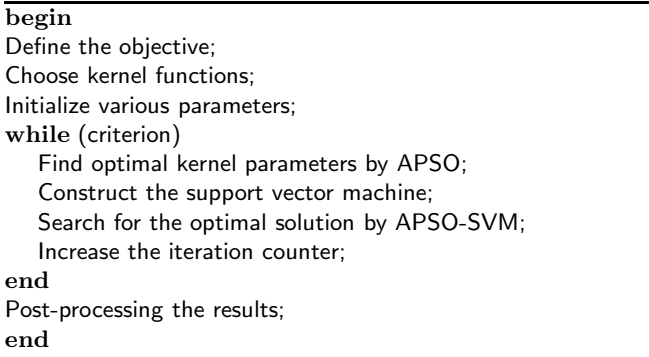
Others have attempted to carry out parameter optimization associated with neural networks and SVM. For example, Liu et al. have used SVM optimized by PSO for tax forecasting [13]. Lu et al. proposed a model for finding optimal parameters in SVM by PSO optimization [14]. However, here we intend to propose a generic framework for combining efficient APSO with SVM, which can be extended to other algorithms such as firefly algorithm [28,31].

## 4.2 APSO-SVM

Support vector machine has a major advantage, that is, it is less likely to overfit, compared with other methods such as regression and neural networks. In addition, efficient quadratic programming can be used for training support vector machines. However, when there is noise in the data, such algorithms are not quite suitable. In this case, the learning or training to estimate the parameters in the SVM becomes difficult or inefficient.

Another issue is that the choice of the values of kernel parameters  $C$  and  $\sigma^2$  in the kernel functions; however, there is no agreed guideline on how to choose them, though the choice of their values should make the SVM as efficiently as possible. This itself is essentially an optimization problem.

Taking this idea further, we first use an educated guess set of values and use the metaheuristic algorithms such as accelerated PSO or cuckoo search to find the best kernel parameters such as  $C$  and  $\sigma^2$  [27,29]. Then, we used these parameters to construct the support vector machines which are then used for solving the problem of interest. During the iterations and optimization, we can also modify kernel parameters and evolve the SVM accordingly. This framework can be called a metaheuristic support vector machine. Schematically, this Accelerated PSO-SVM can be represented as shown in Fig. 1.



**Fig. 1.** Metaheuristic APSO-SVM.

For the optimization of parameters and business applications discussed below, APSO is used for both local and global search [27][32].

## 5 Business Optimization Benchmarks

Using the framework discussed earlier, we can easily implement it in any programming language, though we have implemented using Matlab. We have validated our implementation using the standard test functions, which confirms the correctness of the implementation. Now we apply it to carry out case studies with known analytical solution or the known optimal solutions. The Cobb-Douglas production optimization has an analytical solution which can be used for comparison, while the second case is a standard benchmark in resource-constrained project scheduling [11].

### 5.1 Production Optimization

Let us first use the proposed approach to study the classical Cobb-Douglas production optimization. For a production of a series of products and the labour costs, the utility function can be written

$$q = \prod_{j=1}^n u_j^{\alpha_j} = u_1^{\alpha_1} u_2^{\alpha_2} \cdots u_n^{\alpha_n}, \quad (31)$$

where all exponents  $\alpha_j$  are non-negative, satisfying

$$\sum_{j=1}^n \alpha_j = 1. \quad (32)$$

The optimization is the minimization of the utility

$$\text{minimize } q \quad (33)$$

$$\text{subject to } \sum_{j=1}^n w_j u_j = K, \quad (34)$$

where  $w_j (j = 1, 2, \dots, n)$  are known weights.

This problem can be solved using the Lagrange multiplier method as an unconstrained problem

$$\psi = \prod_{j=1}^n u_j^{\alpha_j} + \lambda \left( \sum_{j=1}^n w_j u_j - K \right), \quad (35)$$

whose optimality conditions are

$$\frac{\partial \psi}{\partial u_j} = \alpha_j u_j^{-1} \prod_{j=1}^n u_j^{\alpha_j} + \lambda w_j = 0, \quad (j = 1, 2, \dots, n), \quad (36)$$

$$\frac{\partial \psi}{\partial \lambda} = \sum_{j=1}^n w_j u_j - K = 0. \quad (37)$$

The solutions are

$$u_1 = \frac{K}{w_1[1 + \frac{1}{\alpha_1} \sum_{j=2}^n \alpha_j]}, \quad u_j = \frac{w_1 \alpha_j}{w_j \alpha_1} u_1, \quad (38)$$

where ( $j = 2, 3, \dots, n$ ). For example, in a special case of  $n = 2$ ,  $\alpha_1 = 2/3$ ,  $\alpha_2 = 1/3$ ,  $w_1 = 5$ ,  $w_2 = 2$  and  $K = 300$ , we have

$$u_1 = \frac{Q}{w_1(1 + \alpha_2/\alpha_1)} = 40, \quad u_2 = \frac{K\alpha_2}{w_2\alpha_1(1 + \alpha_2/\alpha_1)} = 50.$$

As most real-world problem has some uncertainty, we can now add some noise to the above problem. For simplicity, we just modify the constraint as

$$\sum_{j=1}^n w_j u_j = K(1 + \beta\epsilon), \quad (39)$$

where  $\epsilon$  is a random number drawn from a Gaussian distribution with a zero mean and a unity variance, and  $0 \leq \beta \ll 1$  is a small positive number.

We now solve this problem as an optimization problem by the proposed APSO-SVM. In the case of  $\beta = 0.01$ , the results have been summarized in Table 1 where the values are provided with different problem size  $n$  with different numbers of iterations. We can see that the results converge at the optimal solution very quickly.

**Table 1.** Mean deviations from the optimal solutions

size $n$	Iterations	deviations
10	1000	0.014
20	5000	0.037
50	5000	0.040
50	15000	0.009

## 6 Income Prediction

Studies to improve the accuracy of classifications are extensive. For example, Kohavi proposed a decision-tree hybrid in 1996 [10]. Furthermore, an efficient training algorithm for support vector machines was proposed by Platt in 1998 [17,18], and it has some significant impact on machine learning, regression and data mining.

A well-known benchmark for classification and regression is the income prediction using the data sets from a selected 14 attributes of a household from a census

form [10,17]. We use the same data sets at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult for this case study. There are 32561 samples in the training-set with 16281 for testing. The aim is to predict if an individual’s income is above or below 50K ?

Among the 14 attributes, a subset can be selected, and a subset such as age, education level, occupation, gender and working hours are commonly used.

Using the proposed APSO-SVM and choosing the limit value of  $C$  as 1.25, the best error of 17.23% is obtained (see Table 2), which is comparable with most accurate predictions reported in [10,17].

**Table 2.** Income prediction using APSO-SVM

Train set (size)	Prediction set	Errors (%)
512	256	24.9
1024	256	20.4
16400	8200	17.23

### 6.1 Project Scheduling

Scheduling is an important class of discrete optimization with a wider range of applications in business intelligence. For resource-constrained project scheduling problems, there exists a standard benchmark library by Kolisch and Sprecher [11,12]. The basic model consists of  $J$  activities/tasks, and some activities cannot start before all its predecessors  $h$  are completed. In addition, each activity  $j = 1, 2, \dots, J$  can be carried out, without interruption, in one of the  $M_j$  modes, and performing any activity  $j$  in any chosen mode  $m$  takes  $d_{jm}$  periods, which is supported by a set of renewable resource  $R$  and non-renewable resources  $N$ . The project’s makespan or upper bound is  $T$ , and the overall capacity of non-renewable resources is  $K_r^\nu$  where  $r \in N$ . For an activity  $j$  scheduled in mode  $m$ , it uses  $k_{jmr}^\rho$  units of renewable resources and  $k_{jmr}^\nu$  units of non-renewable resources in period  $t = 1, 2, \dots, T$ .

For activity  $j$ , the shortest duration is fit into the time windows  $[EF_j, LF_j]$  where  $EF_j$  is the earliest finish times, and  $LF_j$  is the latest finish times. Mathematically, this model can be written as [11]

$$\text{Minimize } \Psi(\mathbf{x}) \sum_{m=1}^{M_j} \sum_{t=EF_j}^{LF_j} t \cdot x_{jmt}, \tag{40}$$

subject to

$$\sum_{m=1}^{M_h} \sum_{t=EF_j}^{LF_j} t x_{hmt} \leq \sum_{m=1}^{M_j} \sum_{t=EF_j}^{LF_j} (t - d_{jm}) x_{jmt}, (j = 2, \dots, J),$$

$$\sum_{j=1}^J \sum_{m=1}^{M_j} k_{jmr}^\rho \sum_{q=\max\{t, EF_j\}}^{\min\{t+d_{jm}-1, LF_j\}} x_{jmq} \leq K_r^\rho, (r \in R),$$



$$\sum_{j=1}^J \sum_{m=1}^{M_j} k_{jmr}^{\nu} \sum_{t=EF_j}^{LF_j} x_{jmt} \leq K_r^{\nu}, (r \in N), \quad (41)$$

and

$$\sum_{j=1}^{M_j} \sum t = EF_j^{LF_j} = 1, \quad j = 1, 2, \dots, J, \quad (42)$$

where  $x_{jmt} \in \{0, 1\}$  and  $t = 1, \dots, T$ . As  $x_{jmt}$  only takes two values 0 or 1, this problem can be considered as a classification problem, and metaheuristic support vector machine can be applied naturally.

**Table 3.** Kernel parameters used in SVM

Number of iterations	SVM kernel parameters
1000	$C = 149.2, \sigma^2 = 67.9$
5000	$C = 127.9, \sigma^2 = 64.0$

Using the online benchmark library [12], we have solved this type of problem with  $J = 30$  activities (the standard test set j30). The run time on a modern desktop computer is about 2.2 seconds for  $N = 1000$  iterations to 15.4 seconds for  $N = 5000$  iterations. We have run the simulations for 50 times so as to obtain meaningful statistics.

The optimal kernel parameters found for the support vector machines are listed in Table 3, while the deviations from the known best solution are given in Table 4 where the results by other methods are also compared.

**Table 4.** Mean deviations from the optimal solution ( $J=30$ )

Algorithm	Authors	$N = 1000$ 5000	
PSO [22]	Kemmoe et al. (2007)	0.26	0.21
hybrid GA [23]	Valls eta al. (2007)	0.27	0.06
Tabu search [15]	Nonobe & Ibaraki (2002)	0.46	0.16
Adapting GA [4]	Hartmann (2002)	0.38	0.22
<b>Meta APSO-SVM</b>	this paper	<b>0.19</b>	<b>0.025</b>

From these tables, we can see that the proposed metaheuristic support vector machine starts very well, and results are comparable with those by other methods such as hybrid genetic algorithm. In addition, it converges more quickly, as the number of iterations increases. With the same amount of function evaluations involved, much better results are obtained, which implies that APSO is very efficient, and subsequently the APSO-SVM is also efficient in this context. In addition, this also suggests that this proposed framework is appropriate for automatically choosing the right parameters for SVM and solving nonlinear optimization problems.

## 7 Conclusions

Both PSO and support vector machines are now widely used as optimization techniques in business intelligence. They can also be used for data mining to extract useful information efficiently. SVM can also be considered as an optimization technique in many applications including business optimization. When there is noise in data, some averaging or reformulation may lead to better performance. In addition, metaheuristic algorithms can be used to find the optimal kernel parameters for a support vector machine and also to search for the optimal solutions. We have used three very different case studies to demonstrate such a metaheuristic SVM framework works.

Automatic parameter tuning and efficiency improvement will be an important topic for further research. It can be expected that this framework can be used for other applications. Furthermore, APSO can also be used to combine with other algorithms such as neural networks to produce more efficient algorithms [13,14]. More studies in this area are highly needed.

## References

1. Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.* 35, 268–308 (2003)
2. Chatterjee, A., Siarry, P.: Nonlinear inertia variation for dynamic adaptation in particle swarm optimization. *Comp. Oper. Research* 33, 859–871 (2006)
3. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evolutionary Computation* 6, 58–73 (2002)
4. Hartmann, S.: A self-adapting genetic algorithm for project scheduling under resource constraints. *Naval Res. Log.* 49, 433–448 (2002)
5. Howley, T., Madden, M.G.: The genetic kernel support vector machine: description and evaluation. *Artificial Intelligence Review* 24, 379–395 (2005)
6. Goldberg, D.E.: *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison Wesley, Reading (1989)
7. Kennedy, J., Eberhart, R.C.: Particle swarm optimization, in: *Proc. of IEEE International Conference on Neural Networks*, Piscataway, NJ, pp. 1942–1948 (1995)
8. Kennedy, J., Eberhart, R.C.: *Swarm intelligence*. Academic Press, London (2001)
9. Kim, K.: Financial forecasting using support vector machines. *Neurocomputing* 55, 307–319 (2003)
10. Kohavi, R.: Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In: *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 202–207. AAAI Press, Menlo Park (1996), <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult>
11. Kolisch, R., Sprecher, A.: PSPLIB - a project scheduling problem library, OR Software-ORSEP (operations research software exchange program) by H. W. Hamacher. *Euro. J. Oper. Res.* 96, 205–216 (1996)
12. Kolisch, R., Sprecher, A.: The Library PSPLIB, <http://129.187.106.231/psplib/>

13. Liu, L.-X., Zhuang, Y., Liu, X.Y.: Tax forecasting theory and model based on SVM optimized by PSO. *Expert Systems with Applications* 38, 116–120 (2011)
14. Lu, N., Zhou, J.Z., He, Y.: Y., Liu Y., Particle Swarm Optimization for Parameter Optimization of Support Vector Machine Model. In: 2009 Second International Conference on Intelligent Computation Technology and Automation, pp. 283–284. IEEE publications, Los Alamitos (2009)
15. Nonobe, K., Ibaraki, T.: Formulation and tabu search algorithm for the resource constrained project scheduling problem (RCPSP). In: Ribeiro, C.C., Hansen, P. (eds.) *Essays and Surveys in Metaheuristics*, pp. 557–588 (2002)
16. Pai, P.F., Hong, W.C.: Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electric Power Sys. Res.* 74, 417–425 (2005)
17. Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines, Technical report MSR-TR-98014, Microsoft Research (1998)
18. Plate, J.C.: Fast training of support vector machines using sequential minimal optimization, in: In: Scholkopf, B., Burges, C.J., Smola, A.J. (eds.) *Advances in Kernel Methods – Support Vector Learning*, pp. 185–208. MIT Press, Cambridge (1999)
19. Shi, G.R.: The use of support vector machine for oil and gas identification in low-porosity and low-permeability reservoirs. *Int. J. Mathematical Modelling and Numerical Optimisation* 1, 75–87 (2009)
20. Shi, G.R., Yang, X.-S.: Optimization and data mining for fracture prediction in geosciences. *Procedia Computer Science* 1, 1353–1360 (2010)
21. Smola, A. J., Schölkopf, B.: A tutorial on support vector regression, (1998), <http://www.svms.org/regression/>
22. Tchomt e, S.K., Gourgand, M., Quilliot, A.: Solving resource-constrained project scheduling problem with particle swarm optimization. In: *Proceeding of 3rd Multidisciplinary Int. Scheduling Conference (MISTA 2007)*, Paris, August 28 - 31, pp. 251–258 (2007)
23. Valls, V., Ballestin, F., Quintanilla, S.: A hybrid genetic algorithm for the resource-constrained project scheduling problem. *Euro. J. Oper. Res* (2007), doi:10.1016/j.ejor.2006.12.033
24. Vapnik, V.: *Estimation of Dependences Based on Empirical Data*. Springer, New York (1982) (in Russian)
25. Vapnik, V.: *The nature of Statistical Learning Theory*. Springer, New York (1995)
26. Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V.: Comparing support vector machine with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Processing* 45, 2758–2765 (1997)
27. Yang, X.S.: *Nature-Inspired Metaheuristic Algorithms*. Luniver Press (2008)
28. Yang, X.-S.: Firefly algorithms for multimodal optimization. In: Watanabe, O., Zeugmann, T. (eds.) *SAGA 2009*. LNCS, vol. 5792, pp. 169–178. Springer, Heidelberg (2009)
29. Yang, X.-S., Deb, S.: Cuckoo search via L evy flights, in: *Proceedings of World Congress on Nature & Biologically Inspired Computing, NaBIC 2009*, pp. 210–214. IEEE Publications, USA (2009)
30. Yang, X.S., Deb, S.: Engineering optimization by cuckoo search. *Int. J. Mathematical Modelling and Numerical Optimisation* 1, 330–343 (2010)
31. Yang, X.S.: Firefly algorithm, stochastic test functions and design optimisation. *Int. J. Bio-inspired Computation* 2, 78–84 (2010)
32. Yang, X.S.: *Engineering Optimization: An Introduction with Metaheuristic Applications*. John Wiley & Sons, Chichester (2010)

# A Study on the Reliability of Data Transmission of an over the Top Network Protocol on SMS versus UDP/ GPRS (Fragmented)


Alyssa Marie Dykimching, Jan Aaron Angelo Lee, and William Emmanuel Yu

Department of Information Systems and Computer Science,  
Ateneo de Manila University,  
Katipunan Avenue, Loyola Heights,  
Quezon City 1108, Philippines

**Abstract.** The purpose of this study was to identify an aspect of data transmission that could help address the existing gap between the availability of the second and third generations of mobile telephony in the light of increased demand for growth of machine-to-machine applications using mobile data technology. This study examined the reliability of sending data through the SMS interface as an alternative means to transmit data over newer technologies such as GPRS. The research primarily focused on transport performance, and looked into different factors of integrity and reliability, such as transmission time and number of retries. The study was then divided into construction, testing and analysis phases wherein the output was an analysis of the data gathered from testing packet data transmission via SMS and GPRS in key locations around a specific locale. As per the results, SMS in general works in a slow but steady pace of delivery, while GPRS (using the UDP protocol) offers a quicker but slipshod transmission.

**Keywords:** Network reliability, SMS, GPRS, Packet transmission, Network performance.

## 1 Introduction

The rise of the Internet culture along with mobile phones has called for a converging of the two technologies. The availability of the diverse and abundant information the Internet provides, and the accessibility of mobile phones with more than a billion users worldwide, has pushed the advancement of mobile technology. As of the GSM wave of mobile technology, mobile phones used radio signals to transmit and receive data reaching transfer speeds of around 115.2 kb/s. Further advancements of the said technology, including the 3rd and 4th recognized mobile technology generations, fly up to 11mb/s .

With the coming of these newer technologies, the better and more advanced 3G and eventually 4G were commonly thought to replace the already established 2G services. Business factors however, such as infrastructure cost and market demands, hinder 3G from replacing 2G anytime soon. In places like Nigeria, the

demand for higher-end wireless services contributes less than 1% of the total population [14]. With a target margin this small, wireless service carriers are forced to put higher premiums on 3G services and only place 3G capable towers in strategic locations. This kind of inconsistencies between 3G capable areas and 2G only capable areas lead to poor signal strength, connection difficulties, poor quality and low overall customer satisfaction.

While the use of wireless services is growing, with more than a billion SMSs sent everyday [7], it is expected that Philippine telecommunication carriers are better equipped to provide quality SMS services to its clients compared to more premium services such as GPRS. It is therefore more practical to take advantage of this and find innovative ways to broaden the spectrum that SMS currently addresses. The result will most certainly not make SMS replace the higher-end technologies, especially in terms of performance, but simply give a share of the demand to SMS as the market progresses and transitions to the next generation technologies.

### 1.1 Objectives

In order to provide a thorough study on the reliability of data transmission over SMS versus GPRS, this study aims to first design and construct a robust SMS and GPRS data transmission (sending and receiving) interface based on the network protocol proposed in [24]. This research also seeks to define a measure of network reliability and performance for data transmission with variables such as target locations around Metro Manila, SMS and GPRS/EDGE technologies, and data size. To quantify GPRS performance, the UDP protocol is used. Fragmented payload (UDP-F) is also used to allow for a like to like comparison since SMS can only hold smaller payloads. This allows for ease of conversion of existing machine-to-machine applications to use the same PDU size constraint and also makes it easier to compare use cases, wherein the system was originally built for SMS then modified to support GPRS/UDP at a later time (or vice versa). This study, however, does not attempt to compare the reliability of network providers and machines used; hence, such factors will be controlled throughout the study. Lastly, this paper aims to test two (2) over the top networks: SMS and UDP-fragmented (UDP-F).

### 1.2 Significance of the Study

More and more mobile machine-to-machine applications are being developed in the fields of health care, telemetry and navigation among others; and this growth area will definitely benefit from a more pervasive and reliable data transmission [24]. GSM technology worldwide is enjoying a wide coverage for its subscribers; however, the newer and more advanced 3rd and 4th generations' coverage area remain focused only where there is existing demand for such services. Although newer and more advanced mobile technologies exist and allow faster data transmission, these services reach only a small fraction of the user population. Therefore, the need for alternate and robust data transmission methods still exist for the Third World.

## 2 Review of Related Literature

### 2.1 Global System for Mobile Communications (GSM)

With the aid of a subscriber identity module (SIM) that carries the user's personal number, GSM systems have enhanced lives by providing mobility [17]. Since the start of the GSM network operations, a variety of value-added services have been offered. This research will only focus on two of such services, the short message service (SMS) and general packet radio service (GPRS).

SMS enables users to send and/or receive alphanumeric messages up to 140 bytes in length. Messages are stored and transmitted via a short message service center (SMSC) [15], and could be sent as text or binary. General packet radio service (GPRS), on the other hand, aims to enlarge the 9.6 kb/s data transfer rate of GSM services to over 170 kb/s so as to enable multimedia communication [4]. It applies a packet radio principle to improve and simplify wireless access between mobile stations and packet data networks. It then offers data connections with higher transfer rates and shorter access times [3].

In the context of this study, the GPRS mechanism shall use the User Datagram Protocol (UDP) as specified in the Internet Protocol (IP) for data transmission. The reason for this is that GPRS transmission will be done using traditional sockets as usually used in computer networking further specified in the Methodology portion of this paper. Therefore the use of the UDP protocol is more suited in this scenario since the said protocol does not have implicit handshaking mechanisms, which allows flow-control to be handled at the Application Layer of the Open Systems Interconnection Model (OSI Model) for several logging and documentation functions specific to this experiment, as compared to other IP protocols such as TCP/IP, which has functions that handle this at the Transport Layer.

### 2.2 Data Transmission through SMS

Although more advanced technologies, such as GPRS, allow faster data transmission, it is not always reliable in developing countries, where SMS is predominantly used. It is therefore important to look into the feasibility of using SMS in transporting different types of data. Its nationwide coverage, affordability, security, and compatibility of with future technology upgrades, are some of the many factors why the GSM-SMS network has been deemed ideal for information transmission [1] [24].

**Medical.** SMS is widely used in the medical field since it is not only instantaneous, accessible and mobile, but also cost effective. Computer and online access is not needed, and given that physicians have to make rapid decisions, SMS provide the brevity and immediacy they need. Urgent medical advice could now then be sought. In Italy, an SMS was sent to a surgeon stating that a 22 year-old male arrives, coughing blood. A reply was received a minute later advising an angiogram to identify aortic rupture [22]. Similarly, SMS has been also used in collecting diary data from patients for monitoring and self-management of

asthma [2], sending reminders and results to patients, as well as collecting data from outpatients to evaluate overall patient satisfaction. Advantages include convenience, immediate results, reduced costs, and reduced paper use. Confidentiality and the inability to convey a large amount of data, however, are some concerns [20].

**Alerts and Other Services.** Moreover, SMS could be used to enable access to network equipment using a mobile phone. An application of such is a network alarm monitoring system wherein alarm messages generated by a network management program are forwarded to mobile phones of members of the network operations center (NOC) using a simple mail transfer protocol (SMTP) to SMS gateway [23]. Wireless value-added data services based on SMS, such as sending and forwarding emails, and accessing train schedules, could also be deployed. Users could query local schedules via SMS based on train class, time, and target destination. The said schedule service may also be used for other ticket reservation services like the ones for airlines [18].

Likewise, *MobileDeck*, an engaging front-end environment that used SMS as the main communications channel, integrated a graphical user interface (GUI) with a server that feeds the appropriate content through instructions contained in binary SMS. Due to the relatively low cost and accessibility of SMS, *MobileDeck* has proven to be a suitable alternative to services running on top of other mobile data technologies such as GPRS for data transfer and content distribution [19].

SMS has been perceived to continue leading the peer-to-peer mobile communication data service, as it is the most accessible, easiest to use, and not to mention the cheapest. Other applications and data services could also be introduced through SMS, such as personal instant messaging and SMS-based money transfer [6]. However, due to its limited size, using SMS for data transmission will require the division of the total data to be sent into one or more packets with a ratio of 1:1 packets per SMS. In order to fulfill this, a header is used to specify the sequence of the segments [5].

### 2.3 Other Technologies

In this study, the group will use mobile broadband dongles as modems to send and receive data using over the top networks, particularly SMS and UDP-F. Once a serial connection to the dongle has been opened, data could now be transmitted using AT commands, also known as Hayes commands, and the open-source Py-Serial module in Python [16]. Checksums, particularly cyclic redundancy checks (CRCs), are then generated to determine data integrity. These are error detection mechanisms created by summing up all the bytes in a data to create a checksum value, which is appended to the message payload and transmitted with it. If the computed checksum of the data received matches the received checksum value, then there is a high probability that there was no transmission error [13]. The checksum function used in this research is the built-in `crc32` function in Python's `zlib` module, which is calculated using the following formula:

$$x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1. \quad (1)$$

## 2.4 Network Protocol Security and Efficiency

The study uses a simple authentication protocol between the transmitting and receiving ends, which are held as constants, to ensure that ends are operational during a test. It is therefore less secure, as it can easily be mimicked to disrupt proper transmissions, than the traditional Session Initiation Protocol (SIP) that the Internet uses, where it caters to multiple users at the same time, but is nevertheless sufficient for this study's purposes. A recent study seeks to eliminate at least one of the security issues experienced over the Internet when using SIP which can also be applied to a broader implementation of this study [11].

The concern of efficiency also comes to question once a more comprehensive and possibly decentralized application is in view. Another study highlights the comparison of proactive, reactive, and hybrid routing protocols for mobile ad-hoc networks that allows optimal traversing in a decentralized structure [25].

## 2.5 Network Reliability

Reliability is defined as the probability of a tool to perform successfully over a given time interval under specified conditions [12]. It is hence a number between zero and one, or 0% and 100% [9], which may be estimated by:

$$R = \frac{N_t - N_f}{N_t} \quad (2)$$

where  $R$  = estimated reliability;

$N_t$  = total number of trials; and

$N_f$  = number of trials resulting in failure [10].

$R$  is only an estimate due to the finite number of trials taken. In general, a tool is reliable if it yields consistent results when measured across various contexts [21]. The estimated reliability approaches true reliability as the number of trials approach infinity [9].

## 3 Framework

This study will be structured around the network protocol proposed in [24]. The said protocol has two legs: transmitter and receiver. In this research, however, the application will not terminate until all data has been successfully transmitted. After sending each packet, the transmitter application will listen for feedback, and will only continue or terminate once an acknowledgment packet indicating good transmission has been received, or otherwise called the automatic repeat request (ARQ) mechanism. This only happens if the packet received indicates faulty transmission or when the application after a certain amount of time still does not receive anything (a timeout). The process flow for the transmitting and receiving leg of the application is shown in Figures 1 and 2, respectively.



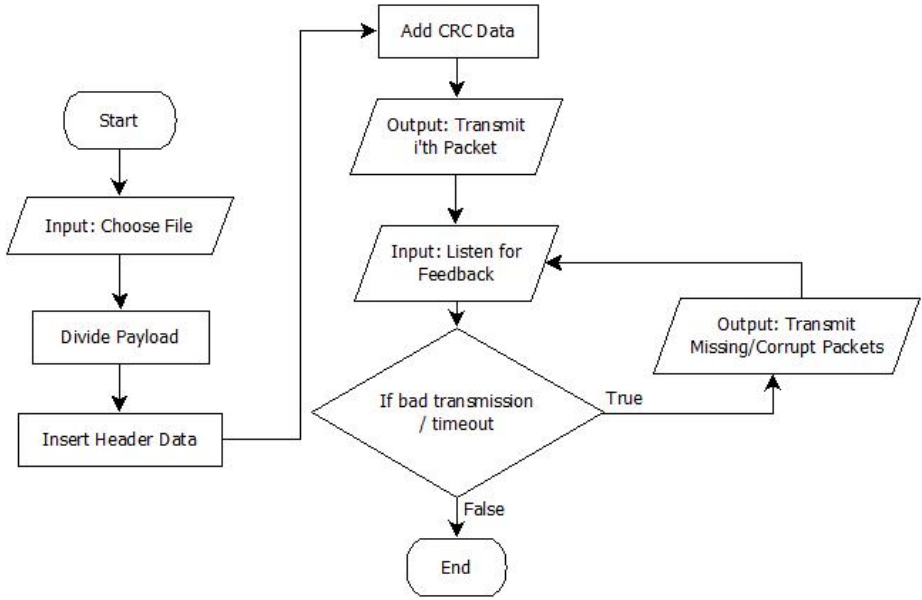


Fig. 1. Transmitter Protocol Flow

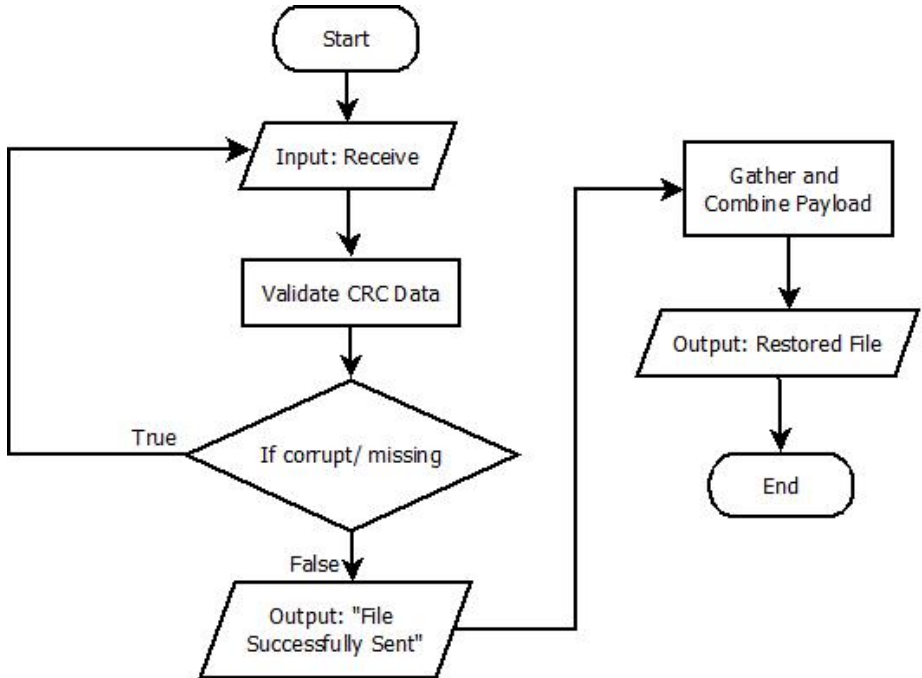


Fig. 2. Receiver Protocol Flow

Transmitting data larger than what a single SMS can conventionally hold will require the data to be split and sent over numerous SMSs depending on the size of the data. For better comparison, the GPRS mechanism will also be structured similarly. The data will be divided into several packets and will form the payload data of every packet. The header data of every packet on the other hand, will basically contain necessary information about the respective packet such as the packet's queue number, the total number of packets to be sent and other necessary data supporting file integrity. The number of packets will then be influenced by variables such as data size other functional schemes like for instance file integrity checking, encryption and compression schemes.

The receiving leg will more or less be the reverse of the transmitting leg of the protocol. The receiving device will wait until it collects all of the packets, check them for file integrity, request for resend of packets if necessary, and finally stitch the data back into its original form. If in the event that the application is waiting for a certain packet and does not receive it in a given amount of time, the receiving application will be deemed to have timed out. It will then stop the current test and proceed with the next one or terminate the application depending on user input.

## 4 Method

### 4.1 Interface

The transmitter and receiver applications were coded in Python 2.7 [16]. The checksum function used is the built-in `crc32` function in Python's `zlib` module. The applications were used to send and receive data over SMS and GPRS, in which UDP-based transport was used. To be able to better compare the reliability of data transmission over the two networks, SMS-based and UDP-fragmented mechanisms were created. Each sending mechanism has its own configuration file that is read upon running the application. It contains specific information necessary for each mechanism to transmit data such as the destination number, SMS Center number and COM port number for SMS, the destination IP address and port numbers for UDP, and the file to be sent. It also allows for multiple tests to be done in a single application run.

### 4.2 SMS-Based Mechanism (SMS)

The SMS-based interface utilizes AT Commands to communicate with the serial devices on a low level scheme. These are sent and received through the built-in serial module of Python as they normally are through other telnet applications such as PuTTY or Windows HyperTerminal. They are used for a number of actions: (1) set the device to PDU-mode, (2) send and receive messages, and (3) check for signal strength.

**SMS sending mechanism.** After reading the configuration file, the application will find the file specified, split and convert it into streams of hexadecimal octets and assign them in an array where each element represents an individual packet

to be sent. The checksum value of each packet is also appended at the beginning of the said packet. Next, it will prepare header data containing the destination number and SMS Center numbers necessary for sending SMS in binary mode.

Data transmission will begin with an invitation, which contains the filename and file size, to check whether the receiving application is ready. Once a confirmation is received, a single data packet will be sent and will be resent on timeout or upon request of the receiving application through an acknowledgement packet; otherwise the application will move on to the succeeding packet and repeat until all data has been transmitted. Once all data has been successfully transmitted, the application will either terminate or move on to the next test if any.

**SMS receiving mechanism.** Upon application start, it will repeatedly loop and wait for an invitation. Once an invitation has been received, it will allocate space for the file, send a confirmation back to the other application and wait for the actual data packets. When a data packet arrives, the header data and payload is extracted and is counterchecked with the calculated checksum. Acknowledgement indicating successful transmission or a request for retransmission is sent afterwards. The receiving application does not timeout and will always just wait to receive data until all packets have been successfully received. After the final acknowledgement is sent, the data is written to a file and the application will once again wait for invitations. The application will only terminate on user intervention.

### 4.3 UDP-Based Mechanism

Network connection is established by use of the given carrier application where it is always explicitly specified to set to GSM-only thus forcing the serial device to use the GPRS network. To validate if the connection has already been established properly, a connection check function is included before the actual packet data transfer. The connection set-up time, or the time it takes for a network connection to finally get established, will be recorded separately from the actual transmission time.

Traditional socket data transfer is then employed to send and receive data. Both applications of both the fragmented and whole versions of the UDP-based mechanisms make use of two sockets. Sending applications use one socket to send data and the other to listen for acknowledgements. Receiving applications use one socket to listen for invitations and data and use the other one to send acknowledgements.

**UDP-fragmented mechanism (UDP-F).** This particular mechanism has exactly the same structure as that of the SMS-based mechanism. This means that the data sent and received has both PDU specific header and payload data that is otherwise required in the SMS-based mechanism but is not at all required in this one. The applications in this mechanism, both transmitting and receiving, also have the exact same structure and flow as that of the SMS-based. This is so that the only difference between this mechanism and the SMS-based mechanism is the medium that the data uses to get from one application to the other.

#### 4.4 Test Instruments

Two files of different sizes, each making a test case, have been prepared. For simpler packet integrity validation, only images were used for the preliminary testing. Two images were sent, snowman.gif (388 bytes) and crab.gif (4713 bytes), respectively. There are three (3) packets for snowman.gif and thirty-seven (37) packets for crab.gif, giving a total of forty (40) packets per test sample. In this study, the transmitting end was situated at St. Charbel Executive Village, while the receiving end remained at New Manila Rolling Hills Village in Quezon City. For easier future reference, the test locations are briefly labeled in Table 1.

**Table 1.** Test Location Reference

Test Locations	Reference
St. Charbel Executive Village	CHARBEL
New Manila Rolling Hills Village	NEWMANILA

#### 4.5 Procedure

The mechanisms created (SMS and UDP-F) are composed of two ends: transmitter and receiver. The test sample (consisting of two images with a total of 40 packets) was transmitted thrice for both mechanisms, while the receiving end collected and waited for all of the sent packets to arrive. It automatically requested for a resending of the missing or corrupt packets, if the transmission was not successful. The time for every significant action, such as dividing or combining packets, and transmitting or receiving packets, was recorded. The number of times the data was resent, if applicable, was also noted.

## 5 Results

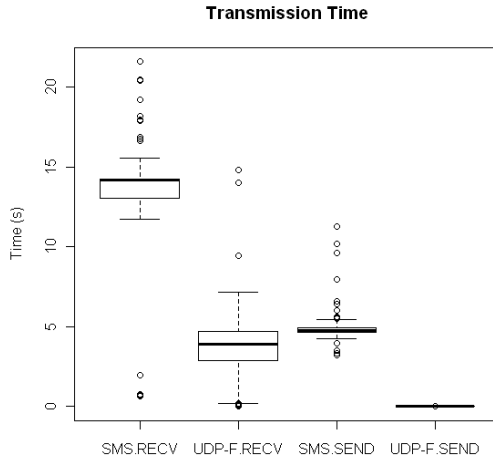
### 5.1 Transmission Time

SMS generally has a greater transmission time and distribution while UDP-F has more constant and smaller ones. In Figure 3, the plots for the sending interface of both SMS and UDP-F refers to how long it takes the application to successfully send a packet to the network while for the receiving applications, the time it takes for them to receive a packet on wait. For better comparison, the transmission time per packet for SMS is compared to GPRS (UDP-F), as presented in Table 2 and Figure 4.

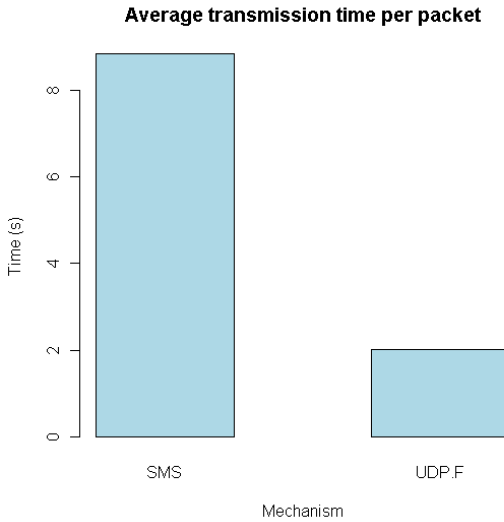
**Table 2.** Transmission time per packet

Mechanism	snowman.gif (388 bytes)	crab.gif (4713 bytes)
SMS	8.75 seconds	8.84 seconds
UDP-F	3.57 seconds	1.88 seconds

The total duration for each file transmission is the time starting from when the sender application begins to send an invitation up until confirmation of a



**Fig. 3.** Transmission time

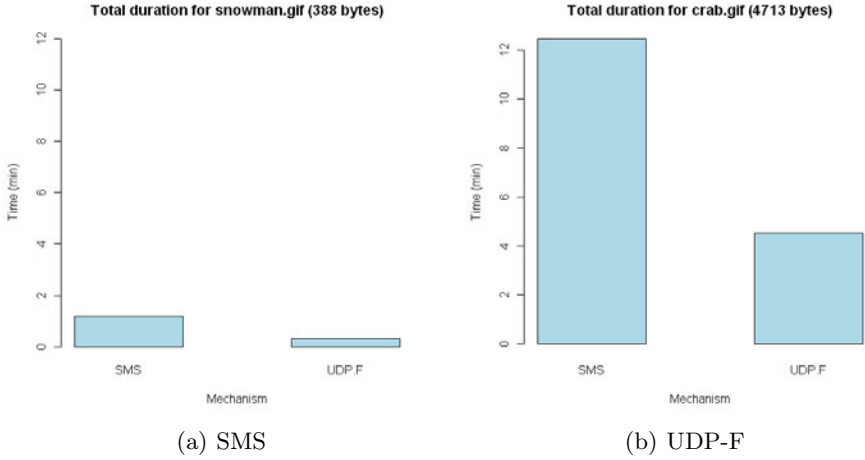


**Fig. 4.** Transmission time per packet

**Table 3.** Average total duration for each file transmission

Mechanism	snowman.gif (388 bytes)	crab.gif (4713 bytes)
SMS	1 minute and 12 seconds	12 minutes and 28 seconds
UDP-F	0 minutes and 19 seconds	4 minutes and 32 seconds

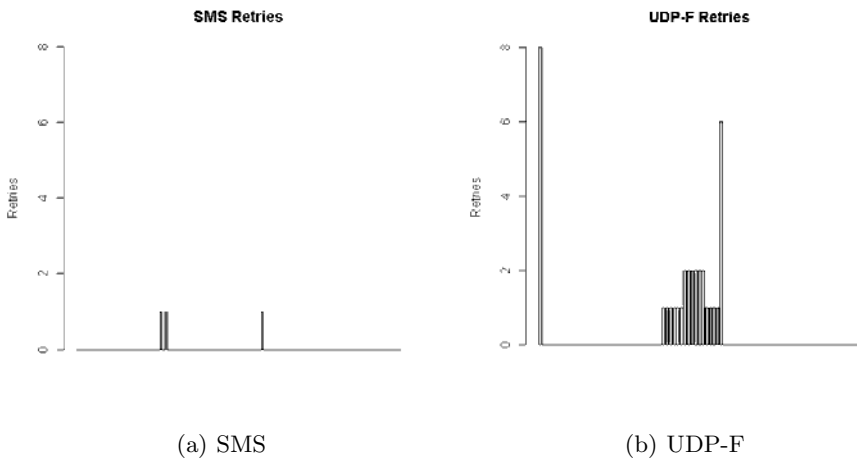
successful file transmission is received. The average values for SMS is 379% and 275%, for snowman.gif and crab.gif respectively, greater than that of UDP-F. These are shown in Table 3 as well as Figures 5(a) and 5(b).



**Fig. 5.** Average total duration for each file transmission

## 5.2 Number of Retries

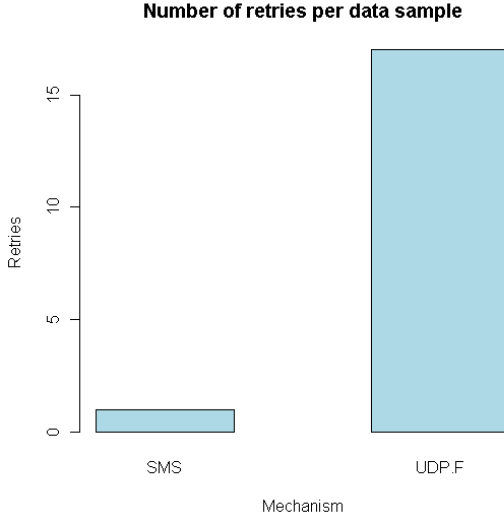
A retry is described as a transmission failure where the sending application has timed out and resent the previous packet and waits once more; that is to say, the sending application has already waited for a certain amount of time and has either received an invalid acknowledgement or none at all which then leads the application to fire a retry. SMS only reached a maximum of 1 retry, while UDP-F reached as much as 6 and 8 retries. Although data transmission over GPRS is faster, the initial test results also show the wide disparity in the number of retries when data was transmitted over SMS versus GPRS (UDP-F), as presented in Table 4 and Figure 7.



**Fig. 6.** Number of retries

**Table 4.** Average number of retries per data sample

Mechanism	snowman.gif (388 bytes)	crab.gif (4713 bytes)
SMS	0	1
UDP-F	3	14

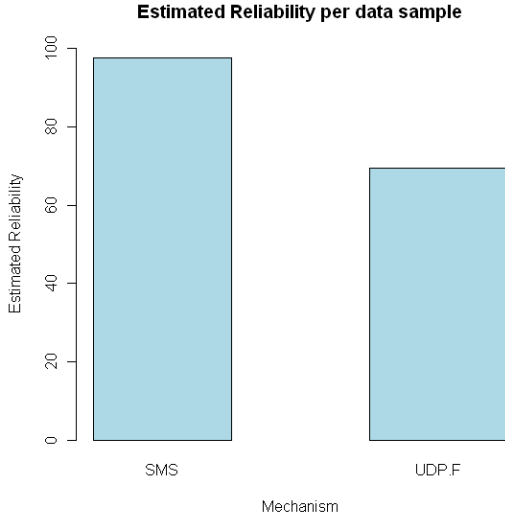
**Fig. 7.** Average number of retries per data sample

### 5.3 Reliability

As given by the reliability formula in [10],  $R = \frac{N_t - N_f}{N_t}$ , where  $R$  = estimated reliability;  $N_t$  = total number of trials; and  $N_f$  = number of trials resulting in failure, we could likewise see the wide disparity in the estimated reliability when data was transmitted over SMS versus GPRS (UDP-F) shown in Table 5 and Figure 8. For more accuracy, these are calculated using the actual number of retries and total packets of 6 tests (3 per file) per mechanism and not the average values used above.  $N_t$  has a base value of  $3 \times 40 = 120$  for both mechanisms where there are  $3 + 37 = 40$  packets, three packets for snowman.gif and thirty-seven packets for crab.gif, for 3 test cycles with no retries.

**Table 5.** Estimated Reliability per data sample

Mechanism	$N_t$	$N_f$	$R = \frac{N_t - N_f}{N_t}$
SMS	123	3	$0.976 = 97.6\%$
UDP-F	173	53	$0.694 = 69.4\%$



**Fig. 8.** Estimated Reliability per data sample

## 6 Conclusion

As per the results, SMS in general works in a slow but steady pace of delivery, while GPRS (using UDP as the protocol) offers a quicker but slipshod transmission. Using the data gathered in sending 40 packets, SMS, with an average transmission time (ATT) of 8.84 seconds per packet, will need 353.6 seconds best case with no retries, while UDP, with an effective ATT of 6.76 seconds<sup>1</sup>, will take 270.4 seconds already with one retry for each packet. SMS therefore can never replace UDP in sending data in large amounts because the time it will take UDP to transmit the correct data even with retries will still outrun that of SMS with perfect (no-retries) transmission. SMS however is generally the ideal choice in sending short and uncompromised data such as that of machine-to-machine transmissions.

## 7 Further Studies

This paper details a study on the comparison between SMS-fragmented and GPRS UDP-fragmented transmissions and its results. Further research into the topic will include data transmission through GPRS UDP-whole (UDP-W) and uniform tests across multiple locations. Adding such dimensions to the study will emphasize reliability in terms of retransmission, transmission time and geographical constraints.

<sup>1</sup> Given 1.88 seconds as the average transmission time for UDP and assuming 3 seconds as the timeout for the transmitting application, a complete *retry cycle*, which comprises of the time it takes for the first transmission, plus the timeout, and the retransmission, will result in an effective ATT of 6.76 seconds.



## References

1. EACOMM Corporation Embedded Systems Division. Wireless Data Transmission over GSM Short Message Service (2003)
2. Anhoj, J., Moldrup, C.: Feasibility of Collecting Diary Data from Asthma Patients through Mobile Phones and SMS. *Journal of Medical Internet Research* (2004)
3. Bettstetter, C., Vogel, H., Eberspacher, J.: GSM Phase 2+ General Packet Radio Service: Architecture, Protocols, and Air Interface. *IEEE Communications Surveys* (1999)
4. Brasche, G., Walke, B.: Concepts, Services, and Protocols of the New GSM Phase 2+ General Packet Radio Service. *IEEE Communications Magazine* (1997)
5. Brown, J., Shipman, B., Vetter, R.: SMS: The Short Message Service. *Computer* (2007)
6. Chau, F.: The Death of SMS - Not Quite Yet. *Wireless Asia* (2007)
7. Einstein, M.: The Philippines: Mobile Market Still Trying to Find its Voice. *Telecom Asia* (2009)
8. Govil, J.: 4G: Functionalities Development and an Analysis of Mobile Wireless Grid. *Emerging Trends in Engineering and Technology* (2008)
9. Kececioglu, D.: A Comprehensive Definition of Reliability. *Maintainability, Availability, and Operational Readiness Engineering Handbook* (2002)
10. Krohn, C.: Reliability Analysis Techniques. In: *Proceedings of the IRE* (1960)
11. Lee, C., Yang, C., Huang, S.: A New Authentication Scheme for Session Initiation Protocol. *Journal of Digital Information Management* 7, 133–136 (2009)
12. Lefebvre, M.: *Basic Probability Theory with Applications*, Springer Undergraduate Texts in Mathematics and Technology (2009)
13. Maxino, T., Koopman, P.: The Effectiveness of Checksums for Embedded Control Networks. *IEEE Transactions on Dependable and Secure Computing* (2009)
14. Olokede, S.: Untapped Capabilities of 2G in Nigeria Telecom Space. *Leonardo Journal of Sciences* (2009)
15. Peersman, G., Cvetkovic, S.: The Global System for Mobile Communications Short Message Service. *IEEE Personal Communications* (2000)
16. Python Programming Language, <http://www.python.org/>
17. Rahnema, M.: Overview of the GSM System and Protocol Architecture. *IEEE Communications Magazine* (1993)
18. Rao, H., Chang, D., Lin, Y.: iSMS: An Integration Platform for Short Message Service and IP Networks. *IEEE Network* (2001)
19. Risi, D., Tefilo, M.: Mobiledeck: Turning SMS into a Rich User Experience. In: *Proceedings of the 6th International Conference On Mobile Technology, Applications, and Systems* (2009)
20. Roberts, A., Gorman, A.: Short Message Service for Outpatient Data Collection. *British Journal of Anaesthesia* (2009)
21. Salkind, N.: Reliability Theory. *Encyclopedia of Measurement and Statistics* (2007)
22. Sherry, E., Colloridi, B., Warkne, P.: Short Message Service (SMS): A Useful Communication Tool for Surgeons. *ANZ Journal of Surgery* (2002)
23. Vougioukas, S., Roumeliotis, M.: A System for Basic-Level Network Fault Management based on the GSM Short Message Service (SMS). In: *Proceedings of International Conference on Trends in Communications* (2001)

24. Yu, W., Tagle, P.: Development of an Over-the-Top Network Protocol for Pervasive, Secure and Reliable Data Transmission over GSM Short Messaging Service. In: 2010 International Conference on Computer and Software Modeling, IACSIT, pp. 1–7 (2010)
25. Zahary, A., Ayesh, A.: A Comparative Study for Reactive and Proactive Routing Protocols in Mobile Ad hoc Networks. *Journal of Intelligent Computing* 1, 20–29 (2010)

# QuickFlood: An Efficient Search Algorithm for Unstructured Peer-to-Peer Networks

Hassan Barjini, Mohamed Othman\*, and Hamidah Ibrahim

Department of Computer Science And Information Technology, Universiti Putra  
Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia.

hassan.barjini@gmail.com, {mothman,hamidah}@fsktm.upm.edu.my

**Abstract.** File sharing is one of the most popular activities in peer-to-peer systems, and the main issue in file sharing is the efficient search for content locations. A search algorithm has to provide low search traffic, optimum search latency, more search coverage, and determinism in returning the results. However, existing search algorithms fail to satisfy all these goals. In this paper we propose, and analyze a novel Hybrid searching algorithm (QuickFlood), based on Flooding-Based searching algorithms. This algorithm combines flooding and teeming search algorithms to benefit both merits and to limit the drawbacks. We provide the analytical results for best threshold point to switch from flooding to teeming. Our analytical results are validated through simulation. The proposed algorithm can reduce up to 90% of redundant messages of the current flooding algorithm and increase more than 3 times of its searching success rates.

**Keywords:** Peer-to-peer, unstructured P2P network, flooding, searching.

## 1 Introduction

The most fundamental searching technique in the unstructured P2P system is flooding. It starts from a query originator by initiated Time to Live (TTL) value (e.g. TTL = 7), and propagates requested messages in hop-to-hop fashion counted by TTL count. TTL is decremented by one when the requested message travels across one hop. A message comes to end either TTL is 0, or because it becomes a redundant message [1]. Flooding has significant merits such as: the simple algorithm, large coverage, high reliability, and moderate latency. Despite these merits, it produces huge redundant messages, which seriously limit system scalability. [2] showed that more than 70% of the generated messages are redundant in flooding within TTL of 7. The most redundant messages are generated in high-hops, particularly in a system with high-connectivity topology.

The current study attempts to develop a new version of flooding based on the hierarchical structure. The proposed algorithm has used two Flood-Based

---

\* The author is also an associate researcher at the Lab of Computational Science and Informatics, Institute of Mathematical Research (INSPEM), University Putra Malaysia.

search algorithms (flooding and teeming) it combined them to receive the high performance search for unstructured peer-to-peer systems.

To accomplish this approach, we developed QuickFlood search algorithm. The algorithm is divided into two steps. In the first step, it follows the flooding with a limited number of hops. In the second step, it implements teeming algorithm for all nodes remained from the first step. We provided analytical estimation for the best threshold point to switch from flooding to teeming. Our analytical results are validated through simulation. Integrating these two steps decreases the most redundant messages and increases the success rate of search in unstructured peer-to-peer networks.

The rest of this paper is organized into six sections: Section 2 reviews related work. Section 3 Flooding and Teeming across hops. Section 4 describes the QuickFlood algorithm. Section 5 discusses the performance evaluation, and Section 6 presents the conclusion.

## 2 Related Work

Flood-based search algorithm can be categories as: 1) TTL Controlled-Based (TCB), 2) Probabilistic Controlled-Based (PCB), and 3) Hybrid Controlled-Based (HCB).

TTL Controlled-Based (TCB), are those which limit the TTL of flooding to gain better performance. Expanding ring search (ERS) is the first TTL control-based algorithm. ERS is successive flooding by different TTL values. [2] shows ERS successfully reins in the TTL as the object's replication increases. Although the ERS algorithm is more efficient than flooding, but still it produces overhead due to repeated query messages caused by inappropriate choice of TTL value. However, if the file is far away from the query originator this approach could even generate sever overshooting messages than flooding [3].

Probabilistic Controlled-Based (PCB), limit the number of immediate neighbors. Random-Walks [2] is an adopted version of flooding, which forward a query message (walker) [2] to a randomly chosen neighbor until the query is found or maximum TTL visited. Standard Random-Walk use one walker (message) in each walk, it can reduce an order of magnitude overshooting messages compared to ERS and flooding. However, there is also an order of a magnitude increase in user-perceived delay. Modified Breadth-First-Search (MBFS) [4] or Teeming [5] is modified version of Random-Walk. The (MBFS) or Teeming are forwarded query to a random subset of its neighbors.

Hybrid Controlled-Based (HCB), combines two structures to benefit their merits and to limit their drawbacks. Local flooding with  $k$  independent random walks [6]. The idea is first perform a (local) flooding until precisely  $k$  new outer nodes have been discovered, if one of these nodes host the object, the search is successful and query source is informed. if no, each of the  $k$  nodes begins an independent random walk. If the file is located close to origin, the local flooding would be sufficient otherwise by independent random walk messages production will be high and the performance will be degraded.

### 3 Flooding and Teeming across Hops

Flooding conducted in a hop-by-hop fashion. By increasing of hops, it gains newer peers, and generates more messages. Part of these messages is redundant messages. This section investigated the trend of coverage growth rate and redundant messages in flooding and teeming.

#### 3.1 Trend of Coverage Growth Rate in Flooding

Assume that an overlay network as a random graph. Each node is represented as a peer, and they are connected to each other by edges. The degree of each peer represents the number of its immediate neighbors. Assume that the graph has  $N$  nodes with the average degree  $d_f$ ,  $d_f$  is greater than 2. The total messages broadcasting from each peer up to hop  $t$  as given in [7], is:

$$TM_{tf} = \sum_{i=1}^t d_f (d_f - 1)^{i-1} \quad (1)$$

(Loop nodes or cyclic paths are grouping of nodes linked together as a ring fashion. In Gnutella and other internet topology, there are many cyclic paths.) If there is no cyclic paths [8] or loop nodes; in the topology, then the total number of new peers visited so far is equal to:

$$TP_{tf} = \sum_{i=1}^t d_f (d_f - 1)^{i-1} \quad (2)$$

Thus, the coverage growth rate of messages [1] in hop  $t$  is equal to:

$$CGR_{tf} = \frac{TP_{tf}}{TP_{t-1}} = 1 + \frac{(d_f - 1)^{t-1}}{\sum_{i=1}^{(t-1)} d_f (d_f - 1)^{i-1}} \quad (3)$$

By taking first order derivation of Eq. (3), thus yields:

$$\frac{\partial(CGR_{tf})}{\partial t} = \frac{(d_f - 1)^{t-1} (-d_f) \log(d_f - 1)}{(d_f - 1)^{t-1} - 1} \quad (4)$$

The value of  $\left(\frac{\partial(CGR_{tf})}{\partial t}\right)$  is always negative. So the  $(CGR_{tf})$  is always in descending order. By increasing the value of  $t$  (hops), the value of  $(CGR_{tf})$  decreases, thus the maximum value of  $CGR_{tf}$  is presented in second hop. Therefore, we can show:

$$CGR_{2f} > CGR_{3f} > CGR_{4f} > CGR_{5f} \dots \quad (5)$$

#### 3.2 Trend of Redundant Messages in Flooding

Redundant messages in unstructured P2P networks are generated by loop or cyclic paths routs [9]. Assume that there is just one loop in each hop of the

default topology. Thus the redundant messages are generated in hop  $t$  is equal to:

$$R_{tf} = \sum_{i=1}^{t-2} (d_f - 1)^{i-1} = \frac{(d_f - 1)^{t-1} - 1}{d_f - 2} \quad (6)$$

Clearly, the total number of redundant messages generated up to hop  $t$  is:

$$TR_{tf} = \sum_{i=2}^t \sum_{k=1}^{i-2} (d_f - 1)^{k-1} \quad (7)$$

The value of  $(TR_{tf})$  depend to number of hop  $t$ , by increases  $t$  this value increases exponentially. Therefore, we can show that:

$$TR_{2f} < TR_{3f} < TR_{4f} < TR_{5f} \dots \quad (8)$$

### 3.3 Trend of Coverage Growth Rate in Teeming

Assume that  $\theta$  is the fixed probability of selecting neighbors. Thus the number of neighbors per peers selected in each hop is equal to  $\theta d_i$ . So the average degree in teeming is equal to:

$$d_t = \frac{\sum_{i=1}^n \theta d_i}{n} = \theta d_f \quad (9)$$

Therefore, the coverage growth rate of messages for teeming can be estimate as:

$$CGR_{tt} = 1 + \frac{(d_t - 1)^{t-1}}{\sum_{i=1}^{(t-1)} d_t (d_t - 1)^{i-1}} \quad (10)$$

Obviously, same as  $(CGR_{tf})$  the trend of  $(CGR_{tt})$  is always in descending order therefore, we can show that:

$$CGR_{2t} > CGR_{3t} > CGR_{4t} > CGR_{5t} \dots \quad (11)$$

Therefore, the coverage growth rate of teeming  $CGR_{tt}$  is  $\theta$  times of  $CGR_{tf}$  in each hop.

### 3.4 Trend of Redundant Messages in Teeming

In teeming algorithm, a node propagates the requested query to each of its neighbors with a fixed probability  $\theta$ . Thus compared to flooding, it decreases the number of messages, which propagated. Thus the number of redundant messages in hop  $t$  can be shown:

$$R_{tt} = \sum_{i=1}^{t-2} (d_t - 1)^{i-1} = \frac{(d_t - 1)^{t-1} - 1}{d_t - 2} \quad (12)$$

And the total number of redundant messages generated up to hop  $t$  is:

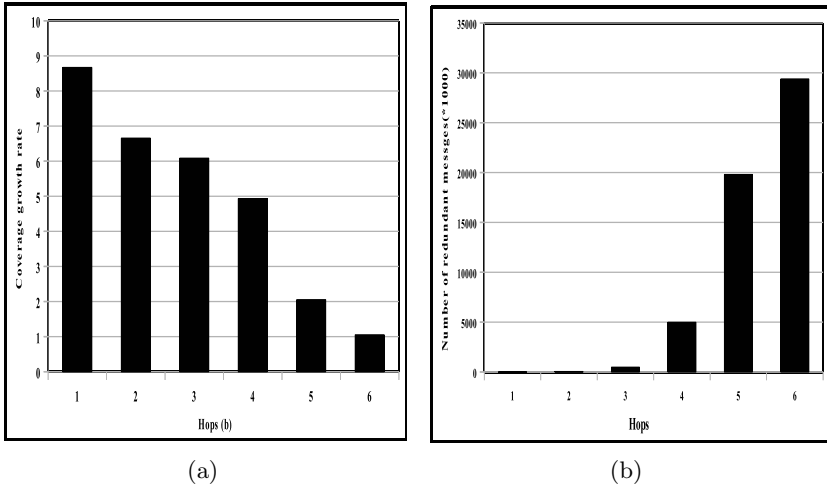
$$TR_{tt} = \sum_{i=2}^t \sum_{k=1}^{i-2} (d_t - 1)^{k-1} = \sum_{i=2}^t (t + 1 - i)(d_t - 1)^{i-2} \quad (13)$$

The trend of  $(TR_{tt})$  same as  $(TR_{tf})$  is always in ascending order. Therefore, we can show that:

$$TR_{2t} < TR_{3t} < TR_{4t} < TR_{5t} \dots \quad (14)$$

So the total number of redundant messages of teeming becomes  $\theta^{(t-2)}$  times of flooding in each hop.

We have examined this fact with sample topology, which is generated by BRITE (topology generator) [10] for 500 sample peers. Figure 1 represented the coverage growth rate of messages and the number of redundant messages in each hop. This observations show that pure flooding is efficient only in low-hops.



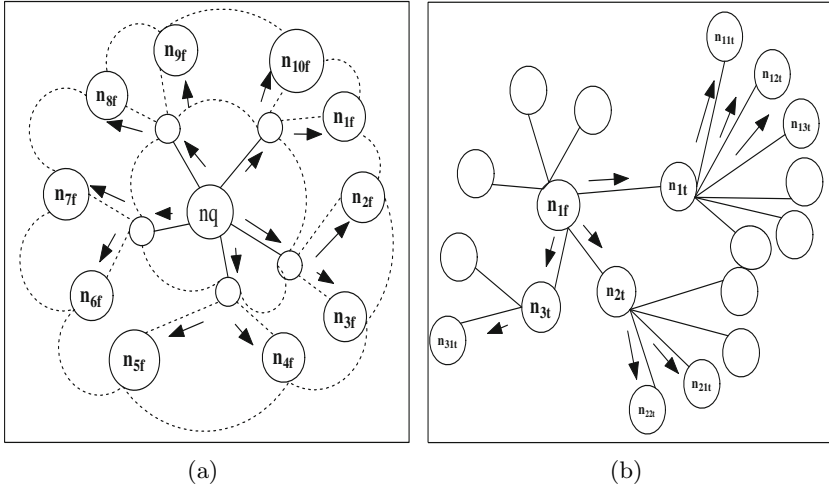
**Fig. 1.** (a) It represents coverage growth rate of messages in each hop. (b) It represents number of redundant messages in each hop.

## 4 QuickFlood Algorithm

The main idea behind the QuickFlood is to retain flooding and teeming merits and to limit their drawbacks. Our algorithm has two steps: at the first it starts by flooding for a limited number of hops. Whenever flooding starts to generate high redundant messages, our algorithm switches from flooding to teeming. Thus, it is important to find the optimum threshold for switch from flooding to teeming.

Figure 2 (a) illustrate the first step of QuickFlood (Flooding  $t_f$  hops), (b) illustrate the second step of QuickFlood (teeming algorithm  $t_t$  hops).

The important contribution of this work is to find the optimum threshold for switches from flooding to the teeming algorithm.



**Fig. 2.** (a) Illustrate the first step of QuickFlood algorithm. It starts from node  $n_q$ , with flooding algorithm after  $f$  hops it reaches to set of nodes such as  $\{n_{1f}, n_{2f}, n_{3f}, \dots, n_{10f}\}$ . (b) Illustrate the second step of QuickFlood (teeming algorithm with *e.g.*  $\emptyset = 50\%$ ) for sample node (*e.g.*  $n_{1f}$ ).

#### 4.1 Comparison Trends of Coverage Growth Rate and Redundant Messages in Flooding and Teeming

We defined the critical value ( $CR_{tx}$ ) for  $x$  algorithm in hop  $t$  as:

$$CR_{tx} = \frac{TR_{tx}}{CGR_{tx}} \quad (15)$$

As far as  $CR_{tx}$  decreases the efficiency of  $x$  algorithm in hop  $t$  increases. The critical value for our QuickFlood algorithm is a combination of  $CR_{tf}$  (flooding) and  $CR_{tt}$  (teeming).

$$CR_{th} = CR_{tf} + CR_{tt} \quad (16)$$

Thus, by substituting  $CGR_{tf} = K$  and  $TR_{tf} = L$  into Eq. (16), we have:

$$CR_{th} = \frac{L}{K} + \frac{L}{K}\theta^{(t-3)} = \frac{L}{K}(1 + \theta^{(t-3)}) \quad (17)$$

Hence, as far as  $\theta$  is less than one the value of  $\frac{L}{K}\theta^{(t-3)}$  is not less than  $\frac{L}{K}$  for  $t = 1, 2,$  and  $3$ . Thus it is not rational to use teeming algorithm in hops  $1, 2,$  and  $3$  in this combination. But for  $t \geq 4$  the value of  $\frac{L}{K}\theta^{(t-3)}$  started to decrease compared to  $\frac{L}{K}$ , the rate of decrease depend to value of  $\theta$ . So the optimum threshold for switching from flooding to teeming is when  $t = 4$ . Thus the best combination for QuickFlood is, to use flooding algorithm in the first three hops, and teeming algorithm in the rest of hops.



## 5 Performance Evaluation

The goal of our evaluation is to study the performance of QuickFlood compared with flooding, ERS, blocking expanding ring search (BERS), and teeming algorithms. Our algorithm is implemented in two steps. In the first step, it performed  $M$  hops with flooding algorithm, and in the second step, it continues  $N$  hops with the teeming algorithm by fixed probability of  $\theta$ . Hence, there is an interesting question which must be investigated by  $(M, N, \text{ and } \theta)$  arrangement.

- What is the effect of increasing or decreasing  $M, N,$  and  $\theta$  in the performance of QuickFlood?

### 5.1 Performance Metrics

The following metrics evaluated by our simulation experiments:

1. Queries success rate
2. Number of redundant messages

The queries success rate is used to measure the user perceived query quality (searching algorithm), i.e., how likely the a query can be solved [11]? As far as the success rate increases, the traffic of peer-to-peer network decrease, and it makes the load of network balance.

The number of redundant messages is used to measured search algorithm quality. The main characteristic of the search algorithm is to generate minimum overhead. The overhead of a search algorithm can be quantified by the number of redundant messages.

### 5.2 Network Topology and Simulation

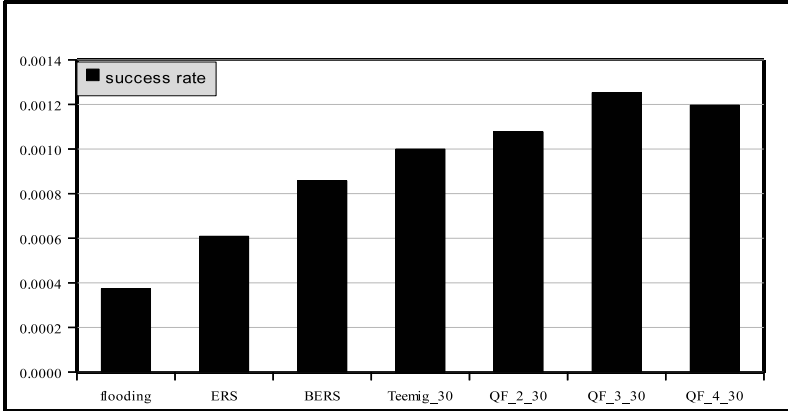
To perform this evaluation, we used Gnutella’s topology collected during the first six month of 2001, which was provided by Clip 2 Distributed Search Solution [12]. The name of this topology is T1, and it consists of 42822 nodes. The average degree of T1 is 3.4 and the average 2-hop neighbor’s peer is 34.1. We used this connectivity graph to simulate the behavior of our algorithm with predefine algorithms such as flooding, teeming, ERS, and BERS.

We set the replication ratio 0.00125, since there are more than 40,000 nodes. The resources were copied uniformly in 50 nodes. Each search was for one result. Simulation was performed 50 times for 40 different nodes.

### 5.3 The Effect of Increasing Hops in First Step of QuickFlood with Fixed Probability

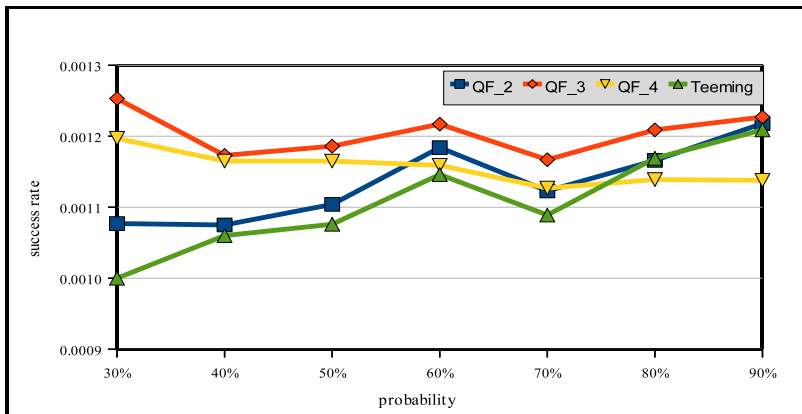
Figures 3 and 5 show the performance of flooding, ERS, BERS, teeming, and QuickFlood with the different arrangements of  $M$  (e.g.  $M = 2, 3, 4$ ) and fix  $\theta = 30\%$ .

Figure 3 shows that the success rate of flooding, ERS, BERS, teeming\_30 and QuickFlood with different arrangements of  $M$  and fixed probability ( $\theta = 30\%$ ).



**Fig. 3.** It compared the average of success rate in flooding, ERS, BERS, teeming\_30, and QuickFlood with different arrangements of  $M$  and  $\theta = 30\%$  for 2000 nodes

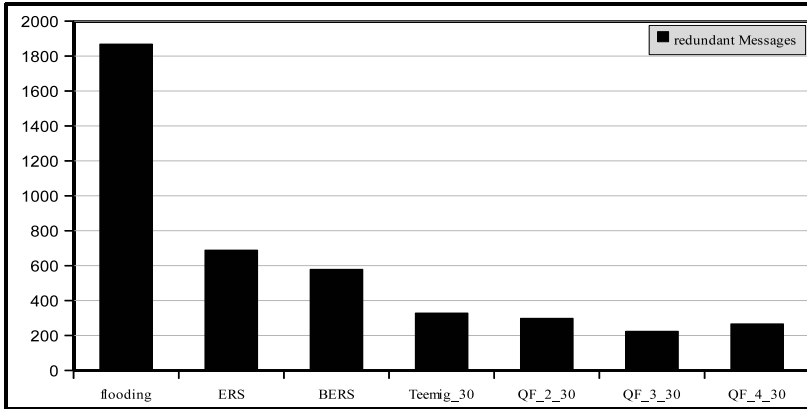
It presented flooding has the lowest and QF\_3\_30 has the highest success rate of all. It shows QF\_3\_30 has more than 3.35 times the success rate, when compared to flooding. Our simulation proved that QuickFlood with different arrangements of  $M$  and fixed probability ( $\theta = 30\%$ ) always gained the better quality of search compared to other algorithms.



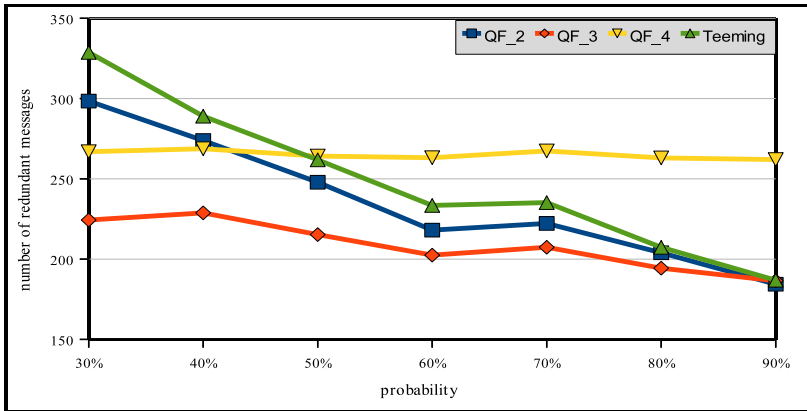
**Fig. 4.** It represented the success rate of different arrangements of QuickFlood and teeming

Figure 4 represent the performance of QuickFlood and teeming with different arrangements. It shows that the success rate of QF\_3\_30 is the highest value compared to other arrangements of QuickFlood and teeming.

In Figure 5 we compared the average number of redundant messages for flooding, ERS, BERS, teeming\_30, and QuickFlood with different arrangements of  $M$



**Fig. 5.** It compared the average number of redundant messages in flooding, ERS, BERS, teeming\_30, and QuickFlood with different arrangements of  $M$  and  $\theta = 30\%$  for 2000 nodes



**Fig. 6.** It represented the redundant messages of different arrangements of QuickFlood and teeming

and fixed probability ( $\theta = 30\%$ ). It shows that flooding has the highest and QF\_3.30 has the lowest redundant messages. It presented that QF\_3.30 reduces about 90% of redundant messages, while ERS reduces about 60%. Figure 5 shows all QuickFlood algorithm had low redundant messages compared to flooding, ERS, and BERS. Figure 6 compared different arrangement of QuickFlood and teeming. It shows QuickFlood with 3 hops flooding and rest teeming has almost the minimum redundant messages.

According to the above performance metrics, the QuickFlood search algorithm receives more efficient than flooding, ERS, BERS, and teeming. It reduces many

redundant messages and traffic load whiles increases success rate and search quality. The best threshold for switching from flooding to QuickFlood base on our analytical result (17) is  $M = 3$  ( at hop = 4 ), because within low-hops of flooding it gain more coverage growth rates and low redundant messages.

## 6 Conclusion

Flooding is a fundamental searching algorithm its overhead imposed on the underlying infrastructure is large and threaten the scalability of distributed systems. To address this, we proposed a new searching algorithm called QuickFlood, which combines flooding with the teeming algorithm. It effectively combines the advantage of flooding, and teeming. QuickFlood received more coverage and low redundant messages from flooding within low-hops, and less overhead within high-hops from teeming. We derive the analytical results for best threshold point to switch from flooding to teeming, which is validated through simulation. Our QuickFlood, represented by its ( $M = 3$  and  $\theta = 30\%$ ) arrangement, provides a simple procedure to control flooding in the cost-effective way upon existing overly. Simulation experiments show that QuickFlood can reduce up to 90 percent of redundant messages and increase up to 3 times success rates of search.

**Acknowledgements.** The authors would like to thank the University Putra Malaysia, under the Research University Grant Scheme (RUGS) 05-03-10-1039RU for financial support

## References

1. Jiang, L.Z.X.W.H., Guo, S.: Lightflood: Minimizing redundant messages and maximizing scope of peer-to-peer search. *IEEE Transactions on Parallel and Distributed Systems* 19, 601–614 (2008)
2. Lv, P.C.E.L.K.S.S., Cao, Q.: Search and replication in unstructured peer-to-peer networks. In: *Proceeding of International Conference on Supercomputing*, vol. 30, pp. 258–259. IEEE, New York (2002)
3. Chou, C.-C.: Techniques for peer-to-peer content distribution over mobile ad hoc networks. Ph.D. dissertation, University of Southern California (December 2007)
4. Vana Kalogeraki, D.Z.-Y., Gunopulos, D.: A local search mechanism for peer-to-peer networks. In: *Conference on Information and Knowledge Management*, pp. 300–307. ACM Press, New York (2002)
5. Dimakopoulos, E., Pitoura, V.V.: On the performance of flooding-based resource discovery. *IEEE Transactions on Parallel and Distributed Systems* 17, 1242–1252 (2006)
6. Dorrigiv, A.L.-O.R.: Search algorithm for unstructured peer-to-peer networks. In: *Proc. of 32nd IEEE Conference of Local Computer Networks*, pp. 343–349 (2007)
7. Aberer, M.H.: An overview on peer-to-peer information systems. In: *Proceedings of WDAS*, pp. 171–188. Carleton Scientific, Ottawa (2002)
8. Zhu, S.Z., Kalnis, P.B.: Dcmp: A distributed cycle minimization protocol for peer-to-peer networks. *IEEE Transactions On Parallel and Distributed Systems* 19, 363–377 (2008)

9. Hung-Chang Hsiao, H.L., Huang, C.-C.: Resolving the topology mismatch problem in unstructured peer-to-peer networks. *IEEE Transactions on Parallel and Distributed Systems* 20, 1668–1681 (2009)
10. Alberto Medina, I.M.J.B.: Brite: A flexible generator of internet topologies. Boston, MA, USA, Tech. Rep. 2000-005
11. Xiao, C.W.L.: An effective p2p search scheme to exploit file sharing heterogeneity. *IEEE Transactions on Parallel and Distributed Systems* 18, 145–157 (2007)
12. Clip2 Distributed Search Solution (2007), <http://www.clip2.com>

# Improved-XY: A High Performance Wormhole-Switched Routing Algorithm for Irregular 2-D Mesh NoC

Ladan Momeni<sup>1</sup>, Arshin Rezazadeh<sup>2</sup>, and Davood Abednejad

Department of Computer Engineering

<sup>1</sup> Islamic Azad University – Ahvaz Branch, Iran

<sup>2</sup> Iran University of Science and Technology

University Road, Narmak, Tehran, Iran 16846-13114

Ladan\_momeni296@yahoo.com, rezazadeh@comp.iust.ac.ir,  
a.rezazadeh@yahoo.com

**Abstract.** Modifying irregular routing algorithms which are based on fault-tolerant algorithms, they can be utilized by irregular networks. These algorithms in general use several virtual channels to pass faults. In this paper, a new wormhole-switched routing algorithm for irregular 2-dimensional (2-D) mesh interconnection Network-on-Chip is evaluated, where not only no virtual channel is used for routing but also no virtual channel is used to pass oversized nodes (ONs). We also improve message passing parameters of ONs as well as comparing simulation results of our algorithm and several state of art algorithms. Simulation results show that our proposed algorithm, i-xy (improved/irregular-xy), has a higher saturation point in comparison with extended-xy and OAPR algorithms. Furthermore, it has less blocked messages and higher routed/switched messages in the network. Moreover, the network uses i-xy has higher utilization compared to other networks which uses e-xy and OAPR from 35 percent to 100 percent, for the irregular 2-D mesh NoC.

**Keywords:** Network-on-Chip, performance, wormhole switching, irregular 2-D mesh, routing, utilization.

## 1 Introduction

As technology scales, Systems-on-Chips (SoCs) are becoming increasingly complex and heterogeneous. One of the most important key issues that characterize such SoCs is the seamless mixing of numerous Intellectual Property (IP) cores performing different functions and operating at different clock frequencies. In just the last few years, Network-on-Chip (NoC) has emerged as a leading paradigm for the synthesis of multi-core SoCs [1]. The routing algorithm used in the interconnection communication NoC is the most crucial aspect that distinguishes various proposed NoC architectures [2], [3]. However, the use of VCs introduces some overhead in terms of both additional resources and mechanisms for their management [4].

Each IP core has two segments to operate in communication and computation modes separately [5]. On-chip packet switched interconnection architectures, called as NoCs, have been proposed as a solution for the communication challenges in these

networks [6]. NoCs relate closely to interconnection networks for high-performance parallel computers with multiple processors, in which each processor is an individual chip.

A NoC is a group of routers and switches that are connected to each other on a point to point short link to provide a communication backbone of the IP cores of a SoC. The most common template that proposed for the communication of NoC is a 2-D mesh network topology where each resource is connected with a router [7]. In these networks, source nodes (an IP-Core), generate packets that include headers as well as data, then routers transfer them through connected links to destination nodes [8].

The wormhole (WH) switching technique proposed by Dally and Seitz [9] has been widely used in the interconnections such as [10], [11], [12], [15] and [16]. In the WH technique, a packet is divided into a series of fixed-size parts of data, called flits. Wormhole routing requires the least buffering (flits instead of packets) and allows low-latency communication. To avoid deadlocks among messages, multiple virtual channels (VC) are simulated on each physical link [12]. Each unidirectional virtual channel is realized by an independently managed pair of message buffers [13].

This paper presents a new routing algorithm for irregular mesh networks by base that enhances a previously proposed technique. The primary distinction between the previous method and the method presented in this paper is passing messages from ONs in the network. Simulation results show that utilization of network by e-xy and OAPR algorithm is worse than the improved one, i-xy. We have been simulated every three algorithms for 5% and 10% of oversized nodes with uniform and hotspot traffic. Results for all situations show that our algorithm has higher utilization and can work in higher message injection rates, with higher saturation point.

The rest of the paper is organized as follows. In section 2 some deterministic-based routing algorithms are discussed. Then the new i-xy irregular routing algorithm is explained followed by Section 3 in which our experimental results are discussed. Finally, Section 4 summarizes and concludes the work.

## 2 Irregular Routing

Routing is the act of passing on data from one node to another in a given scheme [11]. Currently, most of the proposed algorithms for routing in NoCs are based upon deterministic routing algorithms which in the case of oversized nodes, cannot route packets. Since adaptive algorithms are very complex for Network-on-Chips, a flexible deterministic algorithm is a suitable one [14]. Deterministic routing algorithms establish the path as a function of the destination address, always applying the same path between every pair of nodes. This routing algorithm is known as dimension-order routing (x-y routing). This routing algorithm routes packets by crossing dimensions in strictly increasing (or decreasing) order, reducing to zero the offset in one dimension before routing in the next one [13]. To avoid deadlocks among messages, multiple virtual channels (VC) are simulated on each physical channel [12]. But in this paper, we use no VCs in proposed algorithm and introduced a deadlock and live lock-free irregular routing algorithm.

Many algorithms have been suggested to operate in faulty conditions without deadlock and livelock. We can modify these algorithms to use in irregular

interconnection networks. Some of these algorithms like [10], [11], [12], [15] and [16] are based on deterministic algorithms. In [15], Wu proposed a deterministic algorithm. This proposed algorithm uses odd-even turn model to pass the block faults. Also, the algorithm proposed by Lin et al. [16] uses above mentioned method. Since our proposed algorithm is similar to these algorithms (uses no virtual channel), in the next section, we are going to describe how these deterministic algorithms work and how we have improved them. The main idea describes in the rest of this section.

## 2.1 Extended-XY Routing Algorithm

The algorithm presented by Wu [15], extended-xy, uses no VCs by implementing odd-even turn model which is discussed in [17]. Such an algorithm is able to pass faulty ring and orthogonal faulty blocks. This algorithm consists two phases; in phase 1, the offset along the x dimension is reduced to zero and, in phase 2, the offset along the y dimension is reduced to zero [15].

This algorithm has two modes, normal and abnormal mode. The extended-xy routing follows the regular x-y routing (and the packet is in a “normal” mode) until the packet reaches a boundary node of a faulty block. At that point, the packet is routed around the block (and the packet is in an “abnormal” mode) clockwise or counterclockwise based on certain rules: Unlike routing in a fault-free routing, the fault-tolerant routing protocol has to prepare for “unforeseen” situations: a faulty block encountered during the routing process. This is done by three means: 1) the packet should reside in an even column when reaching a north or south boundary node of the routing block in phase 1. 2) In phase 1, the packet should be routed around the west side since, once the packet is east-bound, it cannot be changed to west-bound later. 3) The two boundary lines, one even and one odd, offer just enough flexibility for the packet to make turns for all situations.

In phase 2, to route around the routing block, odd columns (even columns) are used to perform routing along the y dimension when the packet is east-bound (west-bound). The packet is routed around the routing block either clockwise or counterclockwise in phase 2. Note that during the normal mode of routing the packet along the x or y dimension, no 180 degrees turn is allowed. For example, the positive x direction cannot be changed to the negative x direction [15]. Additional information and introduced algorithm about extended-xy algorithm can be found in [15].

## 2.2 OAPR Routing Algorithm

The algorithm presented by S.Y. Lin et al. [16], OAPR, described as follows:

- 1) Avoid routing paths along boundaries of ONs. In the environment of faulty meshes, we can only know the information of faulty blocks in real-time. However, the locations of ONs are known in advance. Therefore, the OAPR can avoid routing paths along boundaries of ONs and reduce the traffic loads around ONs.
- 2) Support f-rings and f-chains for placements of ONs. The OAPR solves the drawbacks of the e-xy and uses the odd-even turn model to avoid deadlock systematically. However, the e-xy cannot support ONs placed at boundaries of irregular meshes. In order to solve this problem, the OAPR applies the concepts of f-rings and f-chains [12]. With this feature, the OAPR can work correctly if ONs are



placed at the boundaries of irregular meshes. Additional information and introduced algorithm about extended-xy algorithm can be found in [16].

### 2.3 Improved-XY Routing Algorithm

This algorithm is based on if-cube2 [10], [11], and similar to extended-xy [15], OAPR algorithm [16] and odd-even turn model [17] uses no virtual channel. Like extended-xy algorithm, able to pass ring blocks of oversized nodes and also chain blocks that not considered in extended-xy routing. Moreover, when a network uses OAPR algorithm, all ONs vertically overlapping must be aligned on the east edge, but in improved-xy this constraint has been removed. Like [11] each message is injected into the network as a row message and its direction is set to null until it reaches to the column of the destination node. Then it would be changed as a column message to reach the destination. A column message could not change its path as a row message, unless it encounters with oversized region. In such a situation, a column message could change its direction into clockwise or counter-clockwise. First, each message should be checked if it has reached to destination node. Else, if this message is a row message and has just reached to the column of destination node, it would be changed as a column message.

For regular meshes, the e-cube provides deadlock-free shortest path routing. At each node during the routing of a message, the e-cube specifies the next hop to be taken by the message. The message is said to be blocked by an oversized node, if its e-cube hop is on an oversized region. The proposed modification uses no virtual channels and tolerates multiple oversized blocks.

To route messages around rings or chains, messages are classified into four types: East-to-West (EW), West-to-East (WE), North-to-South (NS), or South-to-North (SN). EW and WE messages are known as row messages and NS and SN as column messages. A message is labeled as either an EW or WE message when it is generated, depending on its destination. Once a message completes its row hops, it becomes a NS or a SN message to travel along the column. Thus, row messages can become column messages; however, NS and SN messages cannot change their types.

Next, if a message encountered with an oversized region, the Set-Direction(M) procedure would be called to set the direction of the message. The role of this procedure is to pass oversized region by setting the direction of message to clockwise or counter-clockwise. Again, the direction of the message will be set to null when it passed oversized region. While the direction of a message is null, e-cube algorithm used to route messages and it can be use odd/even row/columns. Fig. 1 show the using of odd and even row and columns when a message is passing an oversized node.

Using this modification of passing oversized regions, simulations are performed to evaluate the performance of the enhanced algorithms in comparison with the algorithms proposed in prior work. Simulation results indicate an improvement in the utilization and switched/routed messages for different cases of ONs, and different traffics. Furthermore, the enhanced approach can handle higher message injection rates (i.e., it has a higher saturation rate). In the following of this section, the proposed algorithm, Improved-XY(i-xy), and Set-Direction(M) procedures, have been given.

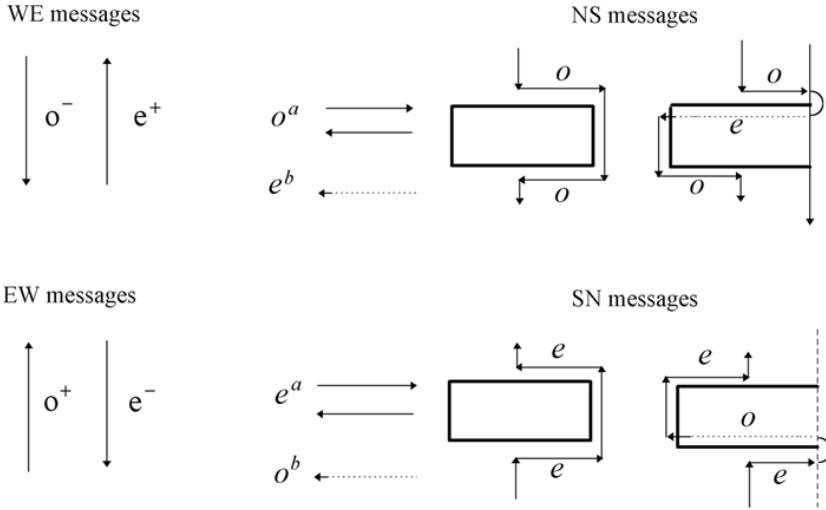


Fig. 1. Usage of odd and even row or columns

#### Algorithm Improved-XY(i-xy)

/\* the current host of message M is  $(s1, s0)$  and its destination is  $(d1, d0)$ . \*/

0. If  $s1 = d1$  and  $s0 = d0$ , consume M and return.
1. If M is a row message and  $s0 = d0$  then change its type to NS, if  $s1 > d1$ , or SN, if  $s1 < d1$ .
2. If the next e-cube hop is not blocked by an oversized node, then set the status of M to normal and set the direction of M to null.
3. Otherwise, set the status of M by Set-Direction(M).
4. If the direction of M is null, then use its x-y hop,
5. Otherwise, route M on the oversized node according to the specified direction.

#### Procedure Set-Direction(M)

0. If M is a column message and its direction is null, then set  $(l1, l0) = (s1, s0)$ .
1. If the direction of M  $\neq$  null and the current node is an end node then reverse the direction of M and return.
2. If M is a column message and  $s0 \neq l0$ , then return.
3. If M is a column message and  $s1 \neq l1$ ,  $s0 = l0$ , then set its direction to null.
4. If the next e-cube hop of M is not faulty, set its direction to null and return.
5. If direction of M is not null, then return.
6. If M is a WE message, set its direction to

- 6.1 clockwise if  $s1 < d1$ , or
- 6.2 counter-clockwise if  $s1 > d1$ , or
- 6.3 either direction if  $s1 = d1$ .
- 7. If M is an EW message, set its direction to
  - 7.1 clockwise if  $s1 > d1$ , or
  - 7.2 counter-clockwise if  $s1 < d1$ , or
  - 7.3 either direction if  $s1 = d1$ .
- 8. If M is an NS message, set its direction to clockwise, if the current node is not located on the EAST boundary of 2D meshes, or counter-clockwise, otherwise, and set  $(l1, l0) = (s1, s0)$ .
- 9. If M is an SN message, set its direction to counter-clockwise, if the current node is not located on the EAST boundary of 2D meshes, or clockwise, otherwise, and set  $(l1, l0) = (s1, s0)$ .

#### 2.4 Deadlock- and Live lock-Freeness

A WE message can travel from north to south or south to north, if its next e-cube hop is an oversized node. A north-to-south (south-to-north) WE message can take south-to-north (north-to-south) hops only if it encounters an end node and takes an u-turn at the end node. No deadlock occurs among EW messages can be assured by similar statements. NS messages can travel from north to south but not from south to north; there can't be a deadlock between NS messages waiting in different rows. NS messages are designed to get around the oversized components in a counterclockwise direction. An NS message can take an u-turn at an end node on the west boundary of 2-D meshes and change its direction to be clockwise, but can't take an u-turn at the east boundary of 2-D meshes, since no entire row of out-of-order components is allowed. Thus, no deadlock can occur between NS messages waiting on the same row. No deadlock can occur among SN messages that are assured by similar statements. Since the number of oversized nodes and broken links is finite and message never visits an oversized node more than once, our routing scheme is also live lock-free.

### 3 Results and Discussions

In this section, we describe how we perform the simulation and obtain results from simulator. Moreover, we show the improvements of the primitive algorithms by our modification. In order to model the interconnection network, an object-oriented simulator was developed base on [10], [11].

Some parameters we have considered are an average number of switched messages (ANSM) and average number of routed messages (ANRM) in each period of time. The other examined parameter in this paper is the utilization of the network which is using our routing algorithm, i-xy. Utilization illustrates the number of flits in each cycle, which passed from one node to another, in any link over bandwidth. Bandwidth

is defined as the maximum number of flits could be transferred across the normal links in a cycle of the network. We have examined utilization over message injection rate (MIR) and average message delay (AMD) over utilization for all sets of cases. The last parameter we have considered is the average number of blocked messages (ANBM) in the network. Simulation methodology describes in the rest of this section.

### 3.1 Simulation Methodology

A flit-level simulator has been designed. We record average message latencies, utilization and some other parameters measured in the network with the time unit equal to the transmission time of a single flit, i.e. one clock cycle. Our study is performed for different rates: 5%, and 10% of oversized nodes. Our generation interval has exponential distribution which leads to Poisson distribution of number of generated messages per a specified interval. In our simulation studies, we assume message length to be equal to 32 flits and we use an 8 x 8 2-dimensional irregular mesh network, and it takes one cycle to transfer a flit on a physical channel. Two different traffic patterns are simulated:

- A) Uniform traffic: The source node sends messages to any other node with equal probability.
- B) Hotspot traffic: Messages are destined to a specific node with a certain probability and are otherwise uniformly distributed.

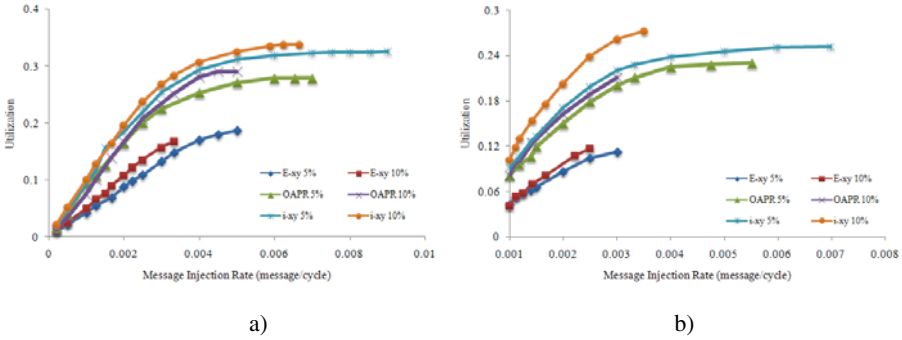
The number of messages generated for each simulation result, depends on the traffic distribution, and is between 1,000,000 to 3,000,000 messages. The simulator has three phases: start-up, steady-state, and termination. The start-up phase is used to ensure the network is in steady-state before measuring message latency. So, we do not gather the statistics for the first 10% of generated messages. All measures are obtained from the remaining of messages generated in steady-state phase. Messages generated during the termination phase are also not included in the results. The termination phase continues until all the messages generated during second phase have been delivered [10], [11]. In the rest of this section we study the effect of using predefined odd/even row and columns on the performance of i-xy. We perform this analysis under a different traffic distribution pattern. It is noted that only parts of simulation results are presented in this paper.

Figures 2, 3, 4, 5, and 6 show the simulation results for two different oversized node cases, 5 percent and 10 percent, with uniform and hotspot ( $p=10\%$ ) traffic.

### 3.2 Comparison of i-xy, e-xy, and OAPR Routing Algorithms

Uniform traffic is the most used traffic model in the performance analysis of interconnection networks [10], [11]. Fig. 2a, 3a, 4a, 5a, and 6a displays the effect of the improvement on the performance of i-xy, e-xy, and OAPR routing algorithms in 2-D irregular mesh interconnection network for this traffic pattern.

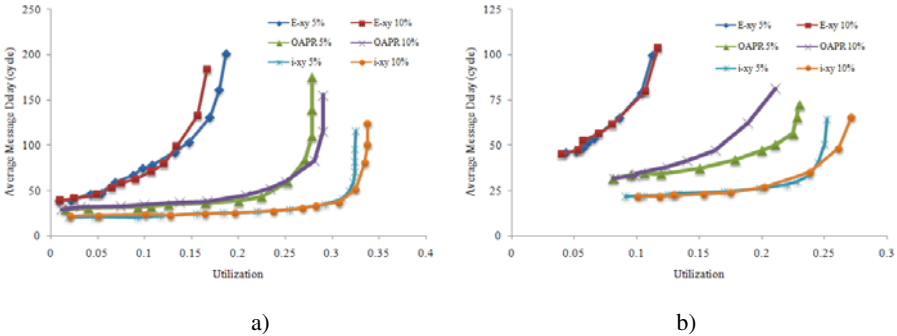
In order to generate hotspot traffic we used a model proposed in [10]. According to this model each node first generates a random number. If it is less than a predefined threshold, the message is sent to the hotspot node. Otherwise, it is sent to other nodes of the network with a uniform distribution.



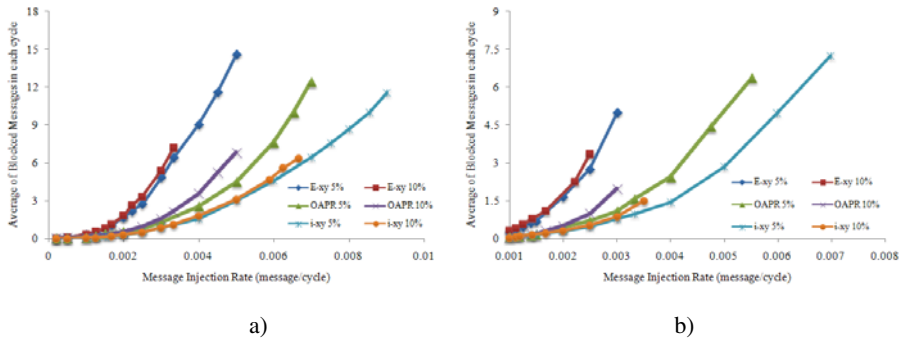
**Fig. 2.** Utilization of i-xy, e-xy, and OAPR routing algorithms for 5% and 10% ONs by 32 flits packets a) Uniform traffic b) Hotspot traffic

As the mesh interconnection network is not a symmetric network, we have considered two types of simulation for hotspot traffic in this network. In one group of simulations, a corner node is selected as the hotspot node and in the other group; a node in the middle of the network is chosen as the hotspot node, and finally averaged. Hotspot rate is also considered in our study, namely 10%. Fig. 2b, 3b, 4b, 5b, and 6b illustrates the effect of the performance of every three above mentioned routing algorithms for hotspot traffic distribution pattern.

We defined utilization as the major performance metric. For an interconnect network, the system designer will specify a utilization requirement. Fig. 2a and 2b shows the utilization over the message injection rate for two cases of oversized nodes with two different traffic patterns, uniform and hotspot traffic, on 8 x 8 irregular 2-dimensional mesh Network-on-Chip. As we can see, the network which uses extended-xy and OAPR algorithm is saturated with low MIR while the improved-xy algorithm has a higher saturation point. As an example in 10% case of extended-xy, the utilization for 0.0033 MIR is lower 16.67% and for OAPR is 25%, yet the other algorithm, improved-xy, works normally even for 0.0067 MIR with 33.8% utilization at 100% traffic load (fig. 2a). In fact our irregular routing algorithm has higher utilization. Additionally, improvement can be found in other traffic pattern.



**Fig. 3.** Performance of i-xy, e-xy, and OAPR routing algorithms for 5% and 10% ONs by 32 flits packets a) Uniform traffic b) Hotspot traffic



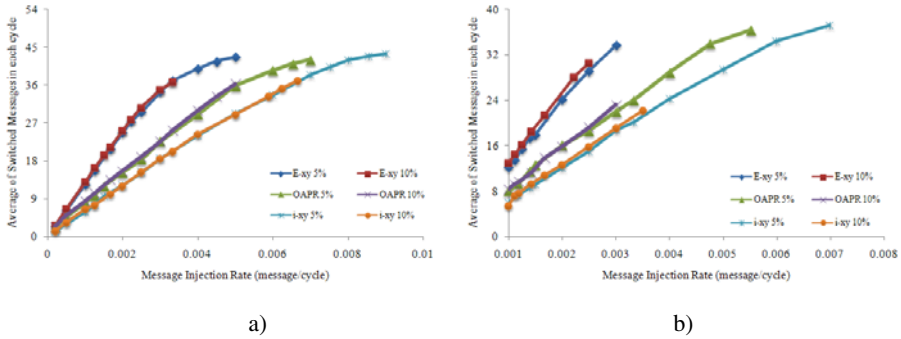
**Fig. 4.** Average number of blocked messages (ANBM) of i-xy, e-xy, and OAPR routing algorithms for 5% and 10% ONs by 32 flits packets a) Uniform traffic b) Hotspot traffic

The most important comparison we have done between these three algorithms is the rate of average message delay over utilization. Comparative performance across different cases in fig. 3a and fig. 3b is specific to the several oversized node sets used. For each case, we have simulated previous sets up to 100% traffic load.

As an example, we consider the amount of average message delay for both algorithms with 16% utilization in 5% mode in uniform traffic (fig. 3a). At this point, the network which uses e-xy has more than 183 AMD at 100% traffic load and the network uses OAPR has more than 38 AMD, while the other network using i-xy, has less than 24 AMD, and it has not been saturated. Comparing the utilization of these algorithms for 100% traffic load, it is obvious the network using i-xy has 32.5% utilization, whereas the OAPR has 27.86% and the other one has just 16.67% utilization. We have improved utilization of network more than 16% by our proposed algorithm at 100% traffic load compared to OAPR for this case, and about twice for extended-xy. Other case is also considerable.

The next parameter we have examined is the average number of blocked messages (ANBM) in each cycle which illustrates average number of blocked messages in the network because no buffer is available to pass to the next node. If nodes of a communication system have more free buffers, messages may deliver simply across the interconnection network.

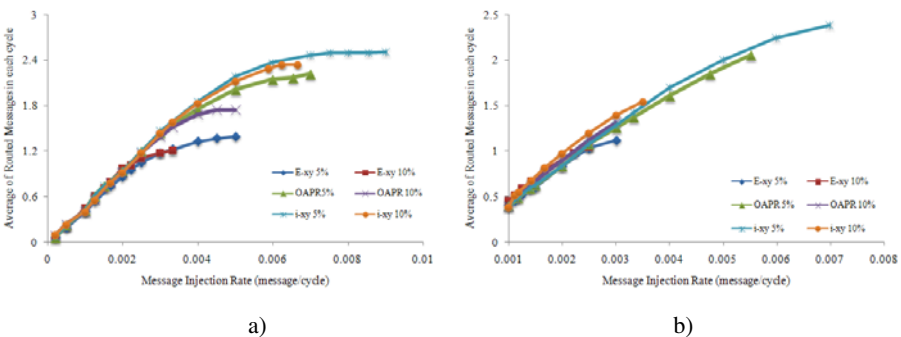
As it is shown in fig. 4a and 4b a fraction of delays which messages are encountered by, is the delay of waiting for an empty buffer for the next hop. For instance, comparing three algorithms in fig. 4b for 10% mode by hotspot traffic in 0.0025 MIR, it is clear that when a network uses e-xy, over 3.35 messages blocked in every cycle and this number for OAPR at this point is more than 0.99 messages, but by using i-xy algorithm, less than 0.55 messages blocked in every cycle. This condition is repeated for the other case shown in fig. 4a for uniform traffic which is substantial.



**Fig. 5.** Average number of switched messages (ANSM) of i-xy, e-xy, and OAPR routing algorithms for 5% and 10% ONs by 32 flits packets a) Uniform traffic b) Hotspot traffic

Fig. 5 shows the average number of switched messages (ANSM) in each cycle over the message injection rate (MIR) for all cases. It is clear; the network uses i-xy algorithm has minor improvement at 100% traffic load compared to the other two above mentioned algorithms. As an example in fig. 5a in 10% mode of extended-xy and OAPR, the ANSM at saturation point is about 36.5, yet ANSM for the other algorithm, improved-xy is more than 37. In fact our irregular routing algorithm has similar behavior for this parameter. But, this parameter for hotspot traffic distribution has better condition.

The last parameter we consider is the average number of routed messages (ANRM) in each cycle. As it is shown in fig. 6a and fig. 6b, the ANRM for improved-xy has higher in comparison to extended-xy and OAPR algorithms. For instance, in fig. 6b in hotspot traffic by 5% mode of extended-xy, the ANRM for 0.003 MIR is 1.12 messages and the network saturated at 0.0055. The network uses OAPR algorithm (at saturation point) has 2.05 ANRM, but this number for improved-xy algorithm is more than 2.37 in 0.007 MIR at saturation point. Also, enhancement can be found by using uniform traffic in fig. 6a.



**Fig. 6.** Average number of routed messages (ANRM) of i-xy, e-xy, and OAPR routing algorithms for 5% and 10% ONs by 32 flits packets a) Uniform traffic b) Hotspot traffic

## 4 Conclusion

Designing a deadlock-free routing algorithm that can tolerate unlimited number of oversized nodes is not an easy job. Oversized blocks are expanded, by disabling good nodes, to be rectangular shapes in existing literature to facilitate the designing of deadlock-free routing algorithms for 2-D irregular mesh networks. The simulation results show the improvement of network utilization (from 35% to 100%), which are needed to work with rectangular oversized nodes, can be recovered if the number of original oversized nodes is less than 10% of the total network.

We have been simulated every three algorithms for the same message injection rates, oversized node situations, message lengths, network size, and the percentage of oversized nodes and in many cases our studies have better results in comparison with the other two algorithms.

We also showed that in various traffics and different number of oversized nodes, these oversized blocks can be handled. The deterministic algorithm is enhanced from the non-adaptive counterpart by utilizing the way of passing oversized nodes by the proposed algorithm when a message is blocked. The method we used for enhancing the extended-xy and OAPR algorithms is simple, easy and its principle is similar to the previous algorithm, if-cube2. Moreover, ANBM and ANRM are improved by our proposed algorithm. In conclusion improved-xy has better performance compared to extended-xy and OAPR and is feasible for Network-on-Chip.

## References

1. Ivanov, A., De Micheli, G.: The Network-on-Chip Paradigm in Practice and Research. *IEEE Design and Test of Computers* 22(5), 399–403 (2005)
2. Bjerregaard, T., Mahadevan, S.: A Survey of Research and Practices of Network-on-Chip. *ACM Computing Surveys* 38(1), 1–51 (2006)
3. Pande, P., Grecu, C., Jones, M., et al.: Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures. *IEEE Trans. Computers* 54(8), 1025–1040 (2005)
4. Palesi, M., Holsmark, R., Kumar, S., et al.: Kumar, Sh., et al.: Application Specific Routing Algorithms for Networks on Chip. *IEEE Trans. on Parallel and Distributed Systems* 20(3), 316–330 (2009)
5. Guerrier, P., Greiner, A.: A generic architecture for on-chip packet-switched interconnections. In: *Proceedings Design Automation and Test in Europe Conference and Exhibition, Paris, France, March 2000*, pp. 250–256 (2000)
6. Srinivasan, K., Chatha, K.S.: A technique for low energy mapping and routing in network-on-chip architectures. In: *ISLPED 2005, California, USA, August 2005*, pp. 387–392 (2005)
7. Ali, M., Welzl, M., Zwicknagl, M., et al.: Considerations for fault-tolerant network on chips. In: *The 17th International Conference on Microelectronics, December 2005*, pp. 178–182 (2005)
8. Matsutani, H., Koibuchi, M., Yamada, Y., et al.: Non-minimal routing strategy for application-specific networks-on-chips. In: *ICPP 2005, International Conference Workshops on Parallel Proceeding, June 2005*, pp. 273–280 (2005)



9. Dally, W.J., Seitz, C.L.: Deadlock-free message routing in multiprocessor interconnection networks. *IEEE Trans. on Computers* 36(5), 547–553 (1987)
10. Rezazadeh, A., Fathy, M.: An Enhanced Fault-Tolerant Routing Algorithm for Mesh Network-on-Chip. In: *International Conference on Embedded Software and Systems*, pp. 505–510 (2009)
11. Rezazadeh, A., Fathy, M.: Throughput Considerations of Fault-Tolerant Routing in Network-on-Chip. In: *2nd International Conference on Contemporary Computing (IC3-2009), Communications in Computer and Information Science (CCIS)*, pp. 81–92. Springer, Heidelberg (2009)
12. Boppana, R.V., Chalasani, S.: Fault-tolerant wormhole routing algorithms for mesh networks. *IEEE Trans. Computers* 44(7), 848–864 (1995)
13. Duato, J., Yalamanchili, S., Ni, L.: *Interconnection networks: An engineering approach*. Morgan Kaufmann, San Francisco (2003)
14. Dally, W.J., Towles, B.: Route packets, not wires: On-chip interconnection networks. In: *Proceedings Design Automation Conference, USA, June 2001*, pp. 684–689 (2001)
15. Wu, J.: A Fault-Tolerant and Deadlock-Free Routing Protocol in 2D Meshes Based on Odd-Even Turn Model. *IEEE Trans. on Computers* 52(9), 1154–1169 (September 2003)
16. Lin, S.Y., Huang, C. H., Chao, C.H., Wu, A.: Traffic-Balanced Routing Algorithm for Irregular Mesh-Based On-Chip Networks. *IEEE Trans. On Computers* 57(9), 1156–1168 (2008)
17. Chiu, G.M.: The Odd-Even Turn Model for Adaptive Routing. *IEEE Trans. on Parallel and Distributed Systems* 11(7), 729–737 (2000)

# XRD Metadata to Make Digital Identity Less Visible and Foster Trusted Collaborations across Networked Computing Ecosystems

Ghazi Ben Ayed and Solange Ghernaoui-Hélie

Faculty of Business and Economics, University of Lausanne, CH-1015,  
Lausanne, Switzerland  
{Ghazi.Benayed, Sgh}@unil.ch

**Abstract.** Distributed computing ecosystems' collaboration and mass integration between partners require extensive digital identities processing in order to better respond to services' consumers. Such processing is increasingly implies loss of user's control over identity, security risks and threats to privacy. Digital identity is represented by a set of linked and disparate documents distributed over computing ecosystems' domains. We suggest an innovative approach based on metadata management, which would make digital identity documents less visible, foster trusted partnership, and therefore encourage trusted collaboration among networked computing ecosystems. Furthermore, an XRD-based implementation of digital identity document metadata is provided and explained.

**Keywords:** Less visible identity, metadata documents management, trusted partnership, XRD, identity-based collaboration.

## 1 From PC to Computing Ecosystems

We still teach our children in schools that PC stands for and means Personal Computer. The "Personal" and "Computer" concepts are evolving. Personal is becoming more personal than before through the use of personal small devices such as PDA, notebooks, and mobile terminals [1]. We could refer to "Personal-case" or "Possessive" for the word "P" but the "C" is radically changing from computer into "Computing-ecosystems". Currently, individuals and organizations are increasingly surrounded by a landscape of not just computers but also embedded sensors, Software-as-a-Service, cloud-hosted applications, m-business services, smart phones, PDAs, entertainment centers, digital cameras and video recorders, Webcams, computers, e-mail, GPS tracking systems, Mashups, mobile storage units, networks, ubiquitous devices, and so on. Networked computing ecosystems need trusted intra- and inter-ecosystems collaboration capabilities in order to provide and consume services, and to demonstrate their *raison d'être*.

Individuals and organizations activities are increasingly planned and performed through the use of collaborating networked computing ecosystems. The "everydayness of the digital" expression [2] is used to refer that the foundation

today's society is constructed upon daily participations through, or tasks' delegations to networked computing ecosystems. Today's young generation, or "Digital Natives" [3], study, work, write, and interact with each other through notebooks. They read blogs rather than newspapers and meet each other online before they meet in person. They get their music online, often for free or illegally, rather than buying it in record stores. They're more likely to send an instant message (IM) than to pick up the telephone to arrange a date. They are constantly connected and they have plenty of friends both in real space and in the virtual worlds. They are frequently editing a profile on MySpace, making encyclopedia entries on Wikipedia, converting a movie format, and downloading a file from P2P file sharing networks. At the organizational level, computing ecosystems are collaborating to allow the achievement of a full coordination and mass integration between partners. The advent of standards is easing the extension of organizations by lowering the barriers to connecting disparate business applications both within and across organizational boundaries [3, 4].

Collaboration requires extensive data processing, which increasingly implies loss of user control over identity and threats to security and privacy. The latter risks could be identified and associated at the individual and organizational levels but in this article we focus only at the individual level. Web users are increasingly leaving trails on the net and most online service providers memorize, access, and exploit 'Web of trails' for their own commercial benefits. Ziki.com offers advertising services through Google ads based on the user's profile. As a consequence, users are losing control over their personal information that could compromise online security, privacy and trust [1, 5-7]. One hundred million worldwide Facebook users are threatened by identity theft, cyber-stalking and cyber-bullying, and digital espionage as a repercussion of Facebook hack case [8], in which personal details have been collated and published on file-sharing service. Fraud is rising rapidly because not only people are posting personal facts on the Web but government agencies are steadily making databases available online. The databases may include birth, marriage and death certificates, credit histories, voter registrations and property deeds [9]. Among others, security, identity theft, incorrect computer records, credit rating destruction privacy, online purchasing and banking, loss of identity, misuse of personal identity, phishing, identity cards, and behavioral monitoring and tracking are current concerns. Security and a loss of control are the current major concerns [10].

Many tools, such as search engines, have been created to turn the Web into more visible and accessible platform but today users are requesting tools and features to provide the right identity in the each context and have control over identity, particularly making identity less visible. People feel concerned and worried about security and/or privacy, but making identity less visible within networked computing ecosystems would establish trust and foster collaboration. "Elements of security in computing begin with identity" [11]. For instance, EBay community of trust lays on users' reputations and Google Web history tool provides more users' control over identity. In this article, we aim to provide a path, several clues, and an implementation towards responding to the vision: how to make digital identity less visible and reduce user's loss of control over it? This would foster trusted partnership and collaboration across networked computing ecosystems.

The article is organized and structured as follows. We introduce the need of collaboration between networked computing ecosystems and associated risks. In the

second section, we present a literature review about definitions, basic concepts and foundations of identity and digital identity. In the third section, we detail loss of control over digital identity as a consequence of its persistence in the context of collaboration. In addition, we explain that making digital identity less visible would reduce such risks. In the fourth section, we highlight added-value, benefits of and possibilities offered by metadata management. We describe our approach that is based on digital identity document metadata in order to contribute making digital identity less visible and give user more control over it. In the fifth section, we propose and explain an XRD-based metadata implementation of digital identity document. Finally, we conclude in the sixth section.

## 2 Identity and Digital Identity: Core Concepts

The notion of identity is evolving over time. Several decades ago, human identity was defined by geography, community, and family relationships. If an individual was born into a well-known and rich family or in a poor remote community, he or she would remain and would typically not be able to change their life pattern or economic status over time. One's geophysical space and one's place in society were inextricably linked and the declaration of an individual's name, sometimes accompanied by the name of their city or village, was sufficient to prove his identity. Today, individuals are having greater choice for participation in different social circles, and more possibilities and freedom of social and economic mobility. In addition, the notion of identity has been extended not only to humans, but animals, machines, organizations, devices, and other objects or resources. We employ the term 'subject' to refer to an individual or a machine to which an individual has delegated a task [12].

### 2.1 Digital Identity Definition

A number of definitions of digital identity have been suggested in the literature. Digitalization is allowing several digital representation of reality, including that of identity. The author [6] calls multiple identities or 'personas' that the subject holds as digital identity 'perspectives' or 'views', which represent different perspectives on who is the subject is and what attributes he processes. They represent also a set of attributes that other entities have and can access to. For instance, a bank sees account attributes and a physician in a hospital sees health record attributes. Digital identity is seen as an intersection of identity and technology in the digital age [5, 6, 13]. More specifically, it represents the data that uniquely describes a person and the data about the person's relationships to other entities. For instance, a car title contains an identification number that uniquely identifies a car to which it belongs and other attributes such as year, model, color and power. The title contains also relationships such as the set of car owners from the time it was made [6]. In OECD report [14] digital identity is defined as 'a thing or an artifact that refers to a person'. Adam's speech and Adam's ID card are two claims of the same person. With the emergence of social networks within participative Web, the author [15] highlights the social side of identity and points out that digital identity is a digital representation of an individual or a machine that presents across all the digital social networks and spaces,

such as avatar, profile or pseudo. The lexicon [16] defines digital identity as “a representation of a set of claims made by one party about itself or another data subject”.

## 2.2 Digital Identity Perceptions and Challenges

Digital identity is perceived differently and faces multiple challenges: a) more people are becoming digital natives and their perception of identity reflects the current reality. Digital Natives, who were born after 1980 and have skills to use networked digital technologies, live much of their lives online and they don't distinguish between online and offline. Instead of thinking of their digital identity and their real-space identity as separate things, they are maintained simultaneously and closely linked to one another. The multiple representations of themselves inform the overall identity; b) new paradox that faces identity in the digital world. Digital identity becomes, in parallel, more dynamic and more persistent. A person living in a remote village during the agrarian age could change many aspects of his personal identity as he wished such as choosing different clothes, expressing himself in a new way, and developing new habits and interests. He has not been able to control her social identity completely because family's status and gossip among neighbors could affect it. The person could change parts of social identity by associating with different people and adjusting social relationships but fellow villagers might still recall earlier versions of it. If he wanted to change or abandon aspects of social identity quickly, he can go beyond the small community where he grew up. Moving to a nearby village, there would likely still be some people who knew him, or knew of him through others and tell stories about him. In the agrarian age, it was possible for the person to completely abandon old social identity and cut off friends and family for good, if he were willing to travel far enough to another city whose inhabitants had little communication with the residents of the town in which he had previously lived. The advent of Internet and digital technologies added new degrees of permanence to identity. He would not be able to change his identity in a complete fashion. A photo of him, with a Photoshop-designed tattoo on his arm, posted in a blog could mark his identity in a persistent way; c) internet does not affect identity. A personal identity today is not that different from what it would have been in the past thus, the digital environment is simply an extension of the physical world. However, in the digital age, social identity may be slightly different from what it would have been in previous ages. Social identity may be shaped by associations that are visible to onlookers at any moment, such as connections in social networks or blogrolls in blogs; and d) identity and digital identity are referential and partial. They are referential because claims must refer to a person; and partial because partial identity refers to a subset of identity information sufficient to identify a person at different moments in time such as nyms, masks or aliases. In either real or digital worlds, a person has multiple identities and partiality is an integrated part of the identity [3, 14, 17, 18]. We consider the use of plural, digital identities, when referring to digital identity of several people. In addition, we align with the distinction, made by [3], between digital identity and digital dossier. Considered as a subset of the digital dossier, digital identity is composed of all attributes that have been disclosed to third parties, whether it is by choice or not. The digital dossier comprises all the personally identifying

information associated with a person, whether that information is accessible or not, and whether it is disclosed to third parties or not. A person's MySpace profile set visible to anyone is part of both digital identity and digital dossier. The medical record held by a doctor is a part of the dossier, but not part of identity because only limited number of people can access it, such as patient, doctor, insurance company and pharmacist.

### 3 Digital Identity Persistence and Loss of Control

Digital identity infrastructure becomes one of the major needs for networked computing ecosystems' collaboration. Building digital identity infrastructure empowers a community of trust. At the corporate level, identity infrastructure should provide security so that interactions with customers, partners, employees, and suppliers become more flexible and richer. The business should not be limited to just transactions, but about relationships with customers, employees, suppliers, and partners and digital identity tends to change this relationship from one-way to a more customized one [6, 11, 15].

Internet users are increasingly losing control over digital identity. They are leaving online trails when browsing the Web and disclosing more personal information, on which many service providers depend. Digital identities are considered as a raw material for social-networking sites. Spock.com is offering people search engine services that would help to find people on the web and more specifically people who have profiles on social networks Live Spaces, Friendster, Hi5, MySpace, and Wikipedia. Spock's mission is to aggregate the world's people information and make it searchable. It is devoted to finding, indexing and profiling people on the Internet. Moreover, Spock provides to people tagging capabilities that could compromise reputations on the internet. Digital identities and user profiles allow to individuals accessing online services and for this reason they become valuable assets. Personal information can be found on websites and in publicly accessible databases. There is more than enough information for an unscrupulous criminal to take over people identity. Companies are using systems that analyze public records such as city's registry, credit files and the register of births, deaths and marriages to build a complete picture of a user online digital footprint. The systems can also analyze the content of social networks to build up a picture of the user relationship to other people. Companies are using applications of semantic tools, designed to bring meaning to large amounts of data [19].

Maintaining control over digital identity fosters collaboration within networked computing ecosystems. "This tension between individuals' interest in protecting their privacy and companies' interest in exploiting personal information could be resolved by giving people more control. They could be given the right to see and correct the information about them that an organization holds, and to be told how it was used and with whom it was shared" [20]. In his book [21], the author points out disclosing and processing personal information through computing ecosystems will be unforgettable like an etched tattoo. He is questioning: "should everyone who self-discloses information lose control over that information forever, and have no say about whether

and when the Internet forgets this information? Do we want a future that is forever unforgiving because it is un-forgetting?” He argues that making identity information less visible, or giving “the right to be let alone” [22], is an efficient way to provide user’s control over identity and revive the forgetting in un-forgetting ecosystems. In his book [2], the author explains the need differently. He argues that the word “trash” implies the remnants of something used but later discarded. It always contains traces and signatures of use such as monthly bills, receipts, personal papers, cellophane wrapping, price tags, and spoiled food. He stresses that future avant-garde practices will be those of trash and nonexistence, which is how does one develop techniques and technologies to make somebody unaccounted for? He illustrates with the example of laser pointer that can blind a surveillance camera when the beam is directed at the lens and as a consequence, the individual is not hiding but simply nonexistent to that node. In the next section, we present an approach based on the use of metadata to make digital identity less visible and therefore gives the subject more control over it.

## **4 Digital Identity Document and Metadata**

### **4.1 Common Use of Metadata**

Metadata, information about information, called also “hidden data” [23] are being democratized and used for various purposes. From antiquity metadata have been created to codify knowledge and classify library materials in the goal to be more accessible. The library classification system in Chinese imperial library is a shining example of metadata usage. As information become more abundant, the main problem is no longer finding it but accessing it easily and quickly. Today, by aiming to organize the world’s information, Google is adding metadata e.g. indexes and PageRank scores when crawling and indexing Web pages. With the advent of Web 2.0, Web users tag web sites, documents, photos and videos helping to label unstructured information so it can be easily found through folkinds such as Delicious URL[24], Diigo URL[25], and Technorati URL[26]. Metadata is becoming a lucrative business opportunity since many companies and consumers are taking advantage of Amazon’s popularity stars, bar codes and RFID labels. Photos uploaded to the website Flickr contain metadata such as when and often where they were taken, as well as the camera model, which could be useful for future buyers [7, 23].

### **4.2 Motivations of Metadata Adoption**

Many motivations and reasons encouraged us to choose metadata as a mean to reduce digital identity loss of control: a) metadata are easy to produce become increasingly available. Ubiquitous and digital technologies are increasingly producing metadata and this will probably fuel the development of metadata management capabilities; b) digital identity document would provide a tracking and responsibility assign capabilities. Word processing and Adobe reader/writer manage metadata related to document updates tracking such as deleted passages, revision numbers and comments. This would help to know what is, when and who added, updated, and deleted digital

identity attributes; and c) metadata usage is a common practice, particularly in Web programming. Web developers and search engine optimization experts are enriching the <meta> tag in the head of Web sites.

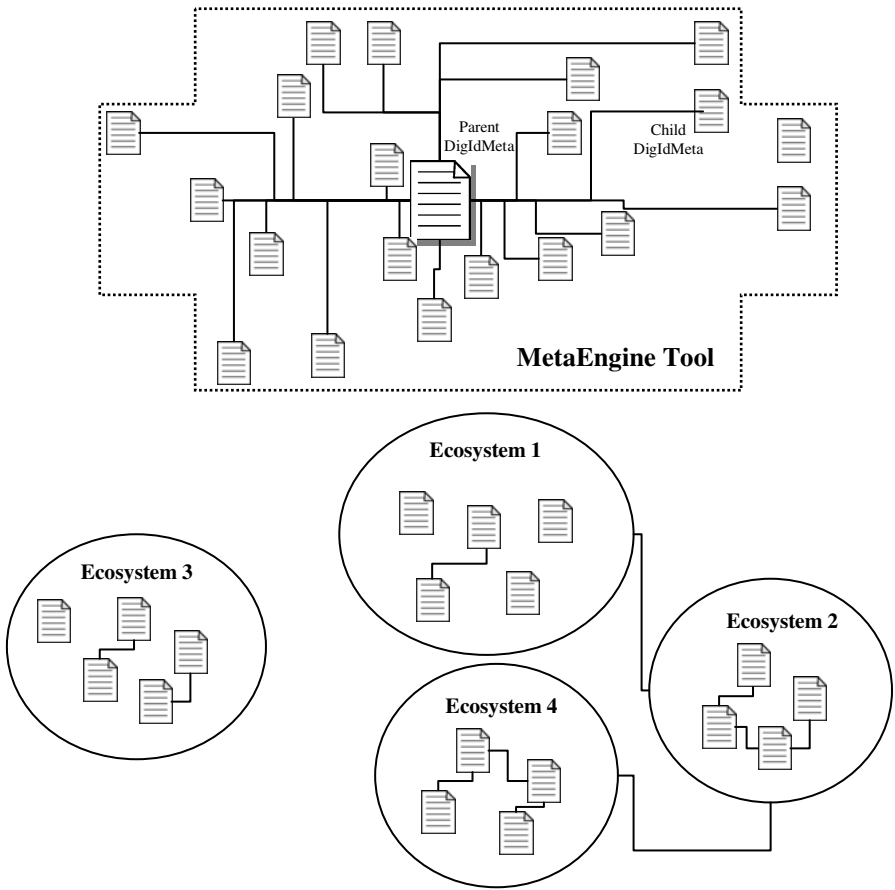
### 4.3 DigIdDoc and MetaEngine

Networked ecosystems collaboration requires processing and exchange of digital identity attributes. Since digital identity is evolving over time, making less visible old or infrequent used one would contribute providing more control to subjects behind computing ecosystems over their identity and encourage networked digital ecosystems community of trust. We consider digital identity as a document (DigIdDoc) comprising attributes' values, a photo, shopping details, etc. because most IDs are given in form of document such as driver's license and Facebook profile. We call a duplicated DigIdDoc is an updated DigIdDoc identity in which attributes' values have changed.

In opposition to the common metadata usage to make digital identity documents more accessible, visible and organized such as casino's data fusion algorithms [23] and Microsoft's MyLifeBits research project [1], we propose to use digital identity metadata (DigIdMeta) toward making persistent digital identity less visible. Figure 1 shows four ecosystems in order to represent multiple distributed computing ecosystems. Behind each eclipse a subject. Links between ecosystems represent an active and constant need of collaboration across different computing ecosystems, such as operations of digital identity aggregation and profiling or a persistent link between two DigIdDocs residing in different ecosystems. Ecosystem 3 is isolated and is not linked to reflect a reality of a person who has a limited set of DigIdDocs. Such as a person who is not using computing ecosystems or striving to conduct anonymous activities. Documents residing inside the eclipses represent DigIdMeta documents attached to DigIdDocs, which are not represented in the figure. At the intra-ecosystem level, the composite DigIdMeta documents could be either linked to each other's or a sub-set of them are linked. The DigIdMeta link could represent the use of the same subject's account to access two or more services such as Google mail and YouTube.

Brain's forgetting mechanism is inspiring research on making digital identity less visible. Researchers are closely studying how the brain forgets information that is stored in long-term memory. Some think that when we forget means that we have lost the link to that information like Web pages URLs. Others reckon and suggest that our brain constantly reconfigures our memory and they say that what we remember is based, at least in part, on our present preferences and needs. Empirical research seems to support the second ideas [21]. Both ideas inspired us to consider adopting an engine that will provide DigIdDoc search, synchronizing, and refresh capabilities. The engine functions could remind a rubber bulb of blood pressure sphygmomanometer. Instead of pushing/pulling air, it will pull DigIdMeta documents from multiple data sources and push them to computing ecosystem's requester. As a result, the latter would receive a specific number of DigIdDocs ordered on a priority basis like any keyword search engine result.





**Fig. 1.** DigIdMeta and MetaEngine Tool

The DigIdDoc priority order is calculated based on the `weight_score`, which is an output of the function, that combines two other scores: `grain_score` and `distance_score`, as follows.

```
Function WeightScore (input grain_score, distance_score):
output weight_score.
```

Whenever a computing ecosystem requests a subject’s digital identity, MetaEngine will collect all DigIdMeta associated with subject’s DigIdDocs and push them into a virtual view. This is similar to data aggregation conducted via virtual directory in which collected data are maintained within non physical settings and the virtual view disappears whenever the operation is no longer needed. The collected DigIdMeta are shown inside the discontinuing-line shape. Besides, MetaEngine tool will calculate the grain score for each DigIdDoc, write it in its DigIdMeta and elect the one that has

the highest score to be the parent, or top-level, document, a shadowed one in figure 2. The parent DigIdMeta will be located in the center and surrounded by other child DigIdMeta. This is like a fact table in a data warehouse's star data schema, which is surrounded by dimension tables. Moreover, the MetaEngine tool will include all the links to the surrounded children in the parent's DigIdMeta and the distance score of each link in the child's DigIdMeta. MetaEngine invokes the function WeightScore to calculate the weight\_scores and writes each weight\_score in its associated child DigIdMeta. The parent DigIdMeta has neither a distance score nor a weight. It has the highest grain\_score and the associated DigIdDoc will appear in the top of the search ordered list like a search engine result. Each of the following DigIdDocs on the list will be ordered on the basis of how high the weight\_score. The distance\_score would empower the "forgetting" capabilities. MetaEngine tool would make a specific number of DigIdDocs, which have higher distance\_scores, easy to access comparing to the ones that have a lower distance\_scores. The latter should be hard to retrieve and to be accessed. For instance, low distance\_score will be on the bottom of search result list, the disclosing decision is followed by the subject's communication of his consent, or the ecosystem should request many times in order to access distant DigIdDocs. MetaEngine tool conducts the refresh operation on on-demand basis, whenever the requester asks for DigIdDocs. It aggregates DigIdMeta, synchronizes the duplicates, recalculates the scores, and reorganizes the links. In the following subsections, we present few parameters that could be used to calculate GrainScore and DistanceScore. We do not intend in this article to provide functions' parameters but we present few clouts that could have a direct or indirect impact on the scores. Work in this area is still in progress and will be subject to further publications.

#### **4.4 Function GrainScore (Input $x_1, x_2, \dots, x_n$ ): Output Grain\_Score**

The central DigIdDoc is the document that has the highest relevance score. The grain\_score is to be calculated on the basis of a set of parameters such as activity and popularity rates. Activity rate represent how actively the subject is using the digital identity document. For instance, the subject could be using frequently the Gmail profile/account more than the Yahoo one, thus the activity rate of the latter is lower than Gmail profile. Popularity rate represents how others perceive subject's identity such as a number of user's tags, a number of users' generated bookmarks on a subject's web page, a number of comments in personal blogs, and a number of blogroll links that point the subject's blog.

#### **4.5 Function DistanceScore (Input $y_1, y_2, \dots, y_n$ ): Output Distance\_Score**

Distance\_score is calculated based on multiple criteria. For instance, DigIdDoc expiration date [21] that could be set by the subject, by computing ecosystem's service provider, or dictated by law. In addition, we can consider forgetting probability and elapsed time from DigIdDoc creation date. As much the distance\_score is higher as far is the child DigIdMeta from the parent one.

## 5 DigIdMeta: XRD Implementation to Support MetaEngine

We present, below, an overview of the XRD document structure and an implementation of DigIdMeta document.

### 5.1 eXtended Resource Description (XRD) Document

Recently published as an OASIS standard, XRD is a simple generic format for describing resources. XRD documents provide machine-readable information about resources for the purpose of promoting interoperability, which is an important need for collaboration across systems. The following XML schema fragment defines the XML namespaces, location of the normative XML Schema file for an XRD document and other header information for the XRD schema [27].

```
<schema targetNamespace="http://docs.oasis-
open.org/ns/xri/xrd-1.0"
  xmlns="http://www.w3.org/2001/XMLSchema"
  xmlns:xrd="http://docs.oasis-open.org/ns/xri/xrd-1.0"
  xmlns:ds="http://www.w3.org/2000/09/xmlldsig#"
  elementFormDefault="unqualified"
  attributeFormDefault="unqualified"
  blockDefault="substitution"
  version="1.0">

<import namespace="http://www.w3.org/2000/09/xmlldsig#"
schemaLocation="http://www.w3.org/TR/2002/REC-xmlldsig-
core-20020212/xmlldsig-core-schema.xsd"/>

<import namespace="http://www.w3.org/XML/1998/namespace"
  schemaLocation="http://www.w3.org/2001/xml.xsd"/>

<annotation>
  <documentation>
    Document identifier: xrd-schema-1.0
    Location: http://docs.oasis-open.org/xri/xrd/v1.0/
  </documentation>
</annotation>
...
</schema>
```

XRD provides XML format for describing meta-documents. XRD DigIdMeta document describes properties of the document itself, as well as the relationships with other DigIdMeta documents. XRD DigIdMeta document can be divided into two main sections: 1) document header section that includes a description of the XRD DigIdMeta document itself, such as document's expiration date [21], and XML namespaces; and 2) resource information section, which is divided into two subsections: resource's description and resource's associated links. The document's description subsection includes properties and aliases of the DigIdDoc, and the next subsection lists links to other DigIdDocs. If a requester's ecosystem wants to know

and learn more about the DigIdDoc, identified by an URI, it retrieves its XRD DigIdMeta document. XRD DigIdMeta provides characteristics and attributes enclosed between <property> tags; and the relationships to other DigIdDocs and available associated services within <links> tags [27, 28].

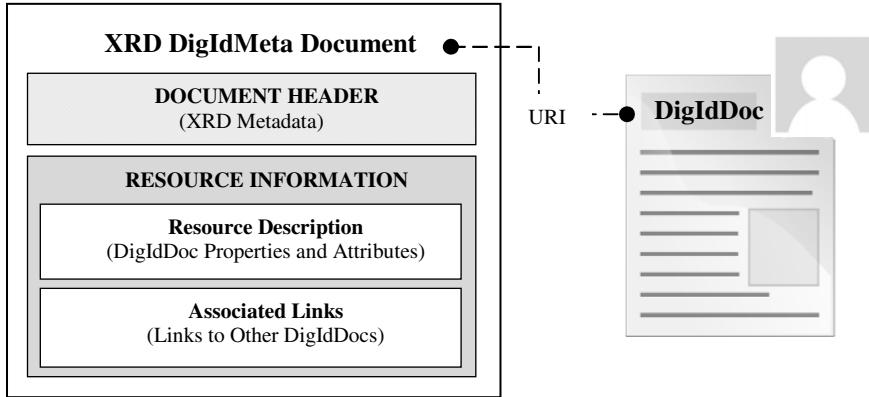


Fig. 2. DigIdDoc and XRD DigIdMeta

XRD DigIdMeta document is bounded to DigIdDoc through either the unique identifier URI or an alias, which is an alternative and human-friendly URI. The <Expires> element defines XRD DigIdMeta document life duration, which could be set by the developer and/or HTTP protocol. The element <property> describes the digital identity document with URI-formatted strings. Finally, XRD DigIdMeta document encapsulates links to other DigIdDocs between <link> tags [27, 28].

## 5.2 DigIdMeta in XRD Format

We present above the XRD implementation of resource information section of the DigIdMeta document. The value between <subject> tags is the unique identifier of the document. Multiple <aliases> could be included to have contextual identifiers and avoiding unique and universal identifier, which could harm privacy. Disk location is enclosed as a property to know the locations of DigIdDoc and its related DigIdMeta document. If the duplicate’s value is set to “Yes” then links to duplicated DigIdDocs are to be added. Subject’s DigIdDoc expiration date [21], recipient’s minimum and maximum expiration dates, and/or legally permissible expiration dates could be either considered as properties in XRD DigIdMeta or as input variables in DistanceScore function. Multiple disclosing dates could be added into the DigIdMeta to ensure a tracking of a few least disclosures. Links to DigIdMeta children are configured by MetaEngine during pulling/pushing operations. DigIdMeta links could add consistency in DigIdDocs search operation and this could be a mean to overcome identity resolution issues associated with having many people with the same full-name.

```

<XRD>
<Subject>http://www.favorite-social.net/gba</Subject>
<Alias> http://www.favorite-socialnet.net/ghazi.benayed
</Alias>
<Alias> http://www.favorite-socialnet.net/ghazibenayed
</Alias>
<Expires>XRD_expiration_date_value</Expires>
<Property type='http://favorite-
social.net/gba/expDate'>DigIdDoc_expiration_date_value</P
roperty>
<Property type='http://favorite-
social.net/gba/location'>DigIdDoc_location</Property>
<Property type='http://favorite-
social.net/gba/duplicate'>Y/N </Property>

// This section is bounded to child's document
<Property type='http://favorite-social.net/gba/gs'>
grain_score_value </Property>
<Property type='http://favorite-social.net/gba/ds'>
distance_score_value </Property>
<Property type='http://favorite-social.net/gba/ws'>
weight_score_value </Property>
<Property type='http://favorite-social.net/gba/cd'>
creation_date</Property>

<Property type='http://favorite-social.net/gba/dd'>
last_disclosing_date </Property>
<Property type='http://favorite-social.net/gba/dexpd'>
discloser_expiration_date </Property>
<Property type='http://favorite-social.net/gba/minrexp'>
min_discloser_expiration_date </Property>
<Property type='http://favorite-social.net/gba/maxrexp'>
max_discloser_expiration_date </Property>

// The Link section is bounded to parent's document
<Link rel='update' type='text/html'
      href='http://favorite-social.net/gba/update'>
  <Title xml:lang='en-us'>Link to Updated DigIdDoc
  </Title> </Link>

<Link rel='duplicate' type='text/html'
      href='http://favorite-social.net/gba/duplicate'>
  <Title xml:lang='en-us'>Link to Duplicated DigIdDoc
  </Title> </Link>

<Link rel='child1' type='text/html'
      href='http://favorite-social.net/gba/child1'>
  <Title xml:lang='en-us'>Link to Child1 DigIdDoc
  </Title> </Link>

```

...

```
<Link rel='childn' type='text/html'
      href='http://favorite-social.net/gba/childn'>
  <Title xml:lang='en-us'>Link to Childn DigIdDoc
</Title> </Link>

</XRD>
```

## 6 Conclusion and Future Work

Digital identity is partial and referential. Partiality is a consequence of context-specific nature of identity. A traveler is asked to provide his passport at the counter of customs or immigration as a proof of his identity and the same person, being a car driver, is asked to show his driving license to a police officer. The context will determine which identity is required to be communicated to other computing ecosystems in order to move forward collaboration. Digital identity is referential because attributes must refer to a subject. We represented partiality through digital identity documents distributed in a computing ecosystem. Referentiality is represented through the subject that is behind each computing ecosystem.

Networked ecosystems collaboration requires processing and exchange of digital identity attributes. Since digital identity is evolving over time, making less visible or infrequent used one would contribute providing more control to subjects behind computing ecosystems over their identity and encourage networked digital ecosystems community of trust. Therefore, we described a metadata management-based mechanism to help making digital identity documents less visible and contribute to give the subject more control over digital identity. The metadata management engine should have write permissions in all digital identity document metadata and data synchronization capabilities in order to establish the virtual view. In the near future, we intend to investigate in more details input parameters of GrainScore, DistanceScore, and WeightScore functions. Moreover, we'll study the opportunity to extend the star data schema into a snowflake one in the virtual view. This data schema could represent more the partiality nature of digital identity.

However, we believe that technical approach is not enough to make digital identity documents less visible, a multidisciplinary and an integrated approach in which we take into consideration several other perspectives, e.g. legal, managerial, user-centricity is needed. Among other perspectives, we notice that besides the drawbacks that accompany the information overabundance age, digital identity could be less visible and increasingly inaccessible in ocean of data. Isn't true that a wealth of information creates a poverty of attention? An exploration, from this perspective, may be worth to consider in the near future.

## References

- [1] Bell, G., Gemmel, J.: A Digital Life, pp. 58–65. Scientific American Magazine (2007)
- [2] Galloway, A.R., Thacker, E.: The Exploit - A Theory of Networks. University of Minnesota Press (2008)
- [3] Palfrey, J., Gasser, U.: Born Digital: Understanding the first generation of digital natives. Basic Books (2008)

- [4] Gardiner, M.: The Business Value of Identity Federation(2007), <http://whitepaper.techworld.com/authentication/4818/the-business-value-of-identity-federation>
- [5] International Telecommunication Union, Digital Life. ITU Internet Report (2006), <http://www.itu.int/osg/spu/publications/digitalife/docs/digital-life-web.pdf>
- [6] Windley, P.J.: Digital Identity: Unmasking identity management architecture (IMA). O'Reilly, Sebastopol (2005)
- [7] Cukier, K.: A special report on managing information. The Economist (2010)
- [8] Facebook hack releases 100 million user details onto filesharing sites. Infosecurity USA (2010), <http://www.infosecurity-us.com/view/11343/facebook-hack-releases-100-million-user-details-onto-filesharing-sites/>
- [9] Fischetti, M.: Scoring Your Identity: New tactics root out the false use of personal data, pp. 27–28. Scientific American (2007)
- [10] Cochrane, P.: Forward of the Book. In: Birch, D.G.W. (ed.) Digital Identity Management: Perspectives on the Technological, Business and Social Implications, Gower Publishing Limited, England (2007)
- [11] Benantar, M.: Access Control Systems: Security, Identity Management and Trust Models. Springer, Heidelberg (2006)
- [12] Noonan, H.: Identity, Stanford Encyclopedia of Philosophy (2009)
- [13] Center for Democracy & Technology. Privacy Principles for Identity in the Digital Age [Draft for Comment - Version 1.4] (2007), [http://www.cdt.org/files/pdfs/20071201\\_IDPrivacyPrinciples.pdf](http://www.cdt.org/files/pdfs/20071201_IDPrivacyPrinciples.pdf)
- [14] Organization for Economic Co-operation and Development (OECD), At Crossroads: Personhood and Digital Identity in the Information Society. The Working Paper series of the OECD Directorate for Science, Technology and Industry (2008), [http://www.oecd.org/LongAbstract/0,3425,en\\_2649\\_34223\\_40204774\\_119684\\_1\\_1\\_1,00.html](http://www.oecd.org/LongAbstract/0,3425,en_2649_34223_40204774_119684_1_1_1,00.html)
- [15] Williams, S.: This is Me Digital Identity and Reputation on the Internet, <http://www.slideshare.net/shirleyearley/this-is-medigital-identity-and-reputation-on-the-internet>
- [16] Identity Gang Group - Working Group of Identity Common. Identity Gang Lexicon, <http://wiki.idcommons.net/Lexicon>
- [17] Damiani, E., et al.: Managing Multiple and Dependable Identities. In: IEEE Internet Computing, pp. 29–37. IEEE Computer Society, Los Alamitos (2003)
- [18] Princeton University Wordnet - Lexical Database for English. Identity Definition, <http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o7=&o5=&o1=1&o6=&o4=&o3=&s=identity&i=1&h=0000#c>
- [19] Fildes, J.: Taking Control of Your Digital ID (2006), <http://news.bbc.co.uk/2/hi/technology/6102694.stm>
- [20] Cukier, K.: New Rules for Big Data: Regulators are having to rethink their brief. The Economist (2010), [http://www.economist.com/specialreports/displaystory.cfm?story\\_id=15557487](http://www.economist.com/specialreports/displaystory.cfm?story_id=15557487)
- [21] Mayer-Schönberger, V.: Delete: The virtue of forgetting in the digital age. Princeton University Press, Princeton (2009)

- [22] Brown, P.: Privacy in an Age of Terabytes and Terror. *Scientific American Magazine*, 46–47 (2008)
- [23] Garfinkel, S.L.: Information of the World, UNITE! *Scientific American Magazine*, 82–87 (2008)
- [24] Delicious, <http://www.delicious.com>
- [25] Diigo, <http://www.diigo.com>
- [26] Technorati, <http://technorati.com>
- [27] OASIS eXtensible Resource Identifier (XRI) TC., Extensible Resource Descriptor (XRD) Version 1.0, OASIS Standard (2010),  
<http://docs.oasis-open.org/xri/xrd/v1.0/xrd-1.0.html>
- [28] Hammer-Lahav, E.: XRD Document Structure (2009),  
<http://hueniverse.com/2009/03/xrd-document-structure/>



# An Overview of Performance Comparison of Different TCP Variants in IP and MPLS Networks

Madiha Kazmi<sup>1</sup>, Muhammad Younas Javed<sup>2</sup>, and Muhammad Khalil Afzal<sup>3</sup>

<sup>1</sup> Dept of Computer Engineering, CE & ME, NUST, Pakistan

<sup>2</sup> Dept of Computer Engineering CE & ME, NUST, Pakistan

<sup>3</sup> COMSATS Institute of Information Technology, Wah Cantt Pakistan  
madihakazmi@yahoo.com, myjaved@ceme.nust.edu.pk,  
khalil\_78\_pk@yahoo.com

**Abstract.** Researchers have shown considerable interest in TCP variants and their behavior under different traffic conditions by conducting research on congestion management of TCP in IP supporting networks. TCP provides a trustworthy end-to-end data transfer under changeable wired networks. To overcome the problem of unreliability of IP network, TCP is used. Many service providers are now moving to MPLS over Inter- net to transfer data, preferring it over traditional transferring strategies. Different variants of TCP show varying behavior in best effort Internet Protocol networks. This paper presents an extensive investigational study of TCP variants under IP and MPLS networks by focusing Tahoe, Reno, New Reno, Sack and Vegas under File Transfer Protocol (FTP).

**Keywords:** MPLS, LDP, LSP, RSVP, FTP.

## 1 Introduction

In transport layer protocol, TCP is the most popular protocol. TCP provides in sequence deliverance of data and an unflinching data transmission among communicating nodes. One of the strengths of TCP is its high responsiveness toward network congestion. TCP is also a defensive protocol as it detects incident congestion and in result to that it tries to lessen the impacts of this congestion, which will prevent collapse of communication.

Nowadays' Internet communication is carried to a large extent using TCP, and as a result a lot of researchers are concentrating on modeling and understanding it on different parameters i.e. time to transmit a file and network consumption. TCP is the fastest growing protocol even in future and we are presenting a comparative study of different TCP variants in MPLS and IP domain. For analytical results of the proposed solution; demonstration of the IP and MPLS network is simulated over a limited number of nodes. The flow of descriptors to maintain network topology determines average delay, throughput, variance of delay, packets sent, packets received, packet dropped. Conclusions are drawn on the basis simulation results, while comparisons

between them have been elaborated. Organization of rest of the paper is as follows: The Section 2 briefly describes the TCP variants. A summary of the work already done in the field of TCP variants, IP and MPLS is presented in the third section. The Section 4 introduces MPLS network. In section 5, we present the simulation result followed by their interpretation. Finally we present the analysis on simulation result.

## 2 Tcp Variants

A brief description of TCP variants is given below. More details of TCP Variants can be found in [13]and [14].

- **TCP Tahoe:** TCP Tahoe has the method to pay compensation for the efficiency plunge caused by congestion after packets are dropped. Tahoe is the very first variant of TCP that uses three mechanisms to organize the flow and handle congestion that is congestion avoidance, slow start, and fast re-transmit [13].
- **TCP Reno:** In 1988 by V. Jacobson proposed a variant of TCP that is typical implementation of TCP protocol, it includes the congestion control algorithm. TCP Reno uses four distinct mechanisms to control the flow and deal with congestion three are those used by Tahoe and fourth algorithm is termed additive increase multiplicative decrease (AIMD) [13].
- **TCP New Reno:** TCP Reno algorithms are efficient in dealing with single packet lost in a congestion window. But in case of multiple packets dropped, it will retransmit the packet whose duplicate acknowledgment was received leading Fast Recovery phase to finish. TCP Reno will re-enter the Fast Recovery phase when it comes to know that more packets are dropped. Effectiveness of protocol is affected by again and again entering the Fast Recovery as TCP New Reno will stay in this phase until all lost packets are retransmitted. New Reno works on the mechanism of partial ACK [13].
- **TCP Sack:** TCP Tahoe, Reno, and New Reno all acknowledge cumulative packets therefore are unable to detect multiple lost packets per round trip time. TCP SACK's selective acknowledgements algorithm deals effectively with multiple packets lost [13].
- **TCP Vegas:** TCP VEGAS detects congestion before it really occurs and follows the AIMD paradigm. TCP Vegas switches to congestion avoidance phase as soon as it senses an early congestion by keep on calculating the difference of current and expected throughput [14].

## 3 Related Work

One of performance comparison research is conducted in [1], that focuses on performance evaluation of certain variants of TCP protocol over IP and MPLS network . Another research was conducted to examine various variants of TCP on two types of traffic, i.e. FTP and Telnet. In Universal Mobile Telecommunications System (UMTS) network there is a significant impact of different type of traffic on as a whole performance of TCP. [2]

Zhong Ren et. al. in [3] integrated mobile IP and MPLS networks. Techniques for controlling and signaling this integration are argued in detail, it also points out some scalability issues of Mobile IP. A similar sort of study is performed by M. Asante in [4] by analyzing the mobile IP and MPLS union architecture. This paper highlighted benefits of this union.

Wierman et. al. [5] gave a framework for analyzing TCP variations i.e. Vegas, Sack and Reno. He analyzed that induced slow start algorithm of Vegas do not help to reduce packet loss but this algorithm wastes a lot of time in slow start phase.

Mazleena Salleh et.al. [6] Compared TCP Tahoe, NewReno, Vegas, and Sack over self-similar traffic. They found that NewReno did better than other TCP variants with respect to efficiency and throughput. TCP Vegas showed better throughput than Reno.[7]

Jeonghoon, et. al. [8] results emphasize on former discussed research results. Go Hasegawa et.al. [9] compared performance of Reno and Vegas sharing bottleneck link on Internet found out Reno to be a better performer. Similar results were concluded by Cheng P. Fu et.al. [10] where they compared performance of Reno and Vegas on asymmetric networks having bottleneck.

Thomas Bonald in [11] compared Reno and Vegas keeping RTT measurement as testing template. They focused on long-term performance criterion i.e. average throughput and average buffer taken up.

Yi-Cheng Chan in [12] has reported a few problematic sides of TCP Vegas while congestion avoidance, which makes its less successful. To trim down impact of Vegas problems authors have also presented congestion avoidance scheme based on router.

## 4 MPLS Network

The unstable growth of the Internet and the introduction of complicated services require an epoch-making change. MPLS was proposed as an alternative. MPLS is a protocol specified by Internet Engineering Task Force (IETF). MPLS provides many services through networks i.e., routing, effective/efficient designation, forwarding of packets and traffic flows switching. The most salient functions of MPLS is to supervise the traffic flows among heterogeneous applications, hardware and machines. MPLS is not reliant on Layer-2 protocols and Layer-3 protocols [15].

Md. Arifur Rahman et. al. in [17] showed the superiority of MPLS over conventional networks. Their Research concluded better throughput and lesser delay on MPLS network, due to its better traffic engineering principles.

## 5 Simulation and Analysis

Simulation has 13 nodes in total divided into IP and MPLS domains. The 7 nodes in MPLS domain are named as LSR1 to LSR7, while there are 6 nodes in IP domain labeled as node0 to node5. The IP domain consists of a sender and a receiver network. Both sender and receiver network having three nodes each. The bandwidth between nodes of IP and MPLS network is set 1 MB with a 5ms link processing delay. All

MPLS enable nodes provides mechanism for label distribution as they are Label Distribution Protocol (LDP) enabled.

The traffic in FTP having packet size of 1500 KB having varying time interval. The simulation runs for 100 seconds. Both source and destination networks are IP-based. Some common networks parameters are revealed in Table 1.

**Table 1.** Network parameter

Network parameters	Values
Number IP nodes	6
Number of MPLS nodes	7
Number of hops	7
Link processing Delay	5ms
Packet size	1500
Bandwidth	1MB

The different TCP variants in MPLS/IP network are analyzed using different scenarios. The numbers of flows of the link were varied to check the effect of different flows on the delay and throughputs. These traffic are run in FTP.

1. Single Traffic
2. Multiple Traffic
  - (a) Two flows Traffic
  - (b) Four flows Traffic

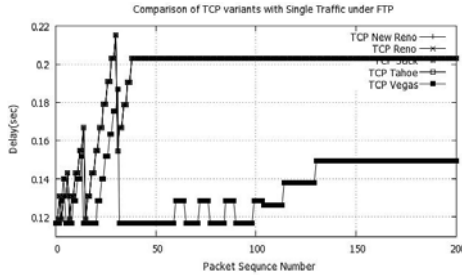
FTP traffic has been run on single and multiple flows and results are recorded thereby. These results are then presented in structure of tables and figures for TCP New Reno, TCP Reno, TCP Sack, TCP Tahoe and TCP Vegas. A FTP connection is set up between node0 and node1 and this simulation is executed for New Reno, Reno, Sack, Tahoe and Vegas. Table 2 shows the values of these simulation average delays in milliseconds. The percentage throughput of different variants on single flow and it can be examined that all variants are giving 100% throughput.

**Table 2.** Delay of TCP Variants on Single Flow FTP

Protocol	Delay (ms)
TCP New Reno	202.551
TCP Reno	202.551
TCP Sack	202.551
TCP Tahoe	202.551
TCP Vegas	149.270

On FTP flow there is no packet loss. Average delay of all the variants is more or less alike and TCP Vegas shows 25% lesser average delay than other variants. The accomplishments of TCP variants can also be observed by recording the delay of

packets. i.e. average delay of simulation traced until 50, 100, 150 and 200, as in Fig.1.



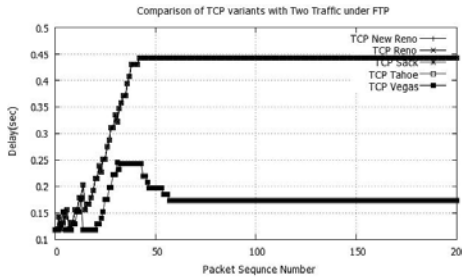
**Fig. 1.** Comparison of TCP Variants with Single Trafcs under FTP

**Table 3.** Delay of TCP Variants on Two Flows FTP

Protocol	Flow1(Delay in ms)	Flow2(Delay in ms)
TCP New Reno	440.777	441.031
TCP Reno	440.777	441.031
TCP Sack	440.777	441.031
TCP Tahoe	440.777	441.031
TCP Vegas	173.062	173.152

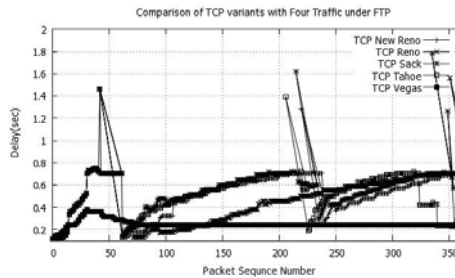
The Table 3 is about average delay in milliseconds on same topology. In this model of FTP flow there is no packet loss. Average delay of all the variants is roughly similar but TCP Vegas shows 40% lesser average delay than other variants. In other words, TCP Vegas performs superior on two flows than other variants keeping average delay under discussion.

Fig.2 shows the behavior of TCP variants observed by recording the delay of the packets plotted with their sequence numbers i.e. average delay after packet number 50, 100, 150 and 200, sent by every variant, is traced.



**Fig. 2.** Comparison of TCP Variants with Two Traffics under FTP

Fig. 3 illustrates the behavior of TCP variants observed by recording the delay of packets on basis of their sequence numbers i.e. average delay of packets after transmission of 50,100,150 and so on packets for every variant under four flows of FTP. It gives a glimpse of behaviors of TCP variants by plotting average delay of packets in seconds across y-axis. Each sharp edge shows the abrupt change in delay of that particular variant. Vegas shows a little jitter in the beginning but later on illustrates a smooth performance, this show that the delay of the packets remains the same throughout simulation time. There is no packet loss in TCP Vegas. TCP Reno has highest delay reaching 1.8 second, until the end of simulation Reno is giving highest peaks of delay.



**Fig. 3.** Comparison of TCP Variants with Four Trafcs under FTP

- Error rate induction in FTP Single Flow

TCP provides reliable end to end transport layer protocol. Because of this feature this a protocol used by 90% traffic on internet approximately. Congestion avoidance in TCP allows the application to increase by one packet whenever an acknowledgement is received; allowing full utilization of available bandwidth.

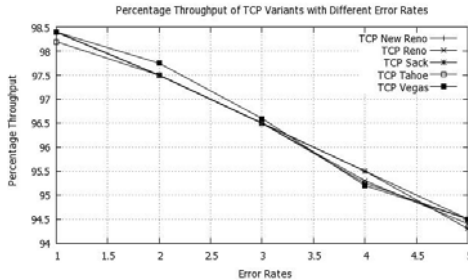
Most important feature of TCP is its Congestion Control strategy. In wired network, whenever there is a packet loss, it indicates that network is congested. Inducing error rate explicitly shows the behavior of different TCP variants, as some variants reduce congestion window unnecessarily. Performance of TCP is affected by various factors like link capacity, RTT, random losses, short flows etc.

As IP is an unreliable network, adding TCP to it is for providing reliability via sliding window scheme, ACK, sequence number and control flow to avoid overflowing of receiver buffer [16].

TCP Tahoe is the scheme of TCP that deploys slow start mechanism to prevent the problem of congestion. It is a reactive mechanism. New Reno is an active variant of TCP which is used for multiple packet losses. It provides the solution for oscillating congestion window to resolve problem faced by TCP Reno. For the solution to the problems of TCP's inability to tell about the multiple packet dropping, TCP Sack was proposed. TCP Vegas is the proactive variant of TCP that anticipates the intended congestion on the basis of round trip times of the data packets.

Fig.4 gives an idea about percentage throughput achieved by TCP variants. There is significant deteriorates in throughput of variants with increase in the probability of packet loss (1 to 5%). Reason behind this decline is that with every lost packet

frequency of dropped packet increases and results in numerous drop and time outs. When random packet loss was introduced, throughput of all variants remained the same i.e. 98, 97, 96, 95 and 94 percent for 1, 2, 3, 4 and 5 percent error rate respectively. Vegas gave lowest end-to-end delay till 3% error rate but dramatically TCP Sack gave delay even lower than Vegas for 4 and 5% error rate.



**Fig. 4.** Percentage throughput of TCP variants with different Error Rates

The congestion window of all the variants rapidly goes down to the smaller value. New Reno had largest value of average congestion window in 1% error rate case. But, Vegas had largest value and Tahoe had smallest value of average congestion window throughout the experiment scenarios, this shows Tahoe's inferiority to Vegas.

## 6 Conclusion and Future Work

TCP is debatably the most significant protocol in the internet today. Congestion control algorithm is special feature of TCP. TCP try to achieve the best bandwidth rate vigorously on any network. It keeps on pushing high transfer rate continuously. It also reduces this transfer rate on detecting errors from time to time. Observing the behavior of TCP is quite a revealing experience about behaviors of different variants of TCP on IP and MPLS network. It is evident that as loss rate increases, throughput decreases. Similarly congestion window size also decreased. Future research will address the Fast Send Protocol's (FSP) implementation to minimize time to transfer data in high-speed bulk to improve the transfer rate [18].

## References

1. Akbar, M.S., Ahmed, S.Z., Qadir, M.A.: Quantitative Analytical Performance of TCP Variants in IP and MPLS Networks. In: IEEE Multitopic Conference, INMIC (2006)
2. Dubois, X.: Performance of Different TCP Versions Common/Dedicated UMTS Channels. Master Thesis, University of Namur (2005)
3. Ren, Z., Tham, C.-K., Foo, C.-C., Ko, C.-C.: Integration of Mobile IP and MPLS. In: IEEE ICC, Finland, vol. 7, pp. 2123–2127 (June 2001)
4. Asante, M., Sherratt, R.S.: Mobile IP Convergence in MPLS-based Switching. In: Proceeding of Wireless and Optical Communication MultiConference (July 2004)

5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an OpenGrid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. Wierman, A., Osogami, T., Olsen, J.: A Unified Framework for Modeling TCP-Vegas, TCP-SACK, and TCP-Reno. In: Proceedings of MASCOTS (2003)
7. Ols' en, J.: Stochastic Modeling and Simulation of the TCP Protocol. PhD thesis, Department of Mathematics, Uppsala University, Sweden (October 2003)
8. Mo, J., La, R.J., Anantharam, V., Walrand, J.: Analysis and comparison of TCP Reno and Vegas. In: Proc. INFOCOM 1999, New York, vol. 3 (March 1999)
9. Fu, C.P., Liew, S.C.: A Remedy for Performance Degradation of TCP Vegas in Asymmetric Networks. In: Communications Letters, vol. 7, pp. 42–44. IEEE Computer Society Press, New York (2003)
10. Thomas, B.: Comparison of TCP Reno and TCP Vegas via Fluid Approximation", Available as INRIA Research Report (November 1998)
11. Chan, Y.-C., Chan, C.-T., Chen, Y.-C., Ho, C.Y.: Performance Improvement of Congestion Avoidance Mechanism for TCP Vegas. In: Proceedings of the 10th International Conference on Parallel and Distributed Systems (ICPADS 2004), IEEE, New York (2004)
12. Xu, K., Tian, Y., Ansari, N.: Improving TCP performance in integrated wireless communications networks. In: Computer Networks, pp. 219–237. Elsevier, Amsterdam (2005)
13. Xu, K., Tian, Y., Ansari, N.: Improving TCP performance in integrated wireless communications networks. Computer Networks 47, 219–237 (2005)
14. Kim, D., Cano, J.-C., Manzoni, P.: A comparison of the performance of TCP-Reno and TCP-Vegas over MANETs. In: C-K. Toh. IEEE, New York (2006)
15. Rosen, E., Viswanathan, A., Callon, R., et al.: Multiprotocol Label Switching Architecture. IETF Internet Draft (August 1999)
16. Floyd, S.: Issues of TCP with SACK, Tech. Report (Cited on pp. 6 and 58.) (January 1996), <http://www.icir.org/floyd/sacks.html>
17. Rahman, M.A., Hassan, Z., Kabir, A.H., Lutfullah, K., Amin, M.R.: Performance Analysis of MPLS Protocols over conventional Network. In: Microwave Conference, China (2008)
18. Koch, C., Rabl, T., Holbling, G., Kosch, H.: Fast Send Protocol - Minimizing Sending Time in High-Speed Bulk Data Transfers.



# Routing in Mobile Ad-Hoc Networks as a Reinforcement Learning Task

Saloua Chettibi and Salim Chikhi

MISC Laboratory, Computer Science Departement,  
University of Mentouri, 25000 Constantine, Algeria  
{sa.chettibi, slchikhi}@yahoo.com

**Abstract.** Communicating nodes in Mobile Ad-hoc NETWORKS (MANETs) must deal with routing in an efficient and adaptive way. Efficiency feature is strongly recommended since both bandwidth and energy are scarce resources in MANETs. Besides, adaptivity is crucial to accomplish the routing task correctly in presence of varying network conditions in terms of mobility, links quality and traffic load. Our focus, in this paper, is on the application of Reinforcement Learning (RL) technique to achieve adaptive routing in MANETs. Particularly, we try to underline the main design-issues that arise when dealing with adaptive-routing as a Reinforcement Learning task.

**Keywords:** Mobile Ad-hoc Networks, Routing, Reinforcement Learning.

## 1 Introduction

Wireless networks mainly break into two sub-classes: infrastructure-based and infrastructure-free networks well known as ad-hoc Networks. In its mobile configuration, the ad-hoc network is called MANET (Mobile Ad-hoc NETWORK). In MANET, nodes can randomly join or leave the network and new links appear or disappear accordingly. Furthermore, the wireless medium is rarely stable and can be easily congested due to the limited bandwidth. Besides, mobile nodes are battery-powered and may fail at any time. Consequently, network topology changes constantly and unpredictably which complicate the routing task.

To deal with constant changing network conditions in terms of mobility, link quality, available energy-resources and traffic load, a routing protocol for MANETs should be adaptive. To design such adaptive protocols, techniques from the field of artificial intelligence have been adopted. Particularly, ACO meta-heuristic, which is a subclass of SI (Swarm Intelligence) algorithms has made the foundation of the majority and the most significant contributions to adaptive routing problem. Thanks to constant path-probing using ants agents, statistical estimates of paths quality are learned and good routing decisions are reinforced. More recently, reinforcement learning has also taken place as an appropriate framework to design routing policies which have the ability to be adapted by trial and error. We have surveyed in a previous paper [1] many routing protocols for MANETs that apply reinforcement learning either to learn routing decisions (i.e. choosing next-hop or path for routing)

or to learn some routing parameters rather than to fix them experimentally. Our focus, in the present paper, is on the modelization of routing-decision making problem in MANETs as a reinforcement learning task. To do so we have selected works that gives the explicit formalization of the routing problem as a MDP or a POMDP. Our ambition is to highlight the main design-challenges that must be addressed when using the RL framework for routing in MANETs.

The remainder of this paper is organized as follows: design issues of routing protocols for MANETs are outlined in section 2. Section 3 introduces briefly the RL framework. Next, in section 4, we describe different RL-models for routing problem in MANETs. In section 5, we conclude the paper by highlighting the emerging issues and challenges when dealing with routing problem in MANETs as a RL-task.

## 2 Routing Issues in MANETs

Required features of routing protocols for MANETs can be summarized as follows:

**Adaptivity.** A routing protocol for MANETs should be adaptive in face of frequent and unpredictable changes in network topology mainly due to wireless links instability and to nodes mobility and failure after their batteries depletion. Moreover, adaptivity in face of changing traffic loads is important to avoid congestion areas in the network.

**Robustness.** Control and data packets can be lost due to the poor quality of wireless connections and to the interference between simultaneous transmissions. Robustness is a crucial feature to keep the routing protocol operating correctly even when such losses occur.

**Efficiency.** Efficiency is important to deal with bandwidth, processing power, memory and energy limitations in MANETs. A Routing protocol should be efficient by optimizing its exploitation of network resources.

**Scalability.** In many deployment scenarios of MANETs, network size can grow to very large sizes. Hence, scalability should be taken in consideration in routing protocols design.

In MANETs literature, several routing protocols that try to streak a balance between two or more among the abovementioned features have been proposed. However, one common deficiency is in assuring the adaptivity feature. This later is generally neglected especially when network topology is assumed to be random or when network-links are considered to be either functional or not [2]. Moreover, using fixed thresholds as routing parameters rather than adjusting them in function of varying network conditions contradict the dynamic nature of MANETs [3][4]. In fact, this deficiency was the main motivation for researchers to exploit artificial intelligence methods such as reinforcement learning to enhance routing protocols adaptivity as will be shown later in this paper.

### 3 Reinforcement Learning

In reinforcement learning, an agent seeks to learn how to make optimal decisions (actions) at each environment-state in a trial and error fashion. Following an action, the environment generates a reinforcement signal that the learning agent exploits to update its current policy. RL- problems are usually modeled as Markov Decisions Process (MDPs) [5]. A MDP is a tuple  $\langle S, A, R, P \rangle$  where:  $S = \{s_1, s_2, \dots, s_N\}$ : a set of  $N$  states;  $A = \{a_1, a_2, \dots, a_M\}$ : a set of  $M$  actions;  $R: S \times A \rightarrow \mathbb{R}$  is the reinforcement function and  $P: S \times A \rightarrow \Pi(S)$  is the state transition distribution function. When the learning agent has not an accurate and a complete view about the environment state then a partially observable MDP (POMDP) arise. This later can be defined as a tuple  $\langle S, Z, A, O, R, P \rangle$  where  $S, A, R$  and  $P$  maintain the same significance as in the MDP definition;  $O: S \times A \rightarrow \Pi(Z)$  is the observation function where  $Z$  is the set of observations.

RL-algorithms that solve RL-problems can be classified into model-based and model-free algorithms [5]. In a model-based method, the environment model is learned first and the optimal policy is calculated later whereas a model-free method learns directly from experience without building any environment model. According to if the RL-algorithm bootstraps or samples we can distinguish MC (Monte Carlo) and TD (Temporal difference) methods [5]. In fact, MC methods are based on averaging sample returns whereas TD methods combine both sampling and bootstrapping.

## 4 Routing Problem in Mobile Ad-Hoc Networks as a Reinforcement Learning Task

Indeed, the first demonstration that network packet-routing can be modeled as a reinforcement learning task is the Q-routing protocol [6] proposed for fixed networks. The authors claim, in the paper [6], that a packet-routing policy answers the question: to which adjacent node should the current node send its packets to get as quickly as possible to its eventual destination? Therefore, the routing problem can be modeled as a RL-task where the environment is the network, the states are the nodes and the learning agent's possible actions are the next-hops it can take to reach the destination. In the following subsections, we describe some MDP/POMDP already proposed in the literature to formalize the routing problem in MANETs.

### 4.1 Mobility Aware Q-Routing

In reference [7], straightforward adaptation of Q-routing to the context of MANETs was proposed. Indeed, the same RL-model of Q-routing is maintained always in the perspective of optimizing packets delivery time:

**States.** In a very intuitive way the set of states is the set of mobile nodes in the network.

**Actions.** At a node  $x$ , available actions are reachable neighbors. A node learns how to choose a next-hop to forward its traffic.

**Reward.** Local information is used to update the routing policy of a source node  $x$ . This includes the min Q-values of its neighbors and the time the current packet spent on the queue,  $b_t^x$ , at node  $x$  before being sent off at period time  $t$  as shown in equation (1), where  $0 < \alpha < 1$  is the learning rate:

$$Q_t^x(d, y) = (1 - \alpha)Q_{t-1}^x(d, y) + \alpha(b_t^x + \min_z Q_{t-1}^y(d, z)) \quad (1)$$

**Action selection rule.** a greedy policy is adopted. When a node  $x$  receives a packet for destination  $d$ , it sends the packet to the neighbor  $y$  with the lowest estimated delivery time  $Q_t^x(d, y)$ .

To take care of nodes mobility, two additional rules are proposed for Q-values updates of neighboring nodes:  $Q^x(d, y) = \infty$  when  $y$  moves out of  $x$  range; and  $Q^x(d, y) = 0$  when  $y$  moves into  $x$  range. Note that the second update rule is made optimistic to encourage exploration of new coming neighbors. Finally, it is worth noticing that the same model described above was used in LQ-routing protocol [8]. This latter have introduced the path-lifetime notion to deal with mobility.

## 4.2 SAMPLE Routing Protocol

In SAMPLE [2], the optimization goal is to maximize network throughput. The routing problem was mapped onto a MDP problem as follows:

**States.** Each node  $n_i$  has a set of states  $S_i = \{B, P, D\}$  where B indicates that a packet is in a buffer waiting to be forwarded (start state), P indicates that a packet has been successfully unicast to a neighbor, and D indicates that a packet has been delivered at node  $n_i$  (terminal state).

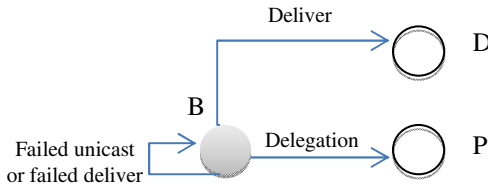


Fig. 1. SAMPLE MDP

**Actions.** The actions available at different states in the MDP are a packet delivery action, a broadcast action to discover new neighbors, links, and routes; and for each neighboring node, a delegation action (unicast).

**Transition model.** A statistical transition model that favors stable links is considered. It acquires information about the estimated number of packets required for a successful unicast as an indication of links quality. In order to build this model, a number of different system events are sampled within a small time window into the past. The monitored events are: attempted unicast transmissions; successful unicast transmissions; received unicast transmissions; received broadcast transmissions; and promiscuously received unicast transmissions.

**Reward model.** A simple static reward model was considered. The rewards are set at values -7 and 1 to model the reward when transmission succeeds under a delegation action and fails, respectively. In effect, these values reflect connection costs in IEEE.802.11 protocol.

**Action selection rule.** Concerning delegation actions, the decision of which next hop to take is chosen probabilistically using the Boltzmann-action selection rule. Variation of temperature value controls the amount of exploration. Furthermore, SAMPLE also uses a simple greedy heuristic in order to restrict exploration to useful areas of the network by only allowing a node to forward to those neighboring nodes with a value that is less than its function value. The discovery action is also allowed with a certain probability in order to explore new routes.

To deal with nodes mobility, authors in [2] have considered Q-values to decay. They suggest configuring the decay-rate to match any estimated mobility model. Note that, the CRL algorithm was used to solve the RL problem of SAMPLE. CRL is a model-based RL-algorithm that extends the conventional RL framework with feedback models for decentralized multi-agent systems. Indeed the same RL-model of SAMPLE was used in SNL-Q routing protocol [9].

### 4.3 RL-Based Secure Routing

Finding secure routes has made the main focus of the RL-based routing protocol proposed in [4]. Learning a policy for selecting neighbors based on their reputation values was simply mapped to a MDP as follows:

**States.** The state of any source node encompasses the reputation values of its neighbors.

**Actions.** The decision maker should select one among its neighbors to for routing.

**Reward.** The destination node can be reached via several paths. Hence, a reward of +1 is assigned to every node on all discovered paths. Otherwise no reward is assigned.

Concerning RL-algorithm, a MC method was adopted. The authors justify their choice by the episodic nature of the routes discovery process which starts when a source node initiates a route discovery and terminates when the destination node is found or when a maximum number of hops is reached.

### 4.4 Selfishness and Energy Aware RL-Based Routing Protocol

In [10], the routing problem is mapped into a partially observable MDP as follows:

**State and observation spaces.** A node state is a vector of its one-hop-neighboring nodes parameters. Those parameters can be about congestion level, selfishness, remaining energy, etc. The overall state space is the  $m \times (n - 1)$  dimensional unit cube, where  $n$  is the number of nodes in the network and  $m$  is the number of considered parameters. However, those parameters are usually unknown to the decision-maker node. To deal with this partial observability, a source node derives estimates about the values from past experiences with its neighboring nodes. Note that, only energy and selfishness parameters were considered in the experiments.

**Actions.** *the actions space*  $A$  is a  $n - 1$  dimensional simplex spanned by the vectors  $g_{max} e_j$  . where  $n$  is the number of nodes in the network,  $g_{max}$  an upper bound on the number of packets a node can process during a single time step and  $e_j$  is the unit vector along the  $j^{th}$  coordinate axis.

**Reward.** A non-linear reward function was used, defined by :

$$r(a(t), s(t)) = \sum_j f^t(\hat{\Theta}_j(t)) (\alpha \cdot a_{ij}(t) - \exp(a_{ij}(t) \cdot C)) \quad (2)$$

Where:  $f^t$  is the learned controller at time step  $t$ ;  $\alpha$  is a predefined constant;  $C$  the energy needed to send a packet;  $a_{ij}(t)$  the number of packets send by node  $i$  through node  $j$  at time  $t$  and  $\hat{\Theta}_j(t)$  is the current estimate of node  $j$  parameters values.

**State transitions.** To update energy and selfishness estimates, two learning rate were used:  $\alpha_w$  when wining and  $\alpha_l$  when losing. When the ratio between the number of packets forwarded by a node  $j$  and the number of packets sent to node  $j$  is greater than the corresponding estimated value than the node is winning otherwise it is losing.

**Action selection rule.** When a source node needs to make decision it calculates the value of the controller for all nodes in the set of one hop neighboring toward a destination  $d$ , given the current nodes parameters estimates. Then, it selects the greedy action i.e. the node that is most likely to forward the packets with probability  $1 - \epsilon_t$  and a randomly selected node, different from the greedy choice, with probability  $\epsilon_t$  where  $\epsilon_t = 1/t$ .

In this work, a stochastic gradient descent based algorithm that allows nodes to learn a near optimal controller was exploited. This controller estimates the forwarding probability of neighboring nodes. Roughly speaking, the idea behind policy search by gradient is to start with some policy, evaluate it and make an adjustment in the direction of the empirically estimated gradient of the aggregate reward, in order to obtain a local optimum policy [11].

## 4.5 RL-Based Energy-Efficient Routing

The RL-based routing protocol presented in [12] has two contrasting optimization objectives, namely: maximizing network lifetime and minimizing energy consumption in MANETs. The routing problem was mapped into a MDP as follows:

**States.** The state space of a source node is given by  $S = \{s | s = [P_{l_e}(i), B_{l_b}(j)], 1 \leq i \leq n, 1 \leq j \leq m\}$  where  $P_{l_e}(i)$  and  $B_{l_b}(j)$  denotes the quantized energy and battery network levels, respectively.

**Actions.** The decision maker should choose a path. The action space includes three actions, namely: minimum-energy routing path " $a(1)$ ", the max-min routing path " $a(2)$ " and the minimum cost routing path " $a(3)$ ". Hence  $A = \{a | a = [a(1), a(2), a(3)], a(j) \in \{0,1\}, \sum a(j) = 1\}$  where selecting a path is indicated by attributing 1 to the corresponding component.

**Cost structure<sup>1</sup>.** Once the source node selects an action at a given state, the following cost incurs:  $c(s, a) = (P_l)^{x_1} (B_l)^{-x_2} (B_{init})^{x_3}$ , where:  $B_{init}$  is the initial level of battery assumed to be constant for all nodes;  $x_1, x_2, x_3$  are weight factors empirically fixed to 1,  $B_l$  and  $P_l$  are ,respectively, the battery bottleneck and the energy consumption along the path l.

Note that the  $\epsilon$ -greedy actions-selection rule was applied. Concerning the RL resolution method, the authors have adopted a MC method.

**Table 1.** Comparison of RL-based routing protocols described in Section 4

	MDP	POMDP	Model Based	Model Free	TD algorithm	MC algorithm	Stochastic Gradient descent
MQ-routing[7]	✓			✓	✓		
LQ-routing[8]	✓			✓	✓		
SAMPLE[2]	✓		✓		✓		
SNL-Q routing[9]	✓		✓		✓		
RL-based Secure routing[4]	✓			✓		✓	
Selfishness and energy aware RL based routing[10]		✓	✓				✓
RL based Energy-efficient routing[12]	✓			✓		✓	

## 5 Conclusion

In this paper, we have selected some representative works from the literature of RL-based routing protocols for MANETs. We have described various MDP and POMDP models formalizing the routing problem with different optimization goals including quality of service (delay and/or throughput), security and energy constrained routing. We can summarize the fundamental design-issues that arise when dealing with routing problem as a RL task in MANETs as follows:

**Model-free versus Model-based.** As indicated in [2], model-based RL is more appropriate in environments where acquiring real-world experiences is expensive. However, model-based methods are characterized by slower execution times. This is, in fact, problematic in the case of real-time applications with strict response-time constraints. Besides, cost of constant probing incurred by model-free RL methods in terms of routing overhead may degrade the routing protocol efficiency. Therefore, a deep analysis of the compromise between convergence time and efficiency is needed to make appropriate recommendations.

<sup>1</sup> Since the optimization goals are minimization problems, we talk about cost rather than reward.

**MDP versus POMDP.** In fact, uncertainty is more pronounced in MANETs due to their very dynamic nature. However, almost all the works presented in this paper neglect this fact and consider the environment state as completely observable. We believe that an accurate model that really reflect MANETs features should deal explicitly with partial observability.

**Exploitation versus exploration.** The tradeoff between exploration and exploitation is well known as a challenging issue in any reinforcement learning algorithm. However, we believe that in presence of mobility the network can be considered somewhat auto-explorative. We think that the relationship between mobility, exploitation and exploration must be investigated.

**RL algorithms.** RL algorithms used to solve the routing problems described in the previous section mainly break into TD or MC methods. However, we do not know which approach is more efficient for MANETs and under which circumstances. We believe that comparative studies in this sense will be of significant interest.

## References

1. Chettibi, S., Chikhi, S.: A survey of Reinforcement Learning Based Routing for Mobile Ad-hoc Networks. In: The third international Conference on Wireless and Mobile Networks, Turkey (2011)
2. Dowling, J., Curran, E., Cunningham, R., Cahill, V.: Using feedback in collaborative reinforcement learning to adaptively optimize MANET routing. *IEEE Trans. Syst. Man. Cybern.* 35, 360–372 (2005)
3. Usaha, W., Barria, J.A.: A reinforcement learning Ticket-Based Probing path discovery scheme for MANETs. *Ad Hoc Networks Journal* 2, 319–334 (2004)
4. Maneenil, K., Usaha, W.: Preventing malicious nodes in ad hoc networks using reinforcement learning. In: The 2nd International Symposium on Wireless Communication Systems, Italy, pp. 289–292 (2005)
5. Sutton, R., Barto, A.: Reinforcement learning. MIT Press, Washington (1998)
6. Boyan, J.A., Littman, M.L.: Packet routing in dynamically changing networks: A reinforcement learning approach. *Advances In Neural Information Processing Systems* 6, 671–678 (1994)
7. Chang, Y.-H., Ho, T.: Mobilized ad-hoc networks: A reinforcement learning approach. In: ICAC 2004: Proceedings of the First International Conference on Autonomic Computing, USA, pp. 240–247. IEEE Computer Society, Los Alamitos (2004)
8. Tao, T., Tagashira, S., Fujita, S.: LQ-Routing Protocol for Mobile Ad-Hoc Networks. In: Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (2005)
9. Binbin, Z., Quan, L., Shouling, Z.: Using statistical network link model for routing in ad hoc networks with multi-agent reinforcement learning. In: International Conference on Advanced Computer Control, pp. 462–466 (2010)
10. Nurmi, P.: Reinforcement Learning for Routing in Ad Hoc Networks. In: 5th Int. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. IEEE Computer Society Press, Los Alamitos (2007)
11. Peshkin, L.: Reinforcement Learning by Policy Search. PhD thesis, Brown University (2001)
12. Naruephiphat, W., Usaha, W.: Balancing tradeoffs for energy-efficient routing in MANETs based on reinforcement learning. In: The IEEE 67th Vehicular Technology Conference, Singapore (2008)



# Identifying Usability Issues in Personal Calendar Tools

Dezhi Wu

Department of Computer Science & Information Systems  
Southern Utah University  
Cedar City, UT 84720, USA  
wu@suu.edu

**Abstract.** This study investigates what types of personal calendar tools people are currently using to manage their time and what usability issues they may experience with their calendar tools. Two sets of in-depth interviews and field observations were designed and carried out at a US university with twenty busy knowledge workers. The preliminary study results indicate that users who used a mixture of mobile and desktop calendar tools reported much higher perceived satisfaction and effectiveness than those who relied on single electronic tools or paper calendars. Furthermore, a number of usability issues are identified and used to propose new features, which can be beneficial to the design of more intelligent electronic calendar systems. Thus, this research not only attempts to understand the current technology-based time management strategies, but to apply this understanding to the development and testing of better calendar tools to more effectively support users' time management tasks.

**Keywords:** Human-Computer Interaction, Time, Calendar, Time Management Systems, Mobile Calendars, Mobile Computing, Usability.

## 1 Introduction

Nowadays people are faced with increasingly complex scheduling demands and multiple deadlines, which dominate a large percentage of their time. People are constantly being called to meetings with management, colleagues and clients at work and also need to coordinate their personal and family endeavors. Secretaries have trouble scheduling meetings with busy and diverse attendees. For people who take on multiple job roles, it becomes extremely difficult to manage their time. Therefore, scheduling ends up being a complex task with good time management skills being a critical factor for a successful professional life.

To manage time, calendar tools have served as a very valuable aid in people's professional lives [1]. In recent years, more and more personal mobile tools provide calendar functions, such as the iPhone, Blackberry, PDA and so on. However, due to the complicated nature of individual time management behavior, individual time management and the calendar tools are under-researched. Many variables impact an individual's time management behavior, such as people's roles, gender, social classes, education background and even biological clock systems. From the existing literature,

little research focused on investigating what types of time management tools people use in the current mobile computing society. It is also not clear how users experienced technical difficulties with their personal calendar tools. Can better information technology be built to help people more effectively manage their time resources? Therefore, the goal of this study is to investigate these research questions.

This is done by first gathering information on the current time management technologies that people use, on the strategies they employ to manage their time and on the problems they encounter in carrying out time management tasks with existing tools. The information collected is used to design a new calendar tool that can be run on both mobile devices and on desktop computers.

The rest of this paper presents a rationale on why developing better time management tools is important. A survey of research has been conducted on time management and calendars to serve as the theoretical foundation for this study. Following a description of two in-depth semi-structured interviews and field observations that were carried out with twenty users, this paper then briefly reports the study findings on the types of calendar tools and usability issues that users experienced. Furthermore, it proposes additional new features for a more intelligent calendar system. Lastly, the study limitations, expected contributions and future research directions are discussed.

## 2 Theoretical Foundation

Time management is an important task for professionals to succeed in today's competitive world. Most people's time management is achieved by interacting with their schedules through their personal calendar tools. As time artifacts (e.g., clocks, calendars) were created, it became possible for people to measure, categorize and manage their invisible time [2]. Time management represents your own rhythm of time, total available time, and in particular, time allocation [3, 4]. As Arndt et al [4] mentioned, "One of the basic social needs of man is to organize this rhythm to solve the regular alternation between work, play and rest" (p. 4). Or, in more details, the allocation of time presents people's perceived freedom of timing and performing three major activity taxonomies: career-oriented, home-oriented, and leisure activities [4]. Thus, how people allocate their time demonstrates how people manage their time. In practice, time management is defined to be "about identifying what's important to you and giving those activities a place in your schedule based on your unique personality needs and goals" [5, p. 12]. Hence, people's scheduling behavior reflects their time management approaches.

Because of increasingly complex schedules, people heavily rely on their calendar tools, for example, to effectively manage various projects, especially for the "follow-the-sun" projects that are dispersed in many different time zones, coordinating project time with multiple global teams becomes very challenging. Wu and Tremaine [6] found that professionals now prefer electronic time management tools because of key features (e.g., search, visualization, sharing etc.) that make them more efficient to use. Palen [7] found that organizational characteristics and work patterns impact the use of groupware calendars and individual time patterns.

The majority of existing prior calendar studies [8, 9, 10, 11, 12, 13, 14, 15] focus on the collaborative calendar user interface design and usability issues. An innovative interface technique called fisheye visualization was implemented in PDAs to overcome the small interface constraints by enlarging focused content in the center and shrinking other contents to the boundaries [16]. A transparency technique [8] was designed and implemented in a hospital scheduling system to better view temporal conflicts for smooth temporal coordination among hospital professionals.

A theory-driven study [15] was conducted by interviewing twenty staff members from the IBM T.J. Watson Research Center. This research found that although a computerized group calendar system was available, people still used a mix of calendars, primarily relying on paper-based calendars.

Another study was conducted by Egger and Wagner [11]. This study focused on time-management as a cooperative task for scheduling surgery in a hospital. Semi-structured interviews and observation were used to collect data on the hospital staff's planning practices and perceived temporal problems. Based on this analysis of the complexity of the surgery scheduling in a large hospital, possibilities of using computer support for strengthening the sharing of information and resources as well as providing participation in decision-making were built into a time management prototype called an "Operation Book." A major conclusion from the study is that cooperation can be supported by a system, but cannot be enforced.

Palen [7] studied the use of collaborative calendar systems with 40 office-workers in a large computer company. She reported that users kept additional individual calendars and that they used both systems for such tasks as scheduling, tracking, reminding, note recording/archiving and retrieval/recall. A key difference between this study and hers is its focus on individual time management rather than collaborative time management.

A similar time management tool study was carried out in a Computer Science department of a British university [9]. They borrowed most interview questions from Palen [7]'s study and interviewed sixteen staff members, who used a public time management tool called "Meeting Maker." This research reported how people used a suite of tools to support their personal and interpersonal time management. These tools included paper, electronic devices and other media. In particular, people expect more seamless integration between time management tools.

A relevant study focusing on personal and private calendar interfaces [17] indicates that users still prefer paper calendars although they have access to PDAs and desktop interfaces. This research provides an interesting design insight to incorporate users' emotional expressions (e.g., diaries and personal relations) to the digital calendar design.

After investigating 44 families' calendaring routines in both America and Canada, Neustaedter et al. [18] reported their findings on how a typology of calendars containing family activities were used by three different types of families – *monocentric*, *pericentric*, and *polycentric* – to plan and coordinate their everyday family activities. This study outlines some guidelines to further enhance the design of digital family calendars and believes that the digital family calendar tools should be designed to fit within the existing families routines.

More recently, another research [19] on the shared online calendar in the modern office environment shows that the representation of real events in calendars can be significantly improved through data fusion with incorporating social network and location data. Through conducting a field study, the performance of online calendar was significantly improved – the number of false events went from 204 using the calendar alone to fewer than 32 using the data fusion methods. The representation of genuine events was also improved when the time and location data were incorporated. The findings uncovered from this research enable the development of new applications or improvements to existing applications. When privacy and security issues become a concern, users' sensitive information has to be carefully managed. Lee et al. [20]'s study proposes a useful authentication scheme for session initiation protocol (SIP) that does not need the password table to control the Internet communication.

The two key differences between this study and other prior calendar studies are: (1) this study focuses on investigating individual time management behavior and usability issues of personal calendar tools in the current mobile society; however, the purpose for most of prior calendar research was to collect system requirements for the early generation of collaborative calendar tools; (2) this study provides up-to-date knowledge of the users' needs to manage their time with the cutting-edge mobile devices, such as the iPhone, Blackberry etc., while the majority of prior studies were accomplished when the paper-based calendars were still pervasive and some recent studies are concerned with incorporating more contextual factors to the design of digital calendars. Therefore, it is necessary to conduct this research in order to understand the current users' needs to manage their time with personal calendar tools.

### **3 A Description of Field Study**

Two sets of semi-structured interviews were designed and conducted with twenty busy knowledge workers in a US university. Some field observations were also carried out to gain insights on how individual users dealt with their daily tasks and managed their interruptions in their offices. The roles of the study participants ranged from receptionist, faculty, administrators and PhD students. In average, they worked about 45 hours/week. This qualitative study took about six months to complete.

With the participants' consent, all interviews were audio-recorded and transcribed by two researchers. The length of the transcripts has over 300 pages, which were segmented to small units that can be coded according to the research questions. Cohen's Kappa coefficient analysis [21] was performed to measure inter-coder reliability, which reached at a satisfying level (>85%).

The first set of interviews focused on examining types of calendar tools people are using and their short-term time management strategies (those involving the current day's scheduling and temporal coordination activities) and the second set of interviews focused on understanding their long-term time management strategies (those involving weekly, monthly and yearly scheduling and long-term time management plans). When the short-term time management interviews were conducted, each interviewee showed the interviewer the schedules recorded in their electronic calendar tools (e.g. iPhone, Desktop Outlook, Google online calendar etc.).

Using the interviewees' personal schedules, they were asked to explain how and why they scheduled and allocated time on specific meetings, events or other items found in their calendars or scheduled for the coming week. Each interviewee was interviewed somewhat differently because of their different personal daily schedules.

Regarding the calendar tools, the following interview questions were used to ask individual participants:

1. What types of time management tools do you use?
2. How do you do your time management with these tools?
3. What are the problems you have using with these tools?
4. How would you evaluate your satisfaction of doing your time management on your tools?
5. How effective would you rate your time management tools for organizing your time?
6. Why did you choose the tools you are using?

Because each individual interviewee worked on various schedules for their daily work, the following interview questions were used as a guide to gather their short-term time management strategies:

1. What are the biggest time wastes in your daily work?
2. Does this daily work mirror most of your ordinary life?
3. Can you please tell me how you get rid of these time wastes?
4. When do you feel you are losing control for your time management? If yes, please indicate some situation.
5. After viewing your time management planned and completed tasks, are you going to change your time management strategies? How?

Individual long-term time management strategies were asked in the second set of interviews which took place a month later. The interview questions used to gather information on long-term time management strategies are as follows:

1. When you have too many things to do, what kind of time management strategies do you use to get your work done on time?
2. When you have important deadlines, how do you usually handle your family demands?
3. When you have too many meetings, how do you deal with more important work?
4. Do you feel you lose control of your time management? If yes, Why? If not, why not?
5. Do you usually participate in any social events? If yes, why? If not, why not?

## **4 Preliminary Study Findings**

This section briefly reports the findings from the above described qualitative study. The interview data were coded and further analyzed. From the first set of semi-structured

interviews, a variety of tools, patterns, and strategies of time management were identified. Seventy-five percent of the participants used electronic tools and only twenty-five percent of them still relied on traditional paper-based calendars (most of them are seniors). More specifically, nine out of twenty people utilized a mixture of electronic time management tools, six used single electronic tools and only five participants still relied on traditional-paper based tools, e.g., one person used a wall calendar and the other four used pocket-sized paper calendars.

**Table 1.** Users' Perception of Calendar Tool Satisfaction and Effectiveness

Type of Calendar Tools	No. of Users	Perceived Satisfaction	Perceived Effectiveness
<b>Paper-based Tools</b>			
<i>Pocket-sized</i>	4	3.5	3.75
<i>Wall-sized</i>	1	5	5
<b>Electronic Calendars</b>			
<b>Single Tools</b>			
Mobile Calendar	4	3.5	3.38
Desktop Calendar	2	3	4
<b>Mixed Tools</b>			
Desktop+Mobile+other Calendars	9	4.33	4.39

Perceived Satisfaction: Least Satisfied:1:2:3:4:5:Very Satisfied

Perceived Effectiveness: Least Effective:1:2:3:4:5:Very Effective

In addition to asking respondents what they used for time management, they were also asked about their satisfaction with each time management tool. Interviewees responded to the question "How would you evaluate your satisfaction of doing your time management on your tools?" on a Likert scale ranging from (1=least satisfied) to (5=very satisfied). In addition, they were asked to give the interviewer an assessment of how effective they felt each of the tools they used was in supporting their time management needs. For the question, "How effective would you rate your time management tools for organizing your time?" users responded on a Likert scale ranging from (1=least effective) to (5=very effective). Each respondent was then asked to give the underlying reasons for their responses. Table 1 shows the types of time management usage and the summarized results from the two Likert-scale questions.

For the paper-based calendar tools, only one person, who used a wall calendar, was happy with his tool and perceived high effectiveness. The other four people who utilized pocket-sized paper calendars reported average satisfaction and perceived tool effectiveness. The following statements indicate how two interviewees who still used paper-based calendars commented their tools:

Interviewee A: *I haven't recently found anything to be better, but I am not satisfied...The main problem is...as I said some of these things are at home...There are calendar tools, and I can sit and make the second copy, but that is too much work. The tradeoff is that I just hope nobody steals it.*

Interviewee B: *One problem with paper calendar is that you cannot delete something that is probably over...But using a computer-based calendar, it is easy to erase things...and the paper calendar is really a mess in this case. I don't need to copy everything again and again in an electronic calendar.*

On the other hand, people who used a mixture of electronic tools reported much higher perceived satisfaction and effectiveness than those who relied on single electronic tools or pocket-sized paper calendars. Several reasons were given for using a mixture of tools: (1) The tools were used collaboratively so that one had to be maintained on a desktop computer; (2) The desktop tool was more convenient but the mobile device, such as a cellphone, provided scheduling information when away from the office; (3) The tools were used to maintain different schedules, one for home and one for work; (4) Private information was kept on the mobile device, which could not be kept on the public desktop calendar. For example, another two interviewees who used electronic calendar tools stated:

Interviewee C: *Yeah. You know, I use my cell phone to manage my time, which is also a computer. Outlook has a calendar. I have it in my desktop computer. They both hold the same information. My cell phone is much more convenient to use when I travel.*

Interviewee D: *I only have an electronic calendar (user only has a Blackberry)...I also use the white board on the wall as my reminder of important things.*

The users had several major complaints about their ability to effectively manage time with their personal calendar tools. Many of the user complaints with calendar tools came from inability to flexibly schedule more complicated time events with multiple people, difficulties to find right time information among the dispersed and inconsistent time resources, and inflexibility to effectively handle time conflicts between important events and coordination with multiple parties at work and at home. More detailed usability issues with personal calendar tools are listed in Table 2.

New employees also indicated that there was no explicit ways for them to learn their job time constraints/requirements, which were mostly implicit to them at the beginning, but it took them for a while to learn more about social-temporal norms at their organization. Overall, most users understood their time constraints. However, due to the reality that the current calendar tools mainly offer schedule recording functions, which cannot support more complicated time management tasks. Additional functions/features are needed to build more intelligent time management systems.

**Table 2.** Identified Usability Issues in Personal Calendar Tools

---

Usability Problems Encountered by Personal Calendar Users
<ul style="list-style-type: none"><li>• Schedules had to be constantly copied from one place to another</li><li>• Dates and times for important events conflict with dates and times for other equally important events</li><li>• Scheduling meetings was very difficult because of different time constraints of individual participants</li><li>• Scheduling events was difficult because all event associated information and activities had to be kept track of</li><li>• Schedules are often flexible but their adjustment requires information that is often not available</li><li>• Calendar tools did not support ambiguous scheduling, that is a flexible time usage that served as a reminder</li><li>• Announcements of new scheduled time usages often came too late to adjust for other time uses</li><li>• Time usage fell into categories and individuals wanted to see the categories separately</li><li>• Important dates had to be learned from experience since they were neither published in prominent places nor made known as important</li><li>• Shifting a repeating event was always difficult due to the numerous conflicts and obstacles</li></ul>

---

## 5 Proposed Key Features for More Intelligent Calendar Systems

Based on the usability study findings (see Table 2), new system requirements for more intelligent calendar systems are identified and developed to support the users' time management needs. The following additional features are therefore proposed to enhance the design of the current calendar systems:

- The capability for users to trivially categorize their scheduled events
- The capability for users to sort, display and download scheduled events based on category
- The capability for the system to display time usage based on category
- The capability for the system to learn scheduling patterns from its user and automatically adjust settings to match these patterns
- The capability for the system to add support documents to the schedule, such as last meeting's minutes, project milestone reports, financial statements etc.
- The capability to transfer reminders to currently available media - that is, instant messaging, a ringing cellphone or a large wall display panel
- The capability to support multiple displays of time usage and schedules depending on the display device capabilities
- The capability to flexibly handle time conflicts based on the importance and priorities of time events
- The capability to automatically combine relevant time events in a personal calendar from the existing online resources



- The capability to advise users of useful time management strategies to improve the quality of individual time management
- The capability to automatically suggest available time solutions for scheduling a meeting with multiple people
- The capability to visually display the time activities among collaborators with the different levels of privacy disclosure based on users' preferences.

The new calendar systems, to be effective, need to have the above proposed intelligent features that allow users to specify what goes where. It is also evident from the interviews that users did not feel that their calendars gave them information on how their time was used. Thus, additional summaries of time usage information were considered invaluable for the busy professionals. Another key factor that came out of the interviews was the difficulties that users found in using electronic calendar systems, either because of their user unfriendliness or due to their privacy concerns of sharing the usage with others. These issues, too, should be addressed in the redesign and implementation of the new calendar system.

## **6 Study Limitations, Expected Contributions and Future Research Directions**

In summary, this qualitative study successfully identifies useful usability issues in personal calendar tools with twenty users, which suggest that a number of additional intelligent features and functions are needed for calendar systems to effectively support personal time management tasks.

One study limitation that should be noted is that this study does not necessarily represent an accurate distribution of the personal calendars in use by general users. The study sample was small and the interviews captured data at a point in time. The other limitation is that the study participants work in an academic institution. Although the institution setting represents a large collection of conflicting time patterns, it would be beneficial to have more diverse organizational settings to ensure the study results more generalizable. This is therefore part of the future research plan.

The value of this work is a quantitative demonstration of the existence and distribution of different types of personal calendar tools and the usability problems that users have encountered in their personal time management practices. The usability findings suggest some important missing features in the current electronic calendar systems. A second value is in recognizing that many of the time constraints that entrain our lives cannot be readily encoded in the current calendar systems indicating that the power of the computer in supporting personal time management is under-utilized.

This paper only reports the preliminary findings and further qualitative data analyses are still ongoing. More theory-based coding schema is being built to analyze the interview data. For example, in a business setting, the organizational-temporal structures including explicit (e.g., project deadlines) and implicit structures (e.g., organizational-temporal norms) may affect personal calendaring behaviors. The in-depth coding analysis could demonstrate how the external temporal structures interact with individual time experiences, how the temporal structures can be effectively

incorporated into both organizational and personal calendar systems, and how the individual time management strategies are tied with personal effective use of calendar tools. The long-term goal of this research is to develop such a design for a more intelligent calendar system that will provide users more flexibility and control to support their time management tasks, and will empower organizations to better coordinate time with their individual workers. Such proposed functionalities are expected to form new requirements for a more intelligent calendar system. Technically, it is important to improve the Internet communication protocol [20], mobile ad hoc networks [22], the design and implementation of the fast sent protocol [23] and the improvement of mobile networks reliability protocols [24] to support the efficiency of the digital calendar design in the future.

**Acknowledgments.** The author would like to cordially thank Dr. Marilyn Tremaine for her guidance and support. The author is also grateful to all the interviewees who offered valuable insights for this study.

## References

1. Kincaid, C.M., Dupont, P.B., Kaye, A.R.: Electronic Calendars in the Office: An Assessment of User Needs and Current Technology. *ACM Transactions on Office Information Systems* 3(1), 89–102 (1985)
2. Bluedorn, A.C., Kaufman, C.F., Lane, P.M.: How Many Times Do You Like To Do At Once? An Introduction to Monochronic and Polychronic Time. *Academy of Management Executive* 6(4), 17–26 (1992)
3. Julkunen, R.: A Contribution to the Categories of Social Time and the Economy of Time. *Acta Sociologica* 20, 5–24 (1977)
4. Arndt, J., Gronmo, S., Hawes, D.K.: The Use of Time as an Expression of Life-Style: A Cross-National Study. *Research in Marketing* 5, 1–28 (1981)
5. Morgenstern, J.: *Time Management From The Inside Out: The Foolproof System for Taking Control of Your Schedule and Your Life*, Henry Holt and Company. In: LLC, vol. 9 (2000) ISBN: 0-8050-6469-9
6. Wu, D., Tremaine, M.: Knowledge Worker Adoption of Time Management Tools: Satisfaction and Perceived Effectiveness. In: *Proceedings of the Tenth Americas Conference on Information Systems* (2004)
7. Palen, L.: Social, Individual & Technological Issues for Groupware Calendar Systems. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 99)*, Pittsburgh, pp. 17–24 (1999)
8. Beard, D., Palaniappan, M., Humm, A., Banks, D., Nair, A., Shan, Y.: A Visual Calendar for Scheduling Group Meetings. In: *Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work*, pp. 279–290 (1990)
9. Blandford, A.E., Green, T.R.G.: Group and Individual Time Management Tools: What You Get is Not What You Need. *Personal and Ubiquitous Computing* 5, 213–230 (2001)
10. Crabtree, A., Hemmings, T., Rodden, T.: Informing the Development of Calendar Systems for Domestic Use. In: *ECSCW 2003: Proceedings of the Eighth European Conference on Computer Supported Cooperative Work*, Helsinki, Finland, pp. 119–138 (2003)
11. Egger, E., Wagner, I.: Time-management: A Case for CSCW. In: *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*, pp. 249–256 (1992)

12. Grudin, J., Palen, L.: Why Groupware Succeeds: Discretion or Mandate? In: Proceedings of European Conference on Computer-Support of Collaborative Work (ECSCW 1995), Dordrecht, the Netherlands, Kluwer, Dordrecht, the Netherlands, Kluwer, pp. 263–278 (1995)
13. Kelley, J.F., Chapanis, A.: How Professional Persons Keep their Calendars: Implications for Computerization. *J Occup. Psychol.* 55, 241–256 (1982)
14. Palen, L.: Calendars on the New Frontier: Challenges of Groupware Technology, Unpublished Ph.D. Dissertation, Information and Computer Science, University of California, Irvine (1998)
15. Payne, S.J.: Understanding Calendar Use. *Human Computer Interaction*, 8(2), 83–100 (1993)
16. Bederson, B., Clamage, A., Czerwinski, M.P., Robertson, G.: A Fisheye Calendar Interface for PDAs: Providing Interviews for Small Displays. In: CHI 2003: Proceedings of ACM Conference on Human-Computer Interaction, pp. 618–619 (2003)
17. Tomitsch, M., Grechenig, T., Wascher, P.: Personal and Private Calendar Interfaces support Private Patterns: Diaries, Relations, Emotional Expressions. In: Proceedings of NordiCHI, pp. 401–404 (2006)
18. Neustaedter, C., Brush, A.J.B., Greenberg, S.: The Calendar is Crucial: Coordination and Awareness through the Family Calendar. *ACM Transaction on Computer-Human Interaction* 6(1), 6:1-6:48 (2009)
19. Lovett, T., O’Neill, E., Irwin, J., Pollington, D.: The Calendar as a Sensor: Analysis and Improvement Using Data Fusion with Social Networks and Location. In: Proceedings of UbiComp, pp. 3–12 (September 26-29, 2010)
20. Lee, C., Yang, C., Huang, S.: A New Authentication Scheme for Session Initiation Protocol. *Journal of Digital Information Management*, 7(3), 133–136 (2009)
21. Kraemer, H.C.: Kappa Coefficient. In: Kotz, S., Johnson, N.L. (eds.) *Encyclopedia of Statistical Sciences*, pp. 352–354. John Wiley and Sons, New York (1982)
22. Mbarushimana, C., Shahrabi, A.: Comparative Study of Reactive and Proactive Routing Protocols Performance in Mobile Ad Hoc Networks. In: Proceedings of Advanced Information Networking and Application Workshop (AINAW 2007), pp. 679–684 (2007)
23. Rabl, T., Koch, C., HÖlbling, G., Kosch, H.: Design and Implementation of the Fast Send Protocol. *Journal of Digital Information Management* 7(2), 120–127 (2009)
24. Zhao, F., Ye, Z.: Contrast Analysis of UDP and TCP and Improvement of UDP in Reliability,  
[http://en.cnki.com.cn/Article\\_en/CJFDTOTAL-WJFZ200609074.htm](http://en.cnki.com.cn/Article_en/CJFDTOTAL-WJFZ200609074.htm)  
(accessed on March 3, 2011)

# Adaptive Query Processing for Semantic Interoperable Information Systems

Benharzallah Saber<sup>1</sup>, Kazar Okba<sup>1</sup>, and Guy Caplat<sup>2</sup>

<sup>1</sup>Department of Computer science, Biskra University, Algeria  
sbharz@yahoo.fr,  
kazarokba@yahoo.fr

<sup>2</sup>INSA de Lyon, Villeurbanne Cedex – France  
guy.caplat@insa-lyon.fr

**Abstract.** We present an adaptive query processing approach for semantic interoperable information systems. The algorithm is self-adapted to the changes of the environment, offers a wide aptitude and solves the various data conflicts in a dynamic way, it reformulates the query using the schema mediation method for the discovered systems and the context mediation for the other systems. Another advantage of our approach consists in the exploitation of intelligent agents for query reformulation and the use of a new technology for the semantic representation.

**Keywords:** Query processing; Semantic mediation; Multi-agent systems and OWL DL.

## 1 Introduction

Interoperability has been a basic requirement for modern information systems environment. The cooperation of systems is confronted with many problems of heterogeneities and must take account of the open and dynamic aspect of modern environments. Querying the distributed ontologies is one major task in semantic interoperable information systems.

Various types of heterogeneity can be encountered cited as follow: technical, syntactic, structural and semantic heterogeneity. The resolution of semantic heterogeneity is becoming more important than before. Its types appear as: naming conflicts (taxonomic and linguistic problems) and values conflicts[28][03].

The high number of the information sources implies the increase and the diversification of the conflicts number, as well as an increase in the time of localization of relevant information. It increases also the time of transmission of the queries towards all these information sources and the time response of the information sources. Therefore, the solutions of semantic interoperability should have an intelligent processor for query processing that allows the adaptation of the environment's changes and solves the various data conflicts in a dynamic way. Each solution provides some advantages to the detriments of others. Each one of them treats just one part of the data conflicts.

We propose an adaptive query processing approach for semantic interoperable information systems. Our algorithm is self-adapted to the changes of the environment, offers a wide aptitude and solves the various data conflicts in a dynamic way. It reformulates the query using the schema mediation method for the discovered systems and the context mediation for the other systems.

In the following, section 2 presents a synthesis of the various existing approaches. Section 3 and 4 describe the architecture of the mediation and the basic concepts of our architecture. The section 5 describes the query processing and the section 6 presents the technical aspects and prototype implementation.

## 2 Related Works

As the query processing problem in distributed systems has been discussed in traditional databases and Semantic Web, two possible orientations have been proposed: the integration guided by the sources (schema mediation), and the integration guided by the queries (context mediation) [21][5] [6][8][4][10][11].

The schema mediation is a direct extension of the federate approach. Data conflicts are statically solved. In the schema mediation; the mediator should be associated with a knowledge set (mapping rules) for locating the data sources. The query processing follows an execution plan established by rules which determine the relevant data in order to treat a query (static resolution of queries). It requires a pre-knowledge on the systems participating in the cooperation. The mediator's role is to divide (according to the global schema) the user query in several sub-queries supported by the sources and gathers the results. The global schema is generally specified by object, logic, XML or OWL interfaces [24][17][3][5][22]. In all these works, the objective is to build a global schema which integrates all the local schemas. When one operates in an evolutionary world where sources can evolve all the time, the elaboration of a global schema is a difficult task. It would be necessary to be able to reconstruct the integrated schema each time a new source is considered or each time an actual source makes a number of changes [4]. Generally, the time response of the queries of this approach is better than the context mediation which requires much time (it uses the semantic reconciliation). In this approach; the transparency (is to give the illusion to the users whom they interact with a local system, central and homogeneous) is assured. The degree of automation of the resolution of the data conflicts is weak, and the scalability (the system effectiveness should be not degraded and the query processing remains independent of the addition or the suppression of systems in a given architecture) and evolutionarity (to control the update, the remove and the addition of information systems) are less respected compared to the context mediation.

Many works are dedicated to the proposition of automatic approaches for schemas/ontologies integration [30][31]. The schemas mapping notion have been particularly investigated in many studies, therefore it leads to the elaboration of several systems such as DIKE [7], COMA [13], CUPID [14]. It is possible to find analyses and comparisons of such systems in [18]. Several ontologies based approaches for integration of information were suggested. In [20] and [4] survey of this subject is presented. Among the many drawbacks of these works is that they do

not describe the integration process in a complete way; they always use assumptions like pre-existence mappings [23][33] from a part, and from another part, they provide methods to calculate mappings between general or specific ontologies [30] and they do not indicate how to really exploit it for automatic integration or for the query reformulation [22][33].

In [21][3] the authors have proposed an extended schema mediation named DILEMMA based on the static resolution of queries. The mediation is ensured by a couple mediator/wrapper and a knowledge base associated with each system that takes part in the cooperation. The mediator comprises a queries processor and a facilitator. This approach provide a better transparency and makes it possible to solve the semantic values conflicts, but in a priori manner. The automation degree of the resolution of the data conflicts is enhanced compared to the schema mediation. This later involves always the recourse of an expert of the domain. It has a low capacity to treat evolutionarity and the scalability.

The role of the mediator in the context mediation approach is to identify, locate, transform and integrate the relevant data according to semantics associated with a query [21][3]. The resolution of data conflicts is dynamic and does not require the definition of a mediation schema. The user's queries are generally formulated in terms of ontologies. The data are integrated dynamically according to the semantic information contained in the description of the contexts. This approach provides a best evolutionarity of the local sources and the automation degree of the resolution of the data conflicts is better compared to schema mediation. Two categories of context mediation are defined: - the single domain approach SIMS [9], COIN [10] working on a single domain where all the contexts are defined by using a universal of consensual speech. The scalability and evolutionarity are respected but remains limited by the unicity of the domain. - Multi-domains approaches Infosleuth [11], Observer [12] they use various means to represent and connect heterogeneous semantic domain: ontologies, hierarchy of ontologies and method of statistical analysis.

In the context mediation approach the data conflicts are dynamically solved during the execution of the queries (dynamic query resolution), allowing the best evolution of the local sources and the automation degree is enhanced compared to the schema mediation, this to the detriment of time response of the queries (it uses the semantic reconciliation). Concerning the semantic conflicts, the majority of the projects solve only the taxonomic conflicts (Coin [10]). The resolution of the values conflicts is either guided by the user (Infosleuth [11]), or unsolved in the majority of cases (Observer [12] [28]).

The agent paradigm gives a new insight for the systems nature development such as: complex, heterogeneous, distributed and/or autonomous [15][34][35][38]. Several works of semantic interoperability use the agent paradigm [16][11][29] [32].

Infosleuth project [11] is used to implement a set of cooperative agents which discover, integrate and present information according to the user or application needs for which they produce a simple and coherent interface. The Infosleuth's architecture project consists of a set of collaborative agents, communicating with each other using the agent communication language KQML (Knowledge Query and Manipulation Language). Users express their queries on a specific ontology using KIF (Knowledge Interchange Format) and SQL. The queries are dispatched to the specialized agents (agent broker, ontological, planner...) to retrieve data on distributed

sources. The resolution of many semantic conflicts remains guided by the user [3]. They use specialized agents seen as threads which are widely different from the usual definition of the cognitive agent given in the distributed artificial intelligence.

In [28], the authors propose a multi-agent system to achieve semantic interoperability and to resolve semantic conflicts related to evolutive ontologies domain. In this approach, the query processing and the validation of the mappings are completely related to the users. In [29] propose an agent based intelligent meta-search and recommendation system for products through consideration of multiple attributes by using ontology mapping and Web services.

This framework is intended for an electronic commerce domain. All the approaches cited above use a approach fixed in advance (schema or context mediation), it can cause problems scalability and adaptation to the changing environment. Our main contribution is what does not fixed in advance for query processing. Our system dynamically adapts to change of the environment and processes queries according to available information on suppliers (or environment).

### 3 Generic Architecture Based Agent for Context and Schema Mediation (GAACSM)

The cooperation suggested in our solution is based on: A preliminary construction of information before its integration in the architecture system and we use the static and dynamics query resolution. An information system can play the role of information supplier and/or consumer (Figure1).

Our architecture consists of two types of agents: intelligent agents (IAs) and routings agents. The integration phase of a new information system (IS) in our proposed mediation system begins with the creation of an IA and continues with the fastening of this last to a routing agent (RA) which is *nearest semantically*.

The new IA integrated into the system of mediation applies the Contract Net protocol and sends an invitation describing its domain. The RAs receiving the call and provide their ability (*semantic proximity rate*). As soon as, the IA receives answers from all RAs, then it evaluates these rates, and makes its choice on a RA which is the nearest semantically. The chosen RA adds the previous IA to its net contacts.

Our approach does not use a global schema or some predefined mappings. Users interrogate the consuming system (the queries formulated in term of the consuming schema). At the beginning, the intelligent agent consuming (IAC) applies the dynamic query resolution protocol (context mediation) because it does not have information on the suppliers systems. This protocol is applied via the RA which is the nearest semantically with the IAC. During the dynamic evaluation of the query, the intelligent agent suppliers (IASs) update their histories and add information (mapping between terms of query ontology and their ontologies) to facilitate their dynamic integration with the IAC.

Each IAS replies with results, the RA updates its KB and reorders the list of IASs that are the most important to previous IAC (in other words; the IASs which contain results are at the head of the list). If no IAS replies, the RA sends the query to other RAs. If there are replies, the RA adds the IASs of other RAs to its KB (auto reorganization).

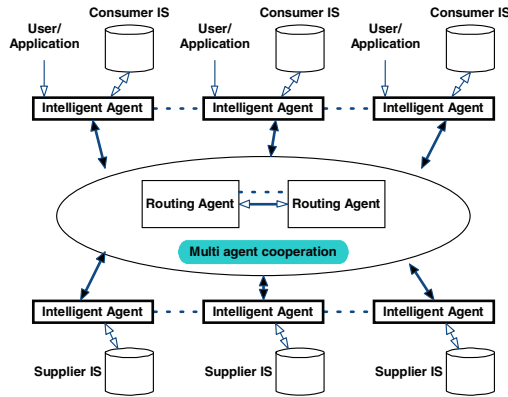


Fig. 1. General architecture of the proposed approach

During the operation of the mediation system, the IAC applies the protocol to discover suppliers which are the nearest semantically to its domain, and to integrate them dynamically in order to use them in the schema mediation. For this aim, it cooperates with the RA. Indeed; the RA updates its KB during its communication with the other agents. Particularly, its KB contains for each IA an ordered list of its IASs which are not discovered yet. These IASs should be near semantically to it. The first IAS in the list is the one which has largest number of responses of IA. After that the first IAS becomes the next supplier solicited to the following dynamic integration done by IA.

After the dynamic integration, the IAC updates its knowledge base by mapping rules.

During the operation of the system, the IAC discovers some suppliers and adapts itself with the environment. So, to treat a query two protocols should be applied: the static query resolution protocol is adopted for the discovered systems and the dynamic query resolution for other systems (algorithm3).

## 4 Basic Concepts of the GAACSM Architecture

In what follows, we present a cooperation scenario which will be used throughout this paper.

### 4.1 Cooperation Scenario

In this section we describe an interoperability example between heterogeneous systems. A given company wishes to provide an information service regarding the concerts of various artists (extension of the example cited in [26]) from the world. We chose this example for reasons of simplification.



<p>The schema of the consuming system is as follows:</p> <p>Class (CS)          FunctionalProperty (nbC domain (CS) range (xsd:integer))          DatatypeProperty (artistN domain (CS) range (xsd:string))          DatatypeProperty (dateC domain (CS) range (xsd:date))          DatatypeProperty (Pfree domain (CS) range (xsd:integer))          DatatypeProperty (Psold domain (CS) range (xsd:integer))          DatatypeProperty (Pprice domain (CS) range (xsd:float))</p> <p>nbC (integer): number identifying a concert, artistN (string): Name of artist, dateC (date): Date of concert, Pfree (integer): number of a free places, Psold (integer): number of sold places, Pprice (float): price of a place (Euro)</p>	<p>The schema of the supplier system 1 is given below:</p> <p>Class (SS1)          Class (Place)          FunctionalProperty (id domain (SS1) range (xsd:integer))          DatatypeProperty (nam domain (SS1) range (xsd:string))          DatatypeProperty (seance domain (SS1) range (xsd:date))          ObjectProperty (EidPlc domain (SS1) range (Place))          DatatypeProperty (ticket domain (SS1) range (xsd:float))          FunctionalProperty (idplc domain (Place) range (xsd:integer))          DatatypeProperty (nbP domain (Place) range (xsd:integer))          DatatypeProperty (totP domain (SS1) range (xsd:integer))          FunctionalProperty (id domain (SS1) range (xsd:integer))</p> <p>id (integer) : number identifying a concert, nam (string) :name of artist, seance (date) : date of a seance, Eidplc (integer) : an identifier reference to the relation Place-idplc, ticket (float) : price of a place (Dinars), idplc (integer) : identifier identifies nbP and totP, nbP (integer) : number of a free places, totP (integer) : number of total places</p>
<p>The schema of the supplier system 2 is the following:</p> <p>Class (SS2)          FunctionalProperty (nomCons domain (SS2) range (xsd:integer))          DatatypeProperty (NamArtist domain (SS2) range (xsd:string))          DatatypeProperty (ConsDate domain (SS2) range (xsd:date))          DatatypeProperty (soldP domain (SS2) range (xsd:integer))          DatatypeProperty (totalP domain (SS2) range (xsd:integer))          DatatypeProperty (Tprice domain (SS2) range (xsd:float))</p> <p>numCons : number identifying a concert, NamArtist : Name of artist, ConsDate: Date of seance, soldP: number of a sold places, totalP : Number of total places, Tprice : Price of a place (Dollars)</p>	

Our approach uses the OWL DL [19] as common data model. The OWL DL enriches the RDF Schemas model by defining a rich vocabulary to the description of complex ontologies. So, it is more expressive than RDF and RDFS which have some insufficiency of expressivity because of their dependence only on the definition of the relations between objects by assertions. OWL DL brings also a better integration, evolution, division and easier inference of ontologies [19].

To build an ontology from a schema; we propose the following steps: a) We use the schema to extract the concepts and the relations between them, in other words; Find the semantic organization of the various concepts (used in the schema) and the relation between them (initial construction). b) We add the synonyms and the antonyms of each name of class in *'label'*, c) We add comments on the name of classes by using *'comment'*, d) We add for each name of a class its sub concepts, its super concepts and its class's sisters.

The construction of this ontology is closely related to the context of the application domain of the information system.

**Example 1**

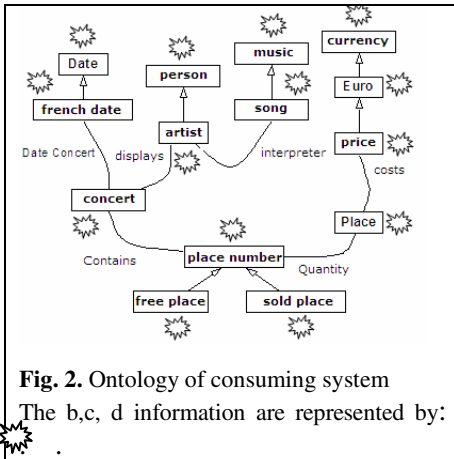
The following example indicates the schema ontologies of the consuming, supplier 1 and 2 systems built by using the preceding steps (it is a concise representation, fig 2, 3 and 4).

**4.2 Definitions**

**Definition 1: Schema-ontology mapping.** Given a schema  $S$  and its ontology  $O$ , a schema-ontology mapping is expressed by the function:

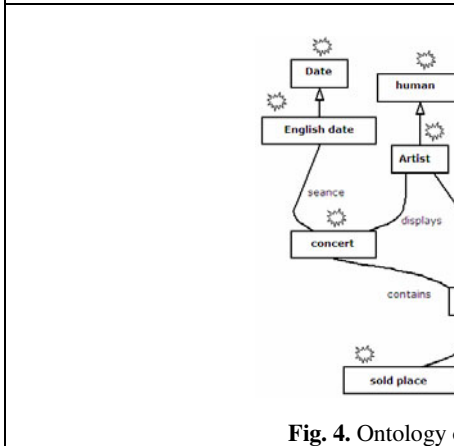
$$MSO: S \rightarrow O$$

$$x \rightarrow e$$



**Fig. 2.** Ontology of consuming system

The b,c, d information are represented by:



**Fig. 3.** Ontology of supplier 1 system

**Example 2**

Mappings  $MSO$  between the consuming schema and its ontology are the following:



**Fig. 4.** Ontology of supplier 2 system

<Concert rdf:ID="nbC"> </Concert>	<freeplace rdf:ID="Pfree"> </ freeplace>
<Concert rdf:ID="CS"> </Concert>	<soldplace rdf:ID="Psold"> </ soldplace>
<Artist rdf:ID="artistN"> </Artist>	<Price rdf:ID="Pprice"> </Price>
<Date rdf:ID="dateC"> </Date>	

**Definition 2: Context.** It describes the assumptions, the explicit information of definition and use of a data. In our approach, the context is defined by  $(S, SCV, O, MSO)$  such as:  $S$  is a schema,  $SCV$  is a semantic conflicts of values,  $O$  defines an ontology and  $MSO$  is a schema-ontology mapping.

**Definition 3: Query language.** We adapted the language defined in [2] as a query language in our architecture. Given  $L$  the set of individuals and values belonging to OWL DL data types. Given  $V$  the set of variables disjoint from those of  $L$ . A query  $Q_i$  in ontology  $O_i$  is of the form  $Q_c^i \wedge Q_p^i$ , where

- $Q_c^i$  is a conjunction of  $C^i(x)$  where  $C^i \in C$  and  $x \in L \cup V$
- $Q_p^i$  is a conjunction of  $P^i(x, y)$  where  $P^i \in P$  and  $x, y \in L \cup V$

### Example 3

This query is formulated in terms of the consuming schema.

$Q = CS(x) \wedge \text{artistN}(x, \text{"artist1"}) \wedge \text{dateC}(x, y)$ . Which means the knowledge of the date or the dates of the concerts of the artist «artist1».

**Definition 4: Semantic similarity.** The calculation of the semantic similarity between two concepts is calculated from the elementary calculations of similarity which take into account the various elements of the environment of a concept in its domain. The various adopted measures are: the terminology of the concept and environment in which the concept is located. These measurements are selected from a deep study of the various similarities measures [1] [36] [34] and from the definition of an ontology of schema in GAACSM architecture. Our algorithm which calculates the semantic similarity between two elements  $e1, e2$  is as follows (figure 5):

#### Algorithm 1. Sim(e1,e2)

**Require:** ontology  $O_1$  and  $O_2$ ,  $e1 \in O_1$ ,  $e2 \in O_2$

1: Calculation SimN of e1,e2,

2: Calculation SimC of e1,e2,

3: Calculation SimV of e1,e2,

4: Calculation SimR of e1,e2,

5:  $SimTer(e1, e2) = \alpha_1 \times SimN + \alpha_2 \times SimC$

6:  $SimStruc(e1, e2) \leftarrow \beta_1 \times SimV + \beta_2 \times SimR$

7:  $Sim(e1, e2) \leftarrow \alpha \times SimTer + \beta \times SimStruc$

**End**

**Fig. 5.** Semantic similarity algorithm

où :  $\alpha \in [0,1], \beta \in [0,1], \alpha_1 \in [0,1], \alpha_2 \in [0,1], \beta_1 \in [0,1]$  et  $\beta_2 \in [0,1]$ . SimTer: terminological similarity. SimStruc: structural similarity. SimN : Similarity of names

using their synonyms and antonyms. SimC: Comments similarity of the two concepts. SimV: Structural similarity vicinity (Our approach is based on the assumption that if the neighbors of two classes are similar, these two classes are also considered as similar). SimR: Roles similarity (The roles are the links between two OWL DL classes).

**Definition 5: Comparison of two ontologies.** The comparison of two ontologies, belonging to different IAs, The comparison is defined by the Comp function as follows:  $\text{Comp} : O \rightarrow O'$  such as  $\text{Comp}(e1) = e'1$  if  $\text{Sim}(e1, e'1) > \text{tr}$  where  $O$  and  $O'$  are two ontologies to be compared,  $\text{tr}$  indicates a minimal level of similarity belonging to the interval  $[0,1]$ ,  $e1 \in O$  and  $e'1 \in O'$ .

**Definition 6: Sub schema Adaptation of an IA.** Given two intelligent agents A, B.

- Given the schema  $S_a$ , the ontology  $O_a$  of A, and the ontology  $O_b$  of the agent B.
- Given the function  $\text{Comp} : O_a \rightarrow O_b$  the comparison between two ontologies  $O_a$  and  $O_b$  of A and B respectively.
- Given  $CO_{ab}$  the set of the elements  $e \in O_a$ , such that  $\text{Comp}(e) = e'$  and  $\text{Sim}(e, e') > \text{tr}$  with  $e' \in O_b$ .
- Given a sub-schema  $S_{S_a}$  the set of elements  $x \in S_a$  such as  $\text{MSO}(x) = e$  with  $e \in CO_{ab}$ .

The adaptation of the sub schema  $S_{S_a}$  of  $S_a$  (of the agent A) on the ontology  $O_b$  of the agent B is the function:

$$\text{Adapt} : S_{S_a} \rightarrow O_b$$

$$X \rightarrow e'$$

with:  $X \in S_{S_a}$ ,  $e' \in O_b$  where there exist  $e \in CO_{ab}$  such that  $\text{Comp}(e) = e'$  and  $\text{Sim}(e, e') > \text{tr}$ ,  $\text{MSO}(x) = e$ .

**Definition 7: Semantic enrichment of a query.** Given the context C represented by  $(S, O, \text{MSO})$  and  $Q = Q_c^i \wedge Q_p^i$ , a query formulated in term of the schema S. The semantic enrichment of this query, by using the ontology O, is defined by the following rules:

- 1) Find using the function MSO, the equivalent classes of  $C^i(x)$  and  $P^i(x, y)$  of the query  $Q_c^i$  and  $Q_p^i$  respectively in the ontology O. They are noted by  $OC^i(x)$  and  $OP^i(x, y)$  respectively.
- 2) Find by using the subsumption relation, the ancestors classes of each class of  $OC^i(x)$  and  $OP^i(x, y)$ . They are noted by  $pOC^i(x)$  and  $pOP^i(x, y)$  respectively..
- 3) Find by using the subsumption relation, the sub classes of each class of  $OC^i(x)$  and  $OP^i(x, y)$ . They are noted by  $cOC^i(x)$  and  $cOP^i(x, y)$  respectively.


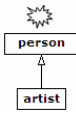

- 4) Find by using the equivalent relation, the equivalent classes of each class of  $OC^i(x)$  and  $OP^i(x, y)$ . They are noted by  $eOC^i(x)$  and  $eOP^i(x, y)$  respectively.
  - 5) We clarify, by using the schema S, the semantic conflicts of values which exist in the query Q. This information is noted by *csvQ*.
- A query Q enriched semantically is composed of :  $Q_C^i \wedge Q_P^i, OC^i(x), \{eOC^i(x)\}, \{pOC^i(x)\}, \{cOC^i(x)\}, OP^i(x, y), \{eOP^i(x, y)\}, \{pOP^i(x, y)\}, \{cOP^i(x, y)\}, csvQ$ . This enrichment is called *query ontology*.


**Example 4**

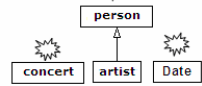
Given the following query formulated in terms of the consuming schema

$$Q = CS(x) \wedge \text{artistN}(x, "artist1") \wedge \text{dateC}(x, y)$$

The semantic enrichment of the query is as follows: We have  $Q = Q_C^i \wedge Q_P^i$ ,

- The Correspondent of the concept  $CS(x)$ , based on the function MSO, is the concept  $OC^i(x) = \text{concert}$ .
- Concepts  $\{eOC^i(x)\}, \{pOC^i(x)\}, \{cOC^i(x)\}$  are represented by: 
- The Correspondent of the concept  $OP^i(x, y) = \text{artistN}(x, "artist1")$ , based on the function MSO, is the concept *Artist*.
- Concepts  $\{eOP^i(x, y)\}, \{pOP^i(x, y)\}, \{cOP^i(x, y)\}$  are represented by: 
- The Correspondent of the concept  $OP^2(x, y) = \text{dateC}(x, y)$ , based on the function MSO, is the concept *Date*.
- Concepts  $\{eOP^2(x, y)\}, \{pOP^2(x, y)\}, \{cOP^2(x, y)\}$  are represented by: 

The semantic enrichment of the query (query  ontology)  $SC(x) \wedge \text{artistN}(x, "artist1") \wedge \text{dateC}(x, y)$  is the ontology:



And  $csvQ = \{ \text{dateC} : < OSCV : \text{Date} > \text{French date} < OSCV : \text{Date} / > \}$

**Definition 8: Semantic evaluation of a query ontology.** The semantic evaluation of a query enriched semantically (query ontology)  $O_Q = (Q_C^i \wedge Q_P^i, OC^i(x), \{eOC^i(x)\}, \{pOC^i(x)\}, \{cOC^i(x)\}, OP^i(x, y), \{eOP^i(x, y)\}, \{pOP^i(x, y)\}, \{cOP^i(x, y)\}, csvQ)$  is defined by the algorithm 2 (figure 6):

**Algorithm 2.**

**Require:** query ontology  $O_Q$  of IAC and ontology  $O_{IAS}$  of IAS

1: Calculation of the function  $Comp: O_Q \rightarrow O_{IAS}$

2: Calculation of the set  $CO_{QIAS}$

3: Calculation of the sub-schema  $SS_Q$

4: Calculation of the function  $Adapt: SS_Q \rightarrow O_{IAS}$

5: /\*Evaluate of the semantic query  $Adapt(SS_Q)$  \*/

6: For each  $X \in SS_Q$  and  $e \in Adapt(X)$  do

7: Search  $Y \in S_{IAS}$  where:  $MSO(Y)=e$

8: If  $Y$  exists then to replace  $X$  by  $Y$  in  $Q_C^i \wedge Q_P^i$  ( $Q_C^i \wedge Q_P^i \in O_Q$ )

9: Endfor

10: the query which will be to evaluate in IAS is  $Q_C^i \wedge Q_P^i$

11: End

**Fig. 6.** Algorithm 2. Semantic evaluation of a query ontology

**Example 5**

Given the previous query  $Q = CS(x) \wedge artistN(x, "artist1") \wedge dateC(x, y)$

The semantic evaluation of its query ontology, on the source of supplier 1 is done by the application of the algorithm 2. The steps are as follows:

- Calculation of the similarities between the query ontology and the ontology of supplier 1.
- Calculation of the set  $CO_{QIAS} = \{Concert, Artist, Person, Date, \dots\}$
- Calculation of the sub schema  $SS_Q = \{CS(x), artistN(x, "artist1"), dateC(x, y)\}$
- Calculation of the function  $Adapt: SS_Q \rightarrow O_{IAS}$ . Its values are :  
 $\{Adapt(SC(x))=Concert, \quad Adapt(artistN(x, "artist1"))=Musician,$   
 $Adapt(dateC(x, y))=Date\}$
- Semantic evaluation of  $Adapt(SS_Q)$ , which requires the calculation of the function  $MSO$  reverse. Hence:  $\{MSO^{-1}(Concert)=SS1, MSO^{-1}(Musician)=nam,$   
 $MSO^{-1}(Date)=seance\}$ .

- Concerning the semantic conflicts of values, the attribute *seance* uses the same format like *dateC*, else it is necessary to take into account the change of the results and to convert the format using the OSCV ontology (a transformation function).

Finally, the query which will be carried out on the level of the source of supplier 1 is as follows:  $Q = SS1(x) \wedge nam(x, "artist1") \wedge seance(x, y)$ .

**Definition 9: Mapping rules.** A schema mapping is a triplet  $(S1, S2, M)$ [2], where: S1 is the source schema; S2 is the target schema; M the mapping between S1 and S2, i.e. a set of assertions  $q_s \mapsto q_r$ , with  $q_s$  and  $q_r$  are conjunctive queries over S1 and S2, respectively, having the same set of distinguished variables  $x$ , and  $\mapsto \in \{\subseteq, \supseteq, \equiv\}$ .

**Definition 10: Query Reformulation.** Let  $Q_i$  be a query in schema  $s_i$  and  $Q_j$  be a query in schema  $s_j$  described by classes and properties in the mapping  $M_{ij}$ .

- $Q_j$  is an equivalent reformulation of  $Q_i$  if  $Q_j \subseteq Q_i$  and  $Q_i \subseteq Q_j$ , which is noted by  $Q_j \equiv Q_i$ .
- $Q_j$  is a minimally-containing reformulation of  $Q_i$  if  $Q_i \subseteq Q_j$  and there is no other query  $Q'_j$  such that  $Q_i \subseteq Q'_j$  and  $Q'_j \subseteq Q_j$ .
- $Q_j$  is a maximally-contained reformulation of  $Q_i$  if  $Q_j \subseteq Q_i$  and there is no other query  $Q'_j$  such that  $Q_j \subseteq Q'_j$  and  $Q'_j \subseteq Q_i$ .

To find the approximate query reformulation we use the mapping rules M (definition 15), we substitute the terms of  $Q_i$  by their correspondents [02].

## 5 Queries Processing

The query processing is divided into several steps, and during this process, the multi-agents system uses a set of protocols. The principal steps are (figure 7):

### 5.1 Static Query Resolution

The static resolution is applied to the systems have been already discovered.

**Step 1: query validation** the IAC checks the validity of the query.

**Step2: query reformulation:** the query is divided into a recombining query of the results and sub queries intended for the IAS which contain data necessary to the execution of the query. The decomposition of the query is done by the use of the mapping rules.

**Algorithm 3. Query processing**

```

Given L the list of discovered agents and their mappings
If QueryValidation() then
1: if L <> empty then
    - QueryReformulation()
    - StaticRecombiningResults()
2: Dynamic query resolution
    - SemanticEnrichmentQuery()
    - TransmissionSemanticallyEnrichedQuery()
    - SemanticEvaluation() /*algorithm 2*/
    - DynamicRecombiningResults()

```

**Fig. 7.** Algorithm 3. Query processing

**Step3: recombining of the results:** the IAC executes the recombining query for the results.

## 5.2 Dynamic Query Resolution

The dynamic resolution makes it possible to take into account the appearance of new IASs. The principal steps are:

**Step 1: Semantic enrichment of a query.** The IAC enriches the query semantically by using the ontology and the links schema-ontology which are in its own knowledge base (definition 7).

**Step 2: Transmission of the semantically enriched query.** The IAC applies the cooperation protocol of dynamic query resolution. So it transmits the semantically enriched query to the routing agent which is nearest semantically. This latter sends it to all IASs of its net contacts.

**Step 3: Semantic evaluation of the semantically enriched query (algorithm 2).** Each IAS answers according to its capacity to treat the query:

- 1) To compare elements of the query with its ontology. The elements of the query and its ontology are compared by using a semantic distance. The identified elements as equivalent are retained.
- 2) The query is rewritten in terms of the equivalent elements of its ontology (then interpreted on its schema) to take into account the semantic conflicts of values (each intelligent agent has library of functions for the conversion of the types).
- 3) The answer is sent latter to the routing agent, indicating the manner of treating the query, so that this letter can build recombining queries of the results.

If no IAS answers, the routing agent sends the query to the other routings agents of other domains and if there are answers the routing agent updates its net contacts.



**Stage 4: Results recombining:** the routing agent recomposes the results obtained by IASs. Then it sends the final result to the IAC, this latter recomposes the results of static and dynamic query resolution.

## 6 Technical Aspects and Prototype Implementation

Our implementation is based on three class libraries: OntoSim [39], Alignment API [25] and Jade [37]. OntoSim provides many similarities measurements between character strings. Alignment API allows to integrate new methods of similarities measurement (between two OWL ontologies) by implementing a Java interface. Jade (Java Agent Development Framework) [37] is used for the construction of the multi agents systems and the realization of applications in conformity with FIPA specifications. The cooperation protocols are implemented using the Jade platform. Concerning the local information systems, the local database of the consuming system and the database of the supplier system 1 are established under the Access DBMS and the Windows XP operating system. The database of the supplier system 2 is implemented in XML files and the same operating system. The scalability and the performances of the transport system of Jade message were treated in [27][28]. The obtained results confirm the fact that Jade deals well with the scalability according to several scenarios intra or inter framework. The figure 8 presents an example of comparison between two ontologies of the consuming system and the supplier system 2. Figure 9 presents the graphical interface, an example of query and the obtained results. In this example, the IAS1 is discovered by agent IAC. This last applies the schema mediation in order to reformulate the query. The IAC applies the context mediation for other agents, which are not yet discovered (IAS2). It communicates with the agent RA.

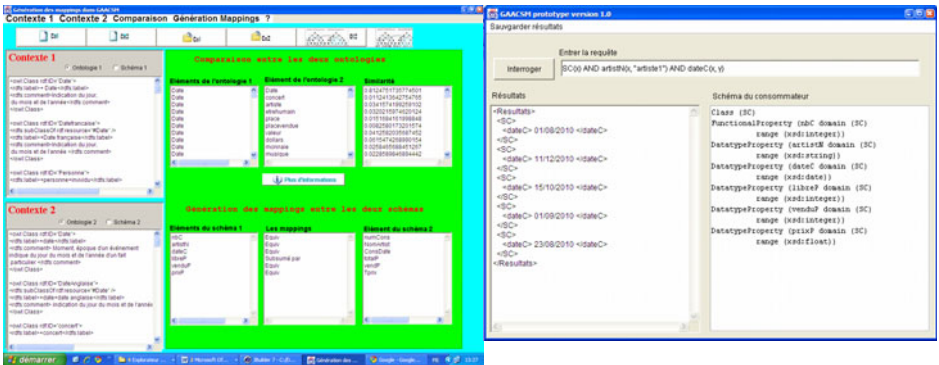


Fig. 8. Automatic mapping generation

Fig. 9. A simple example

## 7 Conclusion and Future Research

In this paper we presented a adaptive query processing approach for semantic interoperable information systems. The main advantage of our query processing

algorithm consists in its robustness with regard to the evolution of systems, adaptation to the changes of environment and the resolution of the most various data conflicts in a dynamic way.

The developed prototype shows the functionality of suggested approach. Our future research consists to employing the intelligent methods (for segmented comparison of large ontologies of arbitrary size) in order to reduce the time of ontologies comparison so we avoid the influence on the scalability of the suggested architecture.

## References

1. Rahm, E., Bernstein, P.A.: 1: A survey of approaches to automatic schema matching. *VLDB Journal The International Journal on Very Large Data Bases* 10(4), 334–350 (2001)
2. Akahani, J.-i., Hiramatsu, K., Satoh, T.: Approximate Query Reformulation for Ontology Integration. In: *Semantic Integration Workshop (SI-2003) Second International Semantic Web Conference, Sanibel Island, Florida, USA, October 20 (2003)*
3. Jouanot, F.: *DILEMMA : vers une coopération de systèmes d'informations basée sur la médiation sémantique et la fusion d'objets université de bourgogne (November 2001)*
4. Schneider, M., Bejaoui, L., Bertin, G.: A Semantic Matching Approach for Mediating heterogeneous Sources, metadata and semantic, pp. 537–548. *springer US, Heidelberg (2009)*; ISBN 978-0-387-77744-3
5. Busse, S., Kutsche, R.-D., Leser, U., Weber, H.: *Federated Information Systems: Concepts, Terminology and Architectures., Technical Report Nr.99-9 Berlin University (1999)*
6. Hakimpour, F.: *thesis: Using Ontologies to Resolve Semantic Heterogeneity for Integrating Spatial Database Schemata, Universität Zürich Zürich (2003)*
7. Palopoli, L., Terracina, G., Ursino, D.: DIKE: a system supporting the semi-automatic construction of cooperative information systems from heterogeneous databases. *Softw., Pract. Exper.* 33(9), 847–884 (2003)
8. Leclercq, E., Benslimane, D., Yétongnon, K.: ISIS: A Semantic Mediation Model and an Agent Based Architecture for GIS Interoperability. In: *Proceedings of the International Database Engineering and Applications Symposium (IDEAS 1999), Montreal, CANADA, August 2-4, pp. 81–92. IEEE Computer Society Press, Los Alamitos (1999) ISBN 0-7695-0265-2*
9. Arens, Y., Hsu, C.-N., Hsu, C.: Query processing in the SIMS Information Mediator. In: *Tate, A. (ed.) Advanced planning technology, AAAI Press, Menlo park (1996)*
10. Bressan, S., Goh, C., Levina, N., Madnick, S., Shah, A., Siegel, M.: Context Knowledge Representation and Reasoning in the Context Interchange System. *Applied Intelligence* 13(2), 165–180 (2000)
11. Bayardo Jr., R.J., et al.: *InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments. Microelectronics and Computer Technology Corporation (1997)*
12. Mena, E., Kashyap, V., Sheth, A., Illarramendi, A.: OBSERVER: An approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In: *International journal on Distributed And Parallel Databases DAPD, vol. 8(2), pp. 223–272 (April 2000)*
13. Do, H.-h., Rahm, E.: COMA - A System for Flexible Combination of Schema Matching Approaches. In: *VLDB 2002, pp. 610–621 (2002)*

14. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic Schema Matching with Cupid. In: VLDB 2001, pp. 49–58 (2001)
15. Girardi, R.: An analysis of the contributions of the agent paradigm for the development of complex systems. In (SCI 2001) and (ISAS 2001), Orlando, Florida (2001)
16. Purvis, M., Cranefield, S., Bush, G., Carter, D., McKinlay, B., Nowostawski, M., Ward, R.: The NZDIS Project: an Agent-Based Distributed Information Systems Architecture. In: Proceedings of the Hawai'i International Conference On System Sciences, Maui, Hawaii, January 4–7 (2000)
17. Suwanmanee, S., Benslimane, D., Thiran, P.: OWL Based Approach for Semantic Interoperability. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications, AINA 2005 (2005)
18. Mohsenzadeh, M., Shams, F., Teshnehlab, M.: Comparison of Schema Matching Systems. In: WEC, vol. (2), pp. 141–147 (2005)
19. Deborah, L., McGuinness, van Harmelen, F.: Owl web ontology language . Technical report (2004)
20. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H.: Ontology-based integration of information - a survey of existing approaches. In: Stuckenschmidt, H. (ed.) IJCAI 2001 Workshop: Ontologies and Information Sharing, pp. 108–117 (2001)
21. Jouanot, F., Cullot, N., Yetongnon, K.: Context Comparison for Object Fusion. In: Eder, J., Missikoff, M. (eds.) CAiSE 2003. LNCS, vol. 2681, pp. 536–551. Springer, Heidelberg (2003)
22. Tous, R.: Thesis: Data Integration with XML and Semantic web Technologies. Universitat Pompeu Fabra, Barcelona (June 2008) ISBN:9783836471381
23. Goasdoué, F., Rousset, M.-C.: Querying Distributed Data through Distributed Ontologies: A Simple but Scalable Approach. IEEE Intelligent Systems, 60–65 (2003)
24. Saw, N.T.H.: OWL-Based Approach for Semantic teroperating and Accessing Heterogeneous XML Sources, 0-7803-9521-2/06/\$20.00 §. IEEE (2006)
25. <http://alignapi.gforge.inria.fr/lib.html>
26. Jouanot, F.: Un modele sémantique pour l'interopérabilité de systemes d'information, laboratoire LE2I. In: INFORSID (2000)
27. Cortese, E., Quarta, F., Vitaglione, G.: Scalability and Performance of Jade Message Transport System. Presented at AAMAS Workshop on AgentCities, Bologna (July 16, 2002)
28. Séguran, M.: Résolution des conflits sémantiques dans les systèmes d'information coopératifs: proposition d'un modèle d'interaction entre agents, Université Jean Moulin, Lyon3 (2003)
29. Kim, W., Choi, D.W., Park, S.: Agent based intelligent search framework for product information using ontology mapping. J. Intell. Inf. Syst. 30, 227–247 (2008) DOI 10.1007/s10844-006-0026-8
30. Mougín, F., Burgun, A., Bodenreider, O., Chabalier, J., Loréal, O., Le Beux, P.: Automatic methods for integrating biomedical data sources in a mediator-based system. In: Bairoch, A., Cohen-Boulakia, S., Froidevaux, C. (eds.) DILS 2008. LNCS (LNBI), vol. 5109, pp. 61–76. Springer, Heidelberg (2008)
31. Hansen, H.L., Ingvaldsen, K.: Dynamic Schema Integration The ACROs Prototype THESIS, The Faculty of Social Sciences, Department of Information Science and Media Studies (May 2005)

32. Gal, A., Segev, A., Tatsiopoulos, C., Sidiropoulos, K., Georgiades, P.: Agent oriented data integration. In: Akoka, J., Liddle, S.W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., Mayr, H.C. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 98–108. Springer, Heidelberg (2005)
33. Peter haase, dissertation, semantic technologies for distributed information systems, universitätsverlag karlsruhe (2007) ISBN 978-3-86644-100-2
34. KnowledgeWeb, Deliverables of KWEB Project: D2.2.3: State of the art on ontology alignment. EU-IST-2004-507482 (2004),  
<http://knowledgeweb.semanticweb.org/>
35. Carbonell, J.G., Siekmann, J.: In: Kolp, M., Henderson-Sellers, B., Mouratidis, H., Garcia, A., Ghose, A.K., Bresciani, P. (eds.) AOIS 2006. LNCS (LNAI), vol. 4898, Springer, Heidelberg (2008)
36. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
37. Bellifemine, F., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. John Wiley & Sons Ltd, Chichester (2007) PO19 8SQ, England, Copyright
38. LI, A.-P., Yan, J., Wu, Q.-Y.: A Study on Organizational Knowledge in Multi-Agent System. JCIT:Journal of Convergence Information Technology 2(1), 61–65 (2007)
39. <https://gforge.inria.fr/projects/ontosim/>

# Statistical Character-Based Syntax Similarity Measurement for Detecting Biomedical Syntax Variations through Named Entity Recognition

Hossein Tohidi, Hamidah Ibrahim, and Masrah Azrifan Azmi

Faculty of Computer Science, and Information Technology  
University Putra Malaysia, Serdang, Malaysia  
Tohidi.h@gmail.com, {hamidah,masrah}@fsktm.edu.my

**Abstract.** In this study an approach for detecting biomedical syntax variations through the Named Entity Recognition (NER) called Statistical Character-Based Syntax Similarity (SCSS) is proposed which is used by dictionary-based NER approaches. Named Entity Recognition for biomedical literatures is extraction and recognition of biomedical names. There are different types of NER approaches, that the most common one is dictionary-based approaches. For a given unknown pattern, Dictionary-Based approaches, search through a biomedical dictionary and finds the most common similar patterns to assign their biomedical types to the given unknown pattern. Biomedical literatures include syntax variations, which means two different patterns, refer to the same biomedical named entity. Hence a similarity function should be able to support all of the possible syntax variations. There are three syntax variations namely: (i) character-level, (ii) word-level, and (iii) word order. The SCSS is able to detect all of the mentioned syntax variations. This study is evaluated based on two measures: recall and precision which are used to calculate a balanced F-score. Result is satisfied as recall is 92.47% and precision is 96.7%, while the f-test is 94.53%.

**Keywords:** Artificial Intelligence, Natural Language Processing, Information Extraction, Named Entity Recognition, Text Mining, Biomedical.

## 1 Introduction

The plethora of material on the WWW is one of the factors that have sustained interest in automatic methods for extracting information from text. Information extraction is an application of Natural Language Processing (NLP). As the term implies, the goal is to extract information from text, and the aim is to do so without requiring the end user to read the text. The most important phase of information extraction is Named Entity Recognition (NER). This phase is the task of recognizing entity-denoting expressions, or named entities, in natural language documents. These name entities, can be person, places, organizations and especially biomedical name entities like gene and protein name.

The aim of this paper is to present a NER approach to extract and recognize the biomedical name entities like gene and protein, by considering the biomedical named entity complexities especially variation. Hereby a hybrid approach over Dictionary-Based and Machine Learning approaches is designed, which is called Guess and Try (GAT).

One of the important phases in GAT is GUESS phase. A named entity is a combination of several names, therefore it can be considered as a noun group (noun phrase) or subset of a bigger noun group. Hence in GUESS phase, GAT extracts the entire noun group from the text, in contrast to some approaches that consider the given text as a chain of words.

The second phase TRY recognizes the extracted noun groups from the GUESS phase, and assigns a biomedical type (gene, protein) to them. For an extracted named entity GAT generates more than one suggestion (biomedical type) with a probability rate between 0 and 1. The strength of GAT is based on a Statistical Character-Based Syntax Similarity function which measures the similarity between two different name groups.

The rest of the paper is organized as follows. Section 2 presents the background concepts related to biomedical named entity and highlights the complexities of this area. Section 3 presents related works, based on the approaches they have developed. Section 4 describes the GENIA corpus which is used in our approach. Section 5 presents our proposed approach. Section 6 presents the evaluation of GAT and this is compared to other approaches and Section 7 is the conclusion of this study.

## 2 Background

In this section, the definition of biomedical entities and their complexities are presented.

### 2.1 Biomedical Named Entity

There are several online databases which include list of biomedical NEs. Each of the online databases in Table 1 contains a comprehensive list of biomedical NEs.

**Table 1.** Biomedical NEs Online Databases

Biomedical class	Online Database
Genes	Human Genome Nomenclature ( <a href="http://www.gene.ucl.ac.uk/nomenclature/">http://www.gene.ucl.ac.uk/nomenclature/</a> )
Proteins	UniProt ( <a href="http://www.expasy.org/sprot/">http://www.expasy.org/sprot/</a> ), IPI ( <a href="http://www.ensembl.org/IPI/">http://www.ensembl.org/IPI/</a> )
Cells	Cell database of Riken Bioresource Center ( <a href="http://www.brc.riken.jp/inf/en/">http://www.brc.riken.jp/inf/en/</a> )
Drugs	MedMaster ( <a href="http://www.ashp.org/">http://www.ashp.org/</a> ), USP DI ( <a href="http://www.usp.org/">http://www.usp.org/</a> )
Chemicals	UMLS Metathesaurus ( <a href="http://www.nlm.nih.gov/research/umls/">http://www.nlm.nih.gov/research/umls/</a> )
Diseases	NCBI Genes and Diseases ( <a href="http://www.ncbi.nlm.nih.gov/disease/">http://www.ncbi.nlm.nih.gov/disease/</a> )

Table 2 shows samples of NEs for each category in the biomedical text.

**Table 2.** Biomedical Entity

Biomedical class	Example
Genes	Tp53, agar
Proteins	p53, galactosidase, alpha (GLA)
Cells	CD4+-cells, Human malignant mesothelioma
Drugs	Cyclosporine, herbimycin
Chemicals	5'-(N-ethylcarboxamido)adenosine (NECA)

Biomedical NEs like gene and protein names in the databases and in the texts show several characteristics in common. For instance, many gene and protein names include special characters of the type as shown in Table 3.

**Table 3.** Special Characters in Biomedical NEs

Character
Upper Case
Comma
Hyphen
Slash
Bracket
Digit

There are some predefined standard rules for naming a biomedical entity, like gene, protein, etc, which can help to recognize these patterns among a biomedical text. Some of these rules are presented in Table 4.

In this paper the name entities or patterns which satisfy one or more of the above rules are called *welled-form* entity. The recognition of gene and protein names in the biomedical text is not straightforward, despite many well-known nomenclatures, such as HUGO and Swiss-Prot. In the following some of these issues are explained.

**Table 4.** Standard Biomedical Naming Rules

Naming Rule	Example
Comma	,
Dot	.
Parenthesis	( ) [ ]
RomanDigit	<i>II</i>
GreekLetter	<i>Beta</i>
ATCGsequence	<i>ACAG</i>
OneDigit	<i>5</i>
AllDigits	<i>60</i>
DigitCommaDigit	<i>1,25</i>
OneCap	<i>T</i>
AllCaps	<i>CSF</i>
CapLowAlpha	<i>All</i>
CapMixAlpha	<i>IgM</i>
LowMixAlpha	<i>kDa</i>
AlphaDigitAlpha	<i>H2A</i>
AlphaDigit	<i>T4</i>
DigitAlphaDigit	<i>6C2</i>
DigitAlpha	<i>19D</i>

## 2.2 Ambiguous Names

Some ambiguous names will be confused with common English words; such as *can*, *for*, *not*, *vamp*, *zip*, *white*, and *cycle*.

Also some times, the ambiguity can be presented as two different meaning for one particular pattern. For example, the name *p21* formerly denoted a macromolecule associated with a cascade of signals from receptors at cell surfaces to the nucleus, which stimulates cell division, but currently it denotes a different protein that inhibits the cell cycle.

## 2.3 Variation

Variation means gene and protein names denote the same entities by definition, but different in name. Interestingly, gene and protein names show a high degree of variations in the text, including *character-level* variations, *word-level* variations and *word-order* variations, as illustrated in Table 5.



**Table 5.** Variation in Biomedical NEs

Variation	Examples
Character-level variations	<ul style="list-style-type: none"> <li>i. D(2) or D2</li> <li>ii. SYT4 or SYT IV</li> <li>iii. IGA or IG alpha</li> <li>iv. S-receptor kinase or S receptor kinase</li> <li>v. Thioredoxin h-type 1 or Thioredoxin h (THL1)</li> </ul>
Word-level variations	<ul style="list-style-type: none"> <li>i. Rnase P protein or Rnase P</li> <li>ii. Interleukin-1 beta precursor INTERLEUKIN 1-beta PROTEIN or INTERLEUKIN 1 beta transcription intermediary factor-2 or transcriptional</li> <li>iii. intermediate factor 2</li> </ul>
Word-order variations	<ul style="list-style-type: none"> <li>i. Collagen type XIII alpha 1 or Alpha 1 type XIII collagen</li> <li>ii. integrin alpha 4 or alpha 4 integrin</li> </ul>

### 3 Related Work

Generally there are three main approaches for NER, as follow:

1. Dictionary-based approaches that find names of the well-known nomenclatures in the text by utilizing a dictionary.
2. Rule-based approaches that manually or automatically construct rules and patterns to directly match them to candidate NEs in the text.
3. Machine learning (statistical) approaches that employ machine learning techniques, such as HMMs and SVMs, to develop statistical models for gene and protein name recognition.
4. Hybrid approaches that merge two or more of the above approaches, mostly in a sequential way, to deal with different aspects of NER.

These approaches are explained in the following subsections. The focus of this research is on a hybrid over machine learning (statistical) and dictionary based approaches that use an annotated biomedical dictionary.

#### 3.1 Dictionary-Based Approaches

Unlike the names of persons and locations in the general domain, gene and protein names have been well-managed through databases by leading organizations, such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) (<http://www.ebi.ac.uk/>). It is a natural consequence that previous approaches to gene and protein name recognition have been heavily dependent on such databases. The approaches usually find the database entry names

directly from the text. However, they have several limitations and lack of a unified resource that covers newly published names.

Tsuruoka and Tsujii [1] address the aforementioned problems of dictionary-based approaches with a two-phase method. The method first scans texts for protein name candidates, using a protein name dictionary expanded by a probabilistic variant generator. The generator produces morphological variations of names in the class ‘Amino Acid, Peptide, or Protein’ of the UMLS meta thesaurus and further gives each variant a generation probability that represents the plausibility of the variant. In the second phase, the method filters out irrelevant candidates of short names by utilizing a Naïve Bayes classifier with the features of words both within the candidates and surrounding the candidates.

Hanisch [11] constructs a comprehensive dictionary of genes and proteins by merging HUGO Nomenclature, OMIM database, and UniProt. They present a method for detecting gene and protein names with the unified dictionary. The method processes tokens in a MEDLINE abstract one at a time and scores each candidate name with two measures; that is, boundary score to control the end of the candidate and acceptance score to determine whether the candidate is reported as a match.

### 3.2 Rule-Based Approaches

The dictionary-based approaches can deal with only morphological variations that correspond to some of the character-level and word-level variations in Table 5. Rule-based approaches can deal with a broader range of variations, even covering a few of the word-order variations in Table 5.

Fukuda [10] presents a method of protein name recognition that utilizes surface clues on character strings. The method first identifies core terms, those that contain special characters in Table 3, and feature terms, and those that describe biomedical functions of compound words (e.g., *protein* and *receptor*). It then concatenates the terms by utilizing handcrafted rules and patterns, and extends the boundaries to adjacent nouns and adjectives.

### 3.3 Machine Learning Approaches

Depending on the nature of biological texts, an approach is needed to find in what probability a pattern can be an entity.

Collier et al. [3] use a supervised training method with Hidden Markov Model (HMM) to overcome the problem of rule-based approaches. The HMM is trained with bigrams based on lexical and character features in a small corpus of 100 MEDLINE abstracts. For each sentence, the model takes an input that consists of a sequence of words in the sentence and their features. The features used in the model include the presence or absence of each special character, and whether a word is a determiner or a conjunction. For the given class, the model then calculates the probability of a word belonging to the class. Finally, it produces the sequence of classes with the highest probabilities for the given sequence of words in the sentence. Domain experts mark up or annotate the corpus that is used to train the model with classes, such as proteins and DNA.

In order to handle the lack of training corpus for gene and protein name recognition, Morgan [4] presents a method for automatically constructing a large quantity of training corpora by utilizing FlyBase, which includes a created list of genes and the MEDLINE abstracts from which the gene entries are drawn. They apply simple pattern matching to identify gene names or their synonyms in each article. The noisy corpus, automatically annotated with gene entries of FlyBase, is used to train an HMM for gene name recognition.

### 3.4 Hybrid Approaches

As the number of features for machine learning systems increases to cover more phenomena in NER, the data sparseness problem becomes more serious. Since the approaches discussed above have their own advantages and disadvantages, there is a clear need for combining them for better performance. In fact, some of the methods introduced in the previous subsections are a hybrid of different kinds of approaches.

For instance Proux [6] applies a machine learning technique for disambiguation of relevant candidate gene names in a rule-based system. Zhou and Su [5] automatically construct rules to deal with cascaded entity names for their machine learning system.

## 4 GENIA 3.0

In this study, the GENIA 3.0 which is the largest annotated corpus in molecular and biology domain available to public [9] is used. The corpus is available in XML format and contains 2000 MEDLINE abstracts of 360K words. Ontology for biomedical entity is based on the GENIA ontology which includes 23 distinct classes that are: *gene, multi-cell, mono-cell, virus, body part, tissue, cell type, cell component, organism, cell line, other artificial source, protein, peptide, amino acid monomer, DNA, RNA, poly nucleotide, nucleotide, lipid, carbohydrate, other organic compound, inorganic, and atom.*

In the corpus the biomedical entities are represented by XML tag called *cons*. Each *cons* tag represents a biomedical named entity by the *lex* attribute while its biomedical type (class label) is represented by the *sem* attribute. For instance the following XML code represents one of the nodes in the corpus which stands for a biomedical named entity *Protein Kinase C*.

```
<cons lex="protein_kinase_C" sem="G#protein_molecule">
  Protein kinase C
</cons>
```

Among these 24 biomedical classes this study focuses on the gene and protein names, since these categories have the most complexity for NER compare to the other classes. Also most of the previous researches have been done over these two categories.

## 5 Proposed Approach

Our proposed approach GAT consists of two major phases and one minor data preprocessing phase. The three phases are GUESS, TRY, and data preprocessing which are explained in the following subsections.

### 5.1 Data Preprocessing

There are three steps in this phase.

#### Generate the BioDic

The first step in data preprocessing is a word-level process which extracts words from the 2000 abstracts. All the GENIA abstracts are scanned and tokenized to transform the given unstructured text into a list of extracted words. The occurrence of each word is counted to compute its frequency distribution.

#### Detect Stop Word

Stop words are words that are filtered out before or after the processing of natural language. In this step all of the stop words are removed from the GENIA abstracts.

#### Generate Training Set

In this step the XML file is scanned and its content is transferred and represented into a tabular structure as presented in Table 6.

**Table 6.** Training Dataset after Transformation

Index	Named entity	Entity Type	Welled-form	Frequency	Main Index
100	Protein PKC_Activator	Protein	Yes	7000, 200, 6543	1
101	Cell_Fat _Activator	Cell	No	8453, 5432, 5411	0

In this study for each biomedical named entity four features are defined as below:

1. *Article Index*: A unique *ID* to indicate the extracted biomedical named entity.
2. *Named entity (Noun Group)*: A named entity may contain one or more nouns, which are separated by *underscore*. Therefore it can be regarded as a noun group. This is retrieved from the *lex* attribute in the corpus XML file.

3. *Entity Type*: The biomedical type or class label, which is retrieved from the *sem* attribute.
4. *Well-formed*: Name entities can be categorized into two genres. These two genres are not mutually exclusive; therefore a named entity can belong to both of them. These genres are:
  - a. *Well-formed named entity* is a named entity which at least one of the rules in Table 4 is satisfied. A noun group is called a *well-formed noun group* if it contains at least one well-formed named entity.
  - b. *Frequency named entity* is a non-well-formed named entity that contains at least one noun with a frequency more than a threshold value over the GENIA abstracts. A noun group is called a *frequency noun group* if it contains at least one frequent named entity.
5. *Frequency*: Indicates the frequency for each of the noun that belongs to a named entity. This frequency is counted over that 2000 GENIA abstracts.
6. *Main Index*: Index of the *main noun* from the related noun group. This index is assigned as follows:
  - a. If the noun group is a *well-formed*, then this value is the index of the noun or word that is well-formed. For example in the first row of Table 6, the well-formed part is “PKC” hence the index is 1.
  - b. For *frequency* noun group, the value is the index of the noun which has the highest frequency. For example in the second row of Table 6, the highest frequency is *Cell* so the index is 0.

**5.2 GEUSS Phase**

All of the noun groups or noun phrases of unstructured biomedical text which are called *candidates* are extracted and checked whether they contain a biomedical named entity. This phase contains three steps which are explained below.

**Tokenization**

This process requires a tokenizer. In this study, a regular expression based algorithm is used to extract words among the sentences. In Table 7 the regular expression used in this study is presented. Based on this regular expression a precise splitting is done and words are extracted.

**Table 7.** Regular Expression for Extracting Words

Regular Expression for Extracting Words
" ( [ \\ \t { } ( ) : ; . & ^ % \$ # @   < > _ , \ \ - ! \ " ? \ n ] ) "

**Extract Noun Groups**

Considering the noun group instead of a complete sentence and check word by word can significantly reduce the time complexity of the algorithm. All of the extracted noun groups are then kept in a list. Some samples of outputs for this algorithm after executing over sample in Table 8 are presented in Table 9.

**Table 8.** Example of Input Text

Example of a Biomedical Text
<i>“Protein kinase C is not a downstream effector of p21ras in activated T cells”</i>

**Table 9.** Sample of Extracted Noun Group from Input Text

Noun Group
Protein kinase C
downstream effector
p21ras
activated T cells

A noun group contains collection of nouns, adjectives, and past participle verbs. Also a word or words which satisfied one of the rules in Table 4 is considered as a noun or a noun group.

In summary all of the extracted noun groups are kept in a list. As mentioned earlier, the main idea of GAT is considering the noun group as a good candidate for a named entity, i.e. a named entity is a sequence of nouns or adjectives and past principle verbs. Therefore, first we extract all of the noun groups by considering “(VBN)\*(JJ)\*(NN)+” as regular expression and by using the corpus, a biomedical type is then identified.

### Generate Candidate Set

GAT categorizes the extracted noun group into *welled-form* and *frequency* categories. These two groups are called *welled-form candidate* and *frequency candidate*, respectively and are explained below.

1. *Welled-form candidate* is a noun group that at least one of its nouns follows one of the rules in Table 4.
2. *Frequency candidate* is a non-welled-form noun group that has one word with a frequency more than a threshold value over the BioDic which is explained in subsection 5.1.1.

These extracted noun groups are arranged in the same format that is explained in subsection 5.1 The process of creating this data set is the same as the process presented in subsection 5.1.

### 5.3 TRY Phase

The aim of this phase is to recognize biomedical type or class (gene and protein) for each candidate by extracting one or more matched through the training data set.

The major challenge in GAT is to recognize when a noun phrase candidate as D2 and a known biomedical named entity like DII designate the same entity.

In this study a syntax similarity measurement called *Statistical Character-Based Syntax Similarity (SCSS)* is proposed which by focusing on the natural language syntax and standard naming rules (Table 4) finds a common part between two patterns. This function accepts two patterns, one is a well known pattern from the training set and the other is a candidate pattern. The function measures how much that the two patterns are similar or refer to the same biomedical named entity.

### SCSS Function

The SCSS is independent from the text domain, computes the similarity probability by using four main functions. These four functions focus on word-character and word-order variation as presented in Table 5. Before explaining the four main functions, it is necessary to define five primitive functions that are used by the four main functions.

#### Function 1 Sign

---

$$\text{Sign}(x) = f(x) = \begin{cases} 1 & x \text{ is true} \\ 0 & x \text{ is false} \end{cases}$$

Sign function accepts a proposition and validates it.

**Example:** Sign (5 > 2) = 1; Sign (6 is Odd) = 0

---

#### Function 2 Len

---

Len(X) = Number of characters in string pattern X.

**Example:** Len("abc") = 3

---

#### Function 3 Indexing

---

$X_i$  = The  $i$ th character of X or the  $i$ th position of the string pattern X. For negative  $i$  the value of  $X_i$  is equal to  $X_0$  or the first character of X.

**Example:** If X = "abcdef",  $X_2 = 'c'$

---

#### Function 4 Common Character

---

$$\varphi(X, Y, i) = j, 1 \leq i \leq \text{Len}(X)$$

Returns the Y-position of the first occurrence of the  $i$ th common character between X and Y. If there is no common character at the given order the function returns -1.

**Example:** If X = 'abcd' and Y = 'bdecjhdf', then  $\varphi(X, Y, 1) = 0$  since the first common character between X and Y is 'b' and its index in string pattern Y is 0. While  $\varphi(X, Y, 2) = 3$  since the second common character between X and Y is 'c' and its index in string pattern Y is 3.

---

**Function 5 Position**


---

$$\beta(X, c) = I, (X_i = c)$$

This function returns the position of the *first* occurrence of the character  $c$  inside  $X$ .

**Example:** If  $X = \text{'abcdefg'}$ ,  $\beta(X, \text{'d'}) = 3$ , in fact  $X_3 = \text{'d'}$  (5)

---

Now, the four main functions that are necessary for SCSS are explained.

**1. Ordered Character Distribution (OCD)**

- i. Returns the ratio of the common characters between two patterns by keeping their order. This ratio is based on length of first string pattern.
- ii. Example: Consider  $X = \text{"CD23"}$  and  $Y = \text{"C3FD24"}$ . By keeping the order, characters 'C', 'D' and '2' are common characters. However '3' is a common character but it should not be considered because of its position in  $Y$  is 1 which is less than 4 the index of the last common
- iii. Therefore, OCD character '2'.  $(X, Y) = 3/4$ .

**2. Unordered Character Distribution (UCD)**

- i. Returns the ratio of the common characters between two patterns without keeping their order. This ratio is based on length of first string pattern.
- ii. Example: Consider  $X = \text{"C3D"}$  and  $Y = \text{"C1FD34"}$ . By keeping the order, characters 'C', 'D' are common characters, also in this case '3' is a common character while it does not follow the order. Therefore  $UCD(X, Y) = 3/3$ .

**3. Distribution Convergence Density (DCD)**

- i. Returns the ratio of convergence of the characters distribution from the first string pattern in the second string pattern.
- ii. Example: Consider  $X = \text{"CD24"}$  and  $Y = \text{"Cell Digestion4-L"}$ . Three characters of  $X$  appear in  $Y$  in position 0, 5, 14 and the total distance is  $(14 - 5) + (5 - 0) = 14$  and the  $DCD(X, Y)$  is  $14/17$ .
- iii. Character order is considered in this function.

**4. Symmetric Character (SC)**

- i. Returns the ratio of the common characters in the same position. They are called *symmetric*.
- ii. Example: For  $\text{"CD24"}$  and  $\text{"Cs2D4"}$  the symmetric characters are 'C' and '2'. So SC is 2 divides by 4 where 4 is the length of  $\text{"CD24"}$ .



## SCSS Algorithm

Basically SCSS computes the dissimilarity probability of two patterns and then returns its complementary which represents similarity ( $P(X) = 1 - P(\bar{X})$ ). The SCSS function can be forced to consider case sensitivity depending on the requirement of the domain. This function by giving two string patterns computes the similarity based on the four mentioned functions in section 5.3.1. For this computation, we use *harmonic mean* over the result of those four functions.

The value returned by the SCSS is compared to a *similarity threshold* and if it exceeds the threshold, the biomedical type or class (gene or protein) of the well-known named entity is assigned to the candidate named entity including the probability of similarity.

## 6 Result

The GENIA corpus includes 2000 abstracts. In this study, these 2000 abstracts are divided into two groups of 1,800 abstracts as training set and 200 abstracts as test set. The evaluation process is done over these well known 200 abstracts. The evaluation method is a method commonly used in most of previous researches as Tsuruoka and Tsujii [1]. The evaluation of GAT performance is based on two measures: *recall* and *precision*.

This study achieves the F-measure of 94.5% on GENIA 3.0 on both *gene* and *protein* names and is summarized in Table 10.

In particular, this study achieves the F-measure of 98.4% on the protein class and is summarized in Table 11 and this is compared to Fukuda [10] and PASTA [12] which focused on the protein names.

**Table 10.** Evaluation Result (Gene and Protein)

Performance	Precision	Recall	F-Test
GAT on GENIA V3.0	96.7%	92.4%	94.5
Tsuruoka and Tsujii [1]	71.7%	62.3%	66.6
Hanisch [11]	95.0%	90.0%	92.4
Morgan [4]	78.0%	88.0%	82.4

**Table 11.** Evaluation Result (Recognize Protein)

Performance	Precision	Recall	F-Test
GAT on GENIA V3.0	97.7%	99.1%	98.4
PASTA [12]	97.0%	87.0%	91.7
Fukuda [10] just for protein	94.7%	98.8%	96.7

## 7 Conclusion

In this paper, an approach for named entity recognition (NER) for the biomedical texts called GAT is described. Various complexities like variation, ambiguity and newly discovered names are incorporated to cope with the special phenomena in biomedical named entities. It is clear that the NER for biomedical texts is complex, but it does not mean that it is not possible to design a domain-independent algorithm for NER over biomedical domain.

This study assumes that each named entity occurs among a noun group. Extracting noun groups is done using the Brill Part of Speech tagger. The main strength of GAT is a Statistical Character-Based Syntax Similarity function which lies on the appropriate similarity definition from the syntax viewpoint (which is the nature of biomedical name entities) that considers the various similarity features, including the ratio of common characters, the ratio of symmetric common characters, and the distribution of common characters between two string patterns.

For future work the algorithm can be improved for other domain especially social science by extracting named entities based on their semantics. These categories of domain in contrast to biomedical do not follow lexical rules for naming and thus a semantic based approach is needed.

## References

1. Tsuruoka, Y., Tsujii, J.: Improving the Performance of Dictionary-Based Approaches in Protein Name Recognition. *Journal of Biomedical Informatics* 37, 461–470 (2004)
2. Krauthammer, M.: Using BLAST for Identifying Gene and Protein Names in Journal Articles. *Journal of Gene* 259(1–2), 245–252 (2000)
3. Collier, N., Nobata, C., Tsujii, J.: Extracting the Names of Genes and Gene Products with a Hidden Markov Model. In: *Proceedings of the 17th International Conferences on Computational Linguistics*, pp. 201–207 (2000)
4. Morgan, A.: Gene Name Identification and Normalization using a Model Organism Database. *Journal of Biomedical Informatics* 37, 396–410 (2004)
5. Zhou, G.D., Su, J.: Named Entity Recognition using an HMM-Based Chunk Tagger. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 473–480 (July 2002)
6. Proux, D.: Detecting Gene Symbols and Names in Biomedical Texts: A First Step Toward Pertinent Information. In: *Proceedings of the 9th Workshop on Genome Informatics*, pp. 72–80 (1998)
7. Chinchor, N.: MUC-7 Information Extraction Task Definition. In: *Proceedings of the 7th Message Understanding Conf.* (1998)
8. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A Brief History. In: *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 466–471 (1996)
9. Kim, J.D.: GENIA Corpus—A Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics* 19(suppl. 1), i180–i182 (2003)
10. Fukuda: Towards Information Extraction: Identifying Protein Names from Biological Papers. In: *Proceedings of the Pacific Symp. on Biocomputing, Wailea, HI*, pp. 707–718 (1998)

11. Hanisch, D.: Playing Biology's Name Game: Identifying Protein Names in Scientific Text. In: Hanisch, D. (ed.) Proceedings of the Pacific Symp. on Biocomputing, pp. 403–414 (2003)
12. Gaizauskas, R., Demetriou, G., Humphreys, K.: Protein Structures and Information Extraction from Biological Texts: The PASTA System. *Journal of Bioinformatics* 19(1), 135–143 (2003)
13. Drymonas, E., Zervanou, K., Petrakis, E.G.M.: Exploiting Multi-Word Similarity for Retrieval in Medical Document Collections: the TSRM Approach. *Journal of Digital Information Management* 8(5), 315–321 (2010)

# Correction of Invalid XML Documents with Respect to Single Type Tree Grammars\*

Martin Svoboda and Irena Mlýnková

Department of Software Engineering, Charles University in Prague  
Malostranske namesti 25, 118 00 Prague 1, Czech Republic  
{svoboda,mlynkova}@ksi.mff.cuni.cz

**Abstract.** XML documents and related technologies represent a widely accepted standard for managing semi-structured data. However, a surprisingly high number of XML documents is affected by well-formedness errors, structural invalidity or data inconsistencies. The aim of this paper is the proposal of a correction framework involving structural repairs of elements with respect to single type tree grammars. Via the inspection of the state space of a finite automaton recognising regular expressions, we are always able to find all minimal repairs against a defined cost function. These repairs are compactly represented by shortest paths in recursively nested multigraphs, which can be translated to particular sequences of edit operations altering XML trees. We have proposed an efficient algorithm and provided a prototype implementation.

**Keywords:** XML, correction, validity, grammar, tree.

## 1 Introduction

XML documents [10] and related standards represent without any doubt an integral part of the contemporary World Wide Web technologies. They are used for data interchange, sharing knowledge or for storing semi-structured data. However, the XML usage explosion is accompanied with a surprisingly high number of documents involving various forms of errors [5].

These errors can cause that the given documents are not well-formed, they do not conform to the required structure or have inconsistencies in data values. Anyway, the presence of errors causes at least obstructions and may completely prevent successful processing. Generally we can modify existing algorithms to deal with errors, or we can attempt to modify invalid documents themselves.

We particularly focus on the problem of the structural invalidity of XML documents. In other words we assume the inspected documents are well-formed and constitute trees, however, these trees do not conform to a schema in DTD [10] or XML Schema [4], i.e. a regular tree grammar with the expressive power at the level of single type tree grammars [6]. Having a potentially invalid XML

---

\* This work was partially supported by the Czech Science Foundation (GAČR), grants number 201/09/P364 and P202/10/0573.

document, we process it from its root node towards leaves and propose minimal corrections of elements in order to achieve a valid document close to the original one. In each node of a tree we attempt to statically investigate all suitable sequences of its child nodes with respect to a content model and once we detect a local invalidity, we propose modifications based on operations capable to insert new minimal subtrees, delete existing ones or recursively repair them.

**Related Work.** The proposed framework is based primarily on ideas from [1] and [9]. Authors of the former paper dynamically inspect the state space of a finite automaton for recognising regular expressions in order to find valid sequences of child nodes with minimal distance. However, this traversal is not effective, requires a threshold pruning to cope with potentially infinite trees, repeatedly computes the same repairs and acts efficiently only in the context of incremental validation. Although these disadvantages are partially handled in the latter paper, its authors focused on documents querying, but not repairing.

Next, we can mention an approximate validation and correction approach [11] based on testers and correctors from the theory of program verification. Repairs of data inconsistencies like functional dependencies, keys and multivalued dependencies are the subject of [8,12].

**Contributions.** Contrary to all existing approaches we consider single type tree grammars instead only local tree grammars. Thus we work both with DTD and XML Schema. Approaches in [1,11] are not able to find repairs of more damaged documents, we are able to always find all minimal repairs and even without any threshold pruning to handle potentially infinite XML trees. Next, we have proposed much more efficient algorithm following only perspective ways of the correction and without any repeated repair computations. Finally, we have a prototype implementation available at [3] and performed experiments show a linear time complexity depending on a number of nodes in documents.

**Outline.** In Section 2 we define a formal model of XML documents and schemata as regular tree grammars. Section 3 introduces the entire proposed correction framework and Section 4 concludes this paper.

## 2 Preliminaries

In this section, we introduce preliminary definitions used in this paper.

### 2.1 XML Trees

Analogously to [1], we represent XML documents as data trees based on underlying trees with prefix numbering of nodes.

**Definition 1 (Underlying Tree).** Let  $\mathbb{N}_0^*$  be the set of all finite words over the set of non-negative integers  $\mathbb{N}_0$ ,  $\epsilon$  be an empty word and  $.$  a concatenation. A set  $D \subset \mathbb{N}_0^*$  is an underlying tree or just tree, if the following conditions hold:

- $D$  is closed under prefixes, i.e. having a binary prefix relation  $\preceq$  (where  $\forall u, v \in \mathbb{N}_0^*$  we define  $u \preceq v$  if  $u.w = v$  for some  $w \in \mathbb{N}_0^*$ ) we require that  $\forall u, v \in \mathbb{N}_0^*$ ,  $u \preceq v$ :  $v \in D$  implies  $u \in D$ .
- $\forall u \in \mathbb{N}_0^*$ ,  $\forall j \in \mathbb{N}_0$ : if  $u.j \in D$  then  $\forall i \in \mathbb{N}_0$ ,  $0 \leq i \leq j$ ,  $u.i \in D$ .

We say that  $D$  is an empty tree, if  $D = \emptyset$ . Elements of  $D$  are called nodes, node  $\epsilon$  is a root node and  $LeafNodes(D) = \{u \mid u \in D \text{ and } \neg \exists i \in \mathbb{N}_0 \text{ such that } u.i \in D\}$  represents a set of leaf nodes.

Given a node  $u \in D$  we define  $fanOut(u)$  as  $n \in \mathbb{N}_0$  such that  $u.(n-1) \in D$  and  $\neg \exists n' \in \mathbb{N}$ ,  $n' > n-1$  such that  $u.n' \in D$ . If  $u.0 \notin D$ , we put  $n = 0$ . Finally, we define  $D_p = \{s \mid s \in \mathbb{N}_0^*, p.s \in D\}$  as a subtree of  $D$  at position  $p$ .

Since we are interested only in elements, we ignore attributes. Data values and element labels are modelled as partial functions on underlying nodes.

**Definition 2 (Data Tree).** Let  $D$  be an underlying tree,  $\mathbb{V}$  a domain for data values and  $\mathbb{E}$  a domain of element labels (i.e. set of distinct element names). Tuple  $\mathcal{T} = (D, lab, val)$  is a data tree, if the following conditions are satisfied:

- $lab$  is a labelling function  $D \rightarrow \mathbb{E} \cup \{\mathbf{data}\}$ , where  $\mathbf{data} \notin \mathbb{E}$ :
  - $DataNodes(\mathcal{T}) = \{p \in D \mid lab(p) = \mathbf{data}\}$  is a set of data nodes.
  - If  $p \in DataNodes(\mathcal{T})$ , then necessarily  $p \in LeafNodes(D)$ .
- $val$  is a function  $DataNodes(\mathcal{T}) \rightarrow \mathbb{V} \cup \{\perp\}$  assigning values to data nodes, where  $\perp \notin \mathbb{V}$  is a special symbol representing undefined values.

Finally, we define  $\mathcal{T}_p = (D', lab', val')$  as a data subtree of  $\mathcal{T}$  at position  $p$ , where  $D' = D_p$  and for each function  $\phi \in \{lab, val\}$ : if  $\phi(p.s)$  is defined, then  $\phi'(s) = \phi(p.s)$ , where  $s \in \mathbb{N}_0^*$ .

*Example 1.* In Figure 1 we can find sample data tree  $\mathcal{T}$  based on an underlying tree  $D = \{\epsilon, 0, 0.0, 1, 1.0, 1.1\}$ . Values of  $lab$  function are inside nodes, an implicit tree structure is depicted using edges. Ignoring  $val$  function this data tree corresponds to an XML fragment:  $\langle a \rangle \langle x \rangle \langle d \rangle \langle /x \rangle \langle d \rangle \langle /d \rangle \langle /a \rangle$ .

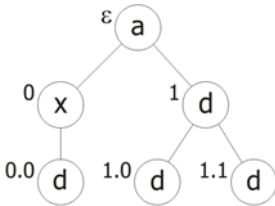


Fig. 1. Sample data tree

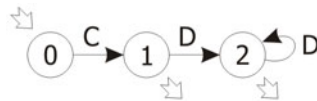


Fig. 2. Glushkov automaton for  $C.D^*$

## 2.2 Regular Expressions

Schemata for XML documents especially restrict nesting of elements through allowed content models. These are based on regular expressions.

**Definition 3 (Regular Expression).** Let  $\Sigma$  be a finite nonempty alphabet and  $S = \{\emptyset, \epsilon, |, \cdot, *, (, )\}$ , such that  $\Sigma \cap S = \emptyset$ . We inductively define a regular expression  $r$  as a word in  $\Sigma \cup S$  and  $L(r)$  as an associated language:

- $r \equiv \emptyset$  and  $L(\emptyset) = \emptyset$ .  $r \equiv \epsilon$  and  $L(\epsilon) = \{\epsilon\}$ .  $\forall x \in \Sigma: r \equiv x$  and  $L(x) = \{x\}$ .
- $r \equiv (r_1|r_2)$  and  $L(r_1|r_2) = L(r_1) \cup L(r_2)$ ,
- $r \equiv (r_1.r_2)$  and  $L(r_1.r_2) = L(r_1).L(r_2)$ ,
- $r \equiv r_1^*$  and  $L(r_1^*) = (L(r_1))^*$ ,

where  $r_1$  and  $r_2$  are already defined regular expressions. Having an expression  $r = s_1 \dots s_n$ , we define  $\text{symbols}(r) = \{s_i \mid \exists i \in \mathbb{N}_0, 1 \leq i \leq n, s_i \in \Sigma\}$ .

Languages of regular expressions can be recognised by finite automata. We use Glushkov automata [2], because they are deterministic for 1-unambiguous regular expressions required by DTD and XML Schema and without  $\epsilon$ -transitions.

**Definition 4 (Glushkov Automaton).** The Glushkov automaton for a 1-unambiguous regular expression  $r$  over an alphabet  $\Sigma$  is a deterministic finite automaton  $\mathcal{A}_r = (Q, \Sigma, \delta, q_0, F)$ , where  $Q = \Sigma' \cup \{q_0\}$  is a set of states,  $\Sigma$  is an input alphabet,  $\delta$  is a partial transition function  $Q \times \Sigma \rightarrow Q$ ,  $q_0 \in Q$  is an initial state and  $F \subseteq Q$  is a set of accepting states.

*Example 2.* The Glushkov automaton  $\mathcal{A}_r$  for regular expression  $r = C.D^*$  over  $N_R = \{C, D\}$  is depicted in Figure 2. This automaton has states  $Q = \{0, 1, 2\}$ , from which  $q_0 = 0$  is the initial state and  $F = \{1, 2\}$  are accepting states. The transition function  $\delta$  is represented by directed edges between states.

## 2.3 Tree Grammars

Adopting and slightly modifying the formalism from [6], we represent schemata in DTD and XML Schema as *regular tree grammars*.

**Definition 5 (Regular Tree Grammar).** A regular tree grammar is a tuple  $\mathcal{G} = (N, T, S, P)$ , where:

- $N$  is a set of nonterminal symbols and  $T$  a set of terminal symbols,
- $S \subseteq N$  is a set of starting symbols,
- $P$  is a set of production rules of the form  $[a, r \rightarrow n]$ , where  $a \in T$ ,  $r$  is a 1-unambiguous regular expression over  $N$  and  $n \in N$ . Without loss of generality, for each  $a \in T$  and  $n \in N$  there exists at most one  $[a, r \rightarrow n] \in P$ .

**Definition 6 (Competing Nonterminals).** Let  $\mathcal{G} = (N, T, S, P)$  be a regular tree grammar and  $n_1, n_2 \in N$ ,  $n_1 \neq n_2$  are two nonterminal symbols. We say that  $n_1$  and  $n_2$  are competing with each other, if there exist two production rules  $[a, r_1 \rightarrow n_1], [a, r_2 \rightarrow n_2] \in P$  sharing the same terminal symbol  $a$ .

The presence of competing nonterminals makes the processing more complicated, thus we define two main subclasses with less expressive power.

**Definition 7 (Tree Grammar Classes).** Let  $\mathcal{G} = (N, T, S, P)$  be a regular tree grammar. We say that  $\mathcal{G}$  is a local tree grammar, if it has no competing nonterminal symbols, and that  $\mathcal{G}$  is a single type tree grammar, if for each production rule  $[a, r \rightarrow n]$  all nonterminal symbols in  $r$  do not compete with each other and starting symbols in  $S$  do not compete with each other too.

As a consequence, we do not need to distinguish between terminal and nonterminal symbols in local tree grammars. DTD schemata correspond to local tree grammars and XML Schema almost to single type tree grammars [6].

*Example 3.* Following the data tree from Example 1 we can introduce grammar  $\mathcal{G}$ , where  $N = \{A, B, C, D\}$  are nonterminals,  $T = \{a, b, c, d\}$  are terminals and  $S = \{A, B\}$  are starting symbols. The set  $P$  contains these transition rules:  $\mathcal{F}_1 = [a, C.D^* \rightarrow A]$ ,  $\mathcal{F}_2 = [b, D^* \rightarrow B]$ ,  $\mathcal{F}_3 = [c, \emptyset \rightarrow C]$  and  $\mathcal{F}_4 = [d, D^* \rightarrow D]$ . Since there are no competing nonterminals, this grammar is a local tree grammar.

## 2.4 Data Trees Validity

The validity of data trees can be defined via the existence of interpretation trees.

**Definition 8 (Interpretation Tree).** Let  $\mathcal{T} = (D, lab, val)$  be a data tree and  $\mathcal{G} = (N, T, S, P)$  be a regular tree grammar. An interpretation tree of a data tree  $\mathcal{T}$  against grammar  $\mathcal{G}$  is a tuple  $\mathcal{N} = (D, int)$ , where  $D$  is the original underlying tree and  $int$  is a function  $D \rightarrow N$  satisfying the following conditions:

- $\forall p \in D$  there exists a rule  $[a, r \rightarrow n] \in P$  such that  $int(p) = n$ ,  $lab(p) = a$  and  $int(p.0).int(p.1) \dots int(p.k) \in L(r)$ , where  $k = fanOut(p) - 1$ .
- If  $p = \epsilon$  is the root node, then we moreover require  $int(p) \in S$ .

**Definition 9 (Data Tree Validity).** We say that a data tree  $\mathcal{T} = (D, lab, val)$  is valid against a regular tree grammar  $\mathcal{G} = (N, T, S, P)$ , if there exists at least one interpretation  $\mathcal{N}$  of  $\mathcal{T}$  against  $\mathcal{G}$ . Given a node  $p \in D$ , we say that  $p$  is locally valid, if  $\mathcal{T}_p^{tree}$  is valid against grammar  $\mathcal{G}' = (N, T, N, P)$ .

By  $L(\mathcal{G})$  we denote a local, single type or regular tree language, i.e. a set of all trees valid against a regular, single type or local tree grammar  $\mathcal{G}$  respectively.

*Example 4.* The data tree from Example 1 represented in Figure 1 is not valid against  $\mathcal{G}$  from Example 3, especially because  $lab(0) \notin T$  and thus there can not exist any production rule to be used for interpretation of node 0.

**Definition 10 (Grammar Context).** Let  $\mathcal{G} = (N, T, S, P)$  be a single type tree grammar and  $\mathcal{F} = [a, r \rightarrow n] \in P$ . We define  $\mathcal{C}_{\mathcal{F}} = (a, n, N_R, P_R, map, r)$  to be a general context of grammar  $\mathcal{G}$  for rule  $\mathcal{F}$ , where:



- $N_R = \{x \mid x \in \text{symbols}(r)\}$  is a set of allowed nonterminal symbols.
- $P_R = \{\mathcal{F}' \mid \mathcal{F}' = [a', r' \rightarrow n'] \in P \text{ and } n' \in N_R\}$  is a set of active rules.
- $\text{map}$  is a function  $T \rightarrow N_R \cup \{\perp\}$  such that  $\forall \mathcal{F}' = [a', r' \rightarrow n'] \in P_R$ :  $\text{map}(a') = n'$  and for all other  $a' \in T$ :  $\text{map}(a') = \perp$  (where  $\perp \notin N$ ).

Next, we define a starting context to be  $\mathcal{C}_\bullet = (\perp, \perp, N_R, P_R, \text{map}, r_\bullet)$ , where  $N_R = S$ , both  $P_R$  and  $\text{map}$  are defined standardly and  $r_\bullet = (n_1 \mid \dots \mid n_s)$  is a starting regular expression meeting  $s = |S|$ ,  $\forall i \in \mathbb{N}$ ,  $1 \leq i \leq s$ ,  $n_i \in S$  and  $\forall i, j \in \mathbb{N}$ ,  $1 \leq i < j \leq s$ ,  $n_i \neq n_j$ . Finally,  $\mathcal{C}_\emptyset = (\perp, \perp, \emptyset, \emptyset, \text{map}, r_\emptyset)$  is an empty context, where  $r_\emptyset = \emptyset$  and  $\text{map}$  is defined standardly again.

*Example 5.* Having production rule  $\mathcal{F}_1$  of grammar  $\mathcal{G}$  from Example 3, we can derive its context  $\mathcal{C}_1 = (a, A, \{C, D\}, \{\mathcal{F}_3, \mathcal{F}_4\}, \{(a, \perp), (b, \perp), (c, C), (d, D)\}, C.D^*)$ . Since  $S = \{A, B\}$  are starting symbols, the starting context is equal to  $\mathcal{C}_\bullet = (\perp, \perp, \{A, B\}, \{\mathcal{F}_1, \mathcal{F}_2\}, \{(a, A), (b, B), (c, \perp), (d, \perp)\}, A|B)$ .

During the data trees correction, being in each particular node, the local correction possibilities are defined by a corresponding grammar context. However, the content model is represented as a regular expression over nonterminal symbols, thus we first need to transform the labels of existing nodes to nonterminals.

**Definition 11 (Node Sequence Imprint).** Let  $\mathcal{T} = (D, \text{lab}, \text{val})$  be a data tree and  $u = \langle u_1, \dots, u_k \rangle$  a sequence of nodes for some  $k \in \mathbb{N}_0$ , where  $\forall i \in \mathbb{N}$ ,  $1 \leq i \leq k$ ,  $u_i \in D$ . We define an imprint of  $u$  in context  $\mathcal{C} = (a, n, N_R, P_R, \text{map}, r)$  to be sequence imprint( $u$ ) =  $\langle \text{map}(\text{lab}(u_1)), \dots, \text{map}(\text{lab}(u_k)) \rangle$ .

*Example 6.* Suppose that  $u = \langle 0, 1 \rangle$  is a sequence of child nodes of the root node in data tree  $\mathcal{T}$  from Example 1. Labels of these nodes are  $\langle x, d \rangle$ . An imprint of  $u$  in  $\mathcal{C}_1$  from Example 5 is a sequence  $\langle \perp, D \rangle$ .

### 3 Corrections

In order to propose a new correction framework, we especially need to introduce a model of allowed data trees transformations and efficient algorithms.

#### 3.1 Edit Operations

First, we define several auxiliary sets of nodes, which become useful in a definition of such allowed transformations, called *edit operations*.

**Definition 12 (Auxiliary Node Sets).** Given a tree  $D$  and a position  $p \in D$ ,  $p \neq \epsilon$ ,  $p = u.i$ ,  $u \in \mathbb{N}_0^*$ ,  $i \in \mathbb{N}$  with  $f = \text{fanOut}(u)$ , we define node sets:

- $\text{PosNodes}(D) = \{u.i \mid i \in \mathbb{N}_0, u.i \notin D, u \in D \text{ and either } i = 0 \text{ or } i > 0 \text{ and } u.(i - 1) \in D\}$ . If  $D$  is an empty tree, then  $\text{PosNodes}(D) = \{\epsilon\}$ .
- $\text{ExpNodes}(D, p) = \{u.k.v \mid k \in \mathbb{N}_0, i \leq k < f, v \in \mathbb{N}_0^*, u.k.v \in D\}$ .

- $IncNodes(D, p) = \{u.(k+1).v \mid k \in \mathbb{N}_0, i \leq k < f, v \in \mathbb{N}_0^*, u.k.v \in D\}$ .
- $DecNodes(D, p) = \{u.(k-1).v \mid k \in \mathbb{N}_0, i+1 \leq k < f, v \in \mathbb{N}_0^*, u.k.v \in D\}$ .

Set  $PosNodes$  together with the underlying tree represent positions ready for insertions, whereas nodes in  $ExpNodes$  are transferred to  $IncNodes$  or  $DecNodes$  after a performed insertion or deletion respectively.

*Example 7.* Having a data tree  $\mathcal{T}$  from Example 1 and  $p = 0$  we can derive  $PosNodes(D) = \{0.0.0, 0.1, 1.0.0, 1.1.0, 1.2, 2\}$ ,  $ExpNodes(D, 0) = \{0, 0.0, 1, 1.0, 1.1\}$  and  $IncNodes(D, 0) = \{1, 1.0, 2, 2.0, 2.1\}$ .

Although we have considered also an internal node insertion/deletion in [7], we focus only on a leaf node insertion/deletion and node renaming in this paper. For the purpose of the following definition of allowed edit operations, we use a symbol  $\leftarrow$  as an assignment conditioned by the definability.

**Definition 13 (Edit Operations).** *An edit operation  $e$  is a partial function transforming a data tree  $\mathcal{T}_0 = (D_0, lab_0, val_0)$  into  $\mathcal{T}_1 = (D_1, lab_1, val_1)$ , denoted by  $\mathcal{T}_0 \xrightarrow{e} \mathcal{T}_1$ . Assuming that  $\phi \in \{lab, val\}$ , we define these operations:*

- $e = addLeaf(p, a)$  for  $p \in D_0 \cup PosNodes(D_0)$ ,  $p \neq \epsilon$ ,  $p = u.i$ ,  $u \in \mathbb{N}_0^*$ ,  $i \in \mathbb{N}_0$ ,  $u \notin DataNodes(D_0)$  and  $a \in \mathbb{E}$ :
  - $D_1 = [D_0 \setminus ExpNodes(D_0, p)] \cup IncNodes(D_0, p) \cup \{p\}$ .
  - $\forall w \in D_0 \setminus ExpNodes(D_0, p): \phi_1(w) \leftarrow \phi_0(w)$ .
  - $lab_1(p) = a$  and if  $lab_1(p) = \mathbf{data}$ , then  $val_1(p) = \perp$ .
  - $\forall (u.(k+1).v) \in IncNodes(D_0, p): \phi_1(u.(k+1).v) \leftarrow \phi_0(u.k.v)$ .
- $e = addLeaf(p, a)$  for  $p = \epsilon$ ,  $D_0 = \emptyset$  and  $a \in \mathbb{E}$ :
  - $D_1 = \{p\}$ ,  $lab_1(p) = a$  and if  $a = \mathbf{data}$ , then  $val_1(p) = \perp$ .
- $e = removeLeaf(p)$  for  $p \in LeafNodes(D_0)$ ,  $p \neq \epsilon$ ,  $p = u.i$ ,  $u \in \mathbb{N}_0^*$ ,  $i \in \mathbb{N}_0$ :
  - $D_1 = [D_0 \setminus ExpNodes(D_0, p)] \cup DecNodes(D_0, p)$ .
  - $\forall w \in D_0 \setminus ExpNodes(D_0, p): \phi_1(w) \leftarrow \phi_0(w)$ .
  - $\forall (u.(k-1).v) \in DecNodes(D_0, p): \phi_1(u.(k-1).v) \leftarrow \phi_0(u.k.v)$ .
- $e = removeLeaf(p)$  for  $p = \epsilon$ ,  $D_0 = \{\epsilon\}$ :
  - $D_1 = \emptyset$ ,  $lab_1$  and  $val_1$  are not defined anywhere.
- $e = renameLabel(p, a)$  for  $p \in D_0$ ,  $a \in \mathbb{E}$  and  $a \neq lab_0(p)$ :
  - $D_1 = D_0$ .
  - $\forall w \in [D_0 \setminus \{p\}]: \phi_1(w) \leftarrow \phi_0(w)$ .
  - $lab_1(p) = a$  and if  $a = \mathbf{data}$ , then  $val_1(p) = \perp$ .

Combining edit operations into *edit sequences*, we obtain complex operations capable to insert new subtrees, delete existing ones or recursively repair them.

*Example 8.* Assume that we have edit sequences  $\mathcal{X}_1 = \langle addLeaf(0, c), renameLabel(1, d) \rangle$ ,  $\mathcal{X}_2 = \langle renameLabel(0, c), removeLeaf(0.0) \rangle$  and  $\mathcal{X}_3 = \langle renameLabel(\epsilon, b), renameLabel(0, d) \rangle$ . Applying these sequences separately to data tree  $\mathcal{T}$  from Example 1, we obtain data trees depicted in Figures 3(a), 3(b) and 3(c) respectively. Auxiliary node sets for  $addLeaf(0, c)$  are derived in Example 7.

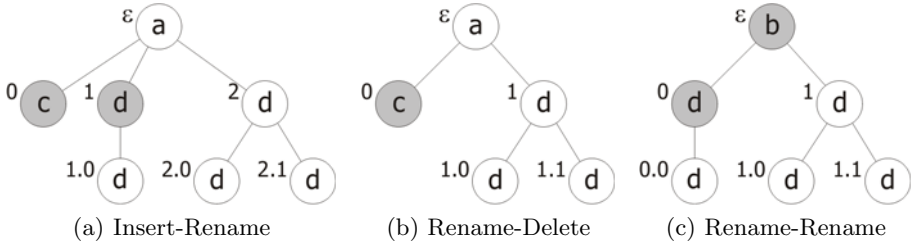


Fig. 3. Transforming sample data tree using edit operations

**Definition 14 (Cost Function).** Given an edit operation  $e$ , we define  $cost(e)$  to be a function assigning to  $e$  its non-negative cost. Having a sequence of edit operations  $E = \langle e_1, \dots, e_k \rangle$  for some  $k \in \mathbb{N}_0$ , we define  $cost(E) = \sum_{i=1}^k cost(e_i)$ .

**Definition 15 (Data Tree Distance).** Assume that  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are two data trees and  $\mathcal{S}$  is a set of all sequences of update operations capable to transform  $\mathcal{T}_1$  to  $\mathcal{T}_2$ . We define distance of  $\mathcal{T}_1$  and  $\mathcal{T}_2$  to be  $dist(\mathcal{T}_1, \mathcal{T}_2) = \min_{E \in \mathcal{S}} cost(E)$ .

Given a regular tree grammar  $\mathcal{G}$  and the corresponding regular tree language  $L(\mathcal{G})$ , we define the distance between a tree  $\mathcal{T}_1$  and language  $L(\mathcal{G})$  as  $dist(\mathcal{T}_1, L(\mathcal{G})) = \min_{\mathcal{T}_2 \in L(\mathcal{G})} dist(\mathcal{T}_1, \mathcal{T}_2)$ .

The goal of the correction algorithm is to find all minimal repairs, i.e. edit sequences of minimal cost. Although the definition of distances talks about all sequences, the algorithm can clearly inspect only the perspective ones.

*Example 9.* Assigning unit costs to all edit operations, we can find out that  $dist(\mathcal{T}, L(\mathcal{G})) = 2$  for  $\mathcal{T}$  from Example 1 and grammar  $\mathcal{G}$  from Example 3.

Our algorithm is always able to find all such sequences and because we would like to represent found repairs compactly, we need to abstract away positions from edit operations. Thus we introduce repairing instructions, which need to be translated later on to edit operations over particular nodes.

**Definition 16 (Repairing Instructions).** For edit operations  $addLeaf(p, a)$ ,  $removeLeaf(p)$  and  $renameLabel(p, a)$  with  $p \in \mathbb{N}_0^*$  and  $a \in \mathbb{E}$  we define associated repairing instructions  $(addLeaf, a)$ ,  $(removeLeaf)$  and  $(renameLabel, a)$  respectively. Each repairing instruction is assigned with the corresponding cost.

### 3.2 Correction Intent

Assume that we are processing a sequence of sibling nodes in order to correct them. For this purpose we statically investigate the state space of the corresponding Glushkov automaton to find edit sequences transforming the original sequence and all nested subtrees with the minimal cost. Being on a given position, we have already considered the given sequence prefix. The notion of a correction intent involves the assignment for this recursive subtree processing.

**Definition 17 (Correction Intent).** Given  $\Omega = \{\text{correct, insert, delete, repair, rename}\}$  we define a correction intent to be a tuple  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$  satisfying the following general constraints:

- $t \in \Omega$  is an intent type,  $p$  is a base node and  $e$  is a repairing instruction.
- $v_I = (s_I, q_I)$ :  $s_I \in \mathbb{N}_0$  is an initial stratum and  $q_I$  is an initial state.
- $v_E = (s_E, q_E)$ :  $s_E \in \mathbb{N}_0$  is an ending stratum and  $q_E$  is an ending state.
- $u = \langle u_1, \dots, u_k \rangle$  is a sequence of nodes to be processed for some  $k \in \mathbb{N}_0$ .
- $\mathcal{C}$  is a grammar context and  $Q_T$  is a set of target states.
- $Y \subseteq \Omega$  is a set of allowed types for nested correction intents.

**Definition 18 (Starting Intent).** Having a data tree  $\mathcal{T} = (D, \text{lab}, \text{val})$  and a single type tree grammar  $\mathcal{G} = (N, T, S, P)$ , we define  $\mathcal{I}_\bullet = (\text{correct}, \perp, \perp, \perp, \perp, u, \mathcal{C}, Q_T, Y)$  to be a starting correction intent, where:

- If  $D$  is not empty, then  $u = \langle \epsilon \rangle$ , else  $u = \langle \rangle$ .
- $\mathcal{C} = \mathcal{C}_\bullet = (\perp, \perp, N_R, P_R, \text{map}, r_\bullet)$  is the starting context.
- $Q_T = F$  from the Glushkov automaton  $\mathcal{A}_r = (Q, N_R, \delta, q_0, F)$  for  $r_\bullet$ .
- $Y = \Omega \setminus \{\text{correct}\}$ .

The data tree correction starts at its root by the starting intent and recursively continues towards leaves by the invocation of nested recursive intents. Authors in [11], contrary to our framework, process data trees from leaves towards a root and attempt to correct only subtrees of locally invalid nodes.

**Definition 19 (Recursive Intents).** Let  $\mathcal{T} = (D, \text{lab}, \text{val})$  be a data tree and  $\mathcal{G} = (N, T, S, P)$  a single type tree grammar. Next, assume that  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$  is an already defined correction intent, where  $u = \langle u_1, \dots, u_k \rangle$ ,  $k \in \mathbb{N}_0$ ,  $\mathcal{C} = (a, n, N_R, P_R, \text{map}, r)$  and  $\mathcal{A}_r = (Q, N_R, \delta, q_0, F)$  is the Glushkov automaton for  $r$ . Finally, let  $\text{imprint}(u) = \langle m_1, \dots, m_k \rangle$ .

Given a position  $v'_I = (s'_I, q'_I)$ , where  $s'_I \in \mathbb{N}_0$ ,  $s'_I \leq k$ ,  $q'_I \in Q$ , we define the following recursive correction intents  $\mathcal{I}' = (t', p', e', v'_I, v'_E, u', \mathcal{C}', Q'_T, Y')$ :

- If  $\text{insert} \in Y$ :  $\forall x \in N_R$ : if  $\delta(q'_I, x)$  is defined, then we define  $\mathcal{I}'$ , where:
  - Let  $\mathcal{F} = [a', r' \rightarrow n'] \in P_R$  such that  $n' = x$  and  $\text{map}(a') = x$ .
  - $t' = \text{insert}$ ,  $p' = \perp$ ,  $e' = (\text{addLeaf}, a')$ ,  $v'_E = (s'_I, \delta(q'_I, x))$ ,  $u' = \langle \rangle$ .
  - $\mathcal{C}' = \mathcal{C}_{\mathcal{F}} = (a', n', N'_R, P'_R, \text{map}', r')$  with  $\mathcal{A}_{r'} = (Q', N'_R, \delta', q'_0, F')$ .
  - If  $r' \neq \emptyset$ , then  $Q'_T = F'$ , else  $Q'_T = \{q'_0\}$ .  $Y' = \{\text{insert}\}$ .

Suppose that  $\langle \mathcal{I}^1, \dots, \mathcal{I}^j \rangle$  is the longest sequence of correction intents for some  $j \in \mathbb{N}_0$ , such that  $\forall i \in \mathbb{N}$ ,  $1 \leq i < j$ ,  $t^i = \text{insert}$ ,  $\mathcal{I}^i$  invokes  $\mathcal{I}^{i+1}$  and  $\mathcal{I}^j = \mathcal{I}$ ,  $t^j = \text{insert}$ . We do not allow the previously described intent  $\mathcal{I}'$ , if  $\exists i$ ,  $1 \leq i \leq j$ :  $a^i = a'$  and  $n^i = x$  with symbols  $a^i$  and  $n^i$  from  $\mathcal{C}^i$ . Finally, we put  $\text{ContextChain}(\mathcal{I}) = \langle (a^1, n^1), \dots, (a^j, n^j) \rangle$ .

- If  $\text{delete} \in Y$  and  $s'_I < k$ , then we define  $\mathcal{I}'$ , where:
  - $t' = \text{delete}$ ,  $p' = u_{s'_I+1}$  and  $e' = (\text{removeLeaf})$ .
  - $v'_E = (s'_I + 1, q'_I)$  and  $u' = \langle u_{s'_I+1}.0, \dots, u_{s'_I+1}.(\text{fanOut}(u_{s'_I+1}) - 1) \rangle$ .
  - $\mathcal{C}' = \mathcal{C}_\emptyset = (\perp, \perp, \emptyset, \emptyset, \text{map}, r_\emptyset)$  with  $\mathcal{A}_\emptyset = (Q', N'_R, \delta', q'_0, F')$ .
  - $Q'_T = \{q'_0\}$  and  $Y' = \{\text{delete}\}$ .

- If **repair**  $\in Y$ ,  $s'_I < k$ ,  $m_{k+1} \neq \perp$  and  $\delta(q'_I, m_{s'_I+1})$  is defined, then:
  - Let  $\mathcal{F} = [a', r' \rightarrow n'] \in P_R$  such that  $n' = m_{s'_I+1}$  and  $a' = \text{lab}(u_{s'_I+1})$ .
  - $t' = \text{repair}$ ,  $p' = u_{s'_I+1}$ ,  $e' = \perp$  and  $v'_E = (s'_I + 1, \delta(q'_I, m_{s'_I+1}))$ .
  - $u' = \langle u_{s'_I+1}.0, \dots, u_{s'_I+1}.(\text{fanOut}(u_{s'_I+1}) - 1) \rangle$ .
  - $\mathcal{C}' = \mathcal{C}_{\mathcal{F}} = (a', n', N'_R, P'_R, \text{map}', r')$  with  $\mathcal{A}_{r'} = (Q', N'_R, \delta', q'_0, F')$ .
  - If  $r' \neq \emptyset$ , then  $Q'_T = F'$ , else  $Q'_T = \{q'_0\}$ .
  - If  $r' \neq \emptyset$ , then  $Y' = \Omega \setminus \{\text{correct}\}$ , else  $Y' = \{\text{delete}\}$ .
- If **rename**  $\in Y$ ,  $s'_I < k$  and  $[m_{s'_I+1} = \perp$  or  $\delta(q'_I, m_{s'_I+1})$  is not defined], then  $\forall x \in N_R$ : if  $\delta(q'_I, x)$  is defined, then we define  $\mathcal{T}'$ , where:
  - Let  $\mathcal{F} = [a', r' \rightarrow n'] \in P_R$  such that  $n' = x$  and  $\text{map}(a') = x$ .
  - $t' = \text{rename}$ ,  $p' = u_{s'_I+1}$  and  $e' = (\text{renameLabel}, a')$ .
  - $v'_E = (s'_I + 1, \delta(q'_I, x))$ ,  $u' = \langle u_{s'_I+1}.0, \dots, u_{s'_I+1}.(\text{fanOut}(u_{s'_I+1}) - 1) \rangle$ .
  - $\mathcal{C}' = \mathcal{C}_{\mathcal{F}} = (a', n', N'_R, P'_R, \text{map}', r')$  with  $\mathcal{A}_{r'} = (Q', N'_R, \delta', q'_0, F')$ .
  - If  $r' \neq \emptyset$ , then  $Q'_T = F'$ , else  $Q'_T = \{q'_0\}$ .
  - If  $r' \neq \emptyset$ , then  $Y' = \Omega \setminus \{\text{correct}\}$ , else  $Y' = \{\text{delete}\}$ .

Finally, we define  $\text{NestedIntents}(\mathcal{I})$  as a set of all nested correction intents invoked by  $\mathcal{I}$ , i.e. all  $\mathcal{T}'$  introduced in this definition and derived from  $\mathcal{I}$ .

*Example 10.* Suppose that within the starting intent  $\mathcal{I}_\bullet$  for data tree  $\mathcal{T}$  from Example 1 and grammar  $\mathcal{G}$  from Example 3 we have invoked a nested **repair** intent  $\mathcal{I}$  on base node  $\epsilon$ . Thus we need to process sequence  $u = \langle 0, 1 \rangle$  of nodes with labels  $\langle x, d \rangle$  in context  $\mathcal{C}_1$  from Example 5. Being at a position  $(0, 0)$ , i.e. at stratum 0 (before the first node from  $u$ ) and in the initial state  $q_0 = 0$  of  $\mathcal{A}_r$  for  $r = C.D^*$  in Example 2, we can derive these nested intents:

$$\begin{aligned} \mathcal{I}_1 &= (\text{insert}, \perp, (\text{addLeaf}, c), (0, 0), (0, 1), \langle \rangle, \mathcal{C}_3, Q_3, \{\text{insert}\}), \\ \mathcal{I}_2 &= (\text{rename}, 0, (\text{renameLabel}, c), (0, 0), (1, 1), \langle 0, 0 \rangle, \mathcal{C}_3, Q_3, \Omega \setminus \{\text{correct}\}), \\ \mathcal{I}_3 &= (\text{delete}, 0, (\text{removeLeaf}), (0, 0), (1, 0), \langle 0, 0 \rangle, \mathcal{C}_\emptyset, Q_\emptyset, \{\text{delete}\}), \end{aligned}$$

where  $Q_3$  is a set of accepting states for  $\mathcal{C}_3$  based on  $\mathcal{F}_3$  and  $Q_\emptyset$  contains only the initial state of  $\mathcal{A}_\emptyset$  for  $r = \emptyset$ .

The recursive nesting terminates, if the node sequence to be processed is empty and the context allows only an empty model.

### 3.3 Correction Multigraphs

Correction intents can be viewed as multigraphs with edges corresponding to nested intents and vertices to pairs of sequence positions and automaton states. The idea of these multigraphs is adopted and extended from [9].

**Definition 20 (Exploration Multigraph).** Assume that  $\mathcal{T}$  is a data tree,  $\mathcal{G}$  a single type tree grammar and  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$  a correction intent with  $u = \langle u_1, \dots, u_k \rangle$ ,  $k \in \mathbb{N}_0$ . We define an exploration multigraph for  $\mathcal{I}$  to be a directed multigraph  $E(\mathcal{I}) = (V, E)$ , where:

- $V = \{(s, q) \mid s \in \mathbb{N}_0, 0 \leq s \leq k, q \in Q\}$  is a set of exploration vertices.
- $E = \{(v_1, v_2, \mathcal{I}') \mid \exists \mathcal{I}' \in \text{NestedIntents}(\mathcal{I}), \mathcal{I}' = (t', p', e', v'_I, v'_E, u', \mathcal{C}', Q'_T, Y') \text{ and } v_1 = v'_I, v_2 = v'_E\}$  is a set of exploration edges.

Extending the exploration multigraph and especially its edges with already evaluated *intent repairs* of nested intents, we obtain a correction multigraph.

**Definition 21 (Correction Multigraph).** *Given an exploration multigraph  $E(\mathcal{I}) = (V, E)$  for correction intent  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$  with  $u$  of size  $k \in \mathbb{N}_0$  and finite automaton  $\mathcal{A}_r = (Q, N_R, \delta, q_0, F)$  for  $r$  from context  $\mathcal{C}$ , we define a correction multigraph to be a tuple  $C(\mathcal{I}) = (V', E', v_S, V_T)$ , where:*

- $V' = V$  is a set of correction vertices.
- $E' = \{(v_1, v_2, \mathcal{I}', \mathcal{R}_{\mathcal{I}'}, c) \mid (v_1, v_2, \mathcal{I}') \in E\}$  is a set of correction edges, where  $\mathcal{R}_{\mathcal{I}'}$  is an intent repair for  $\mathcal{I}'$  and  $c = \text{cost}(\mathcal{R}_{\mathcal{I}'})$  is a cost of  $\mathcal{R}_{\mathcal{I}'}$ .
- $v_S = (0, q_0)$  is a source vertex.
- $V_T = \{v_T \mid v_T = (k, q_T), q_T \in Q_T\}$  is a set of target vertices.

*Example 11.* Continuing with Example 10, we can represent all nested intents derived from  $\mathcal{I}$  by a correction multigraph  $C(\mathcal{I})$  with  $v_S = (0, 0)$  and  $V_T = \{(2, 1), (2, 2)\}$  in Figure 4. For simplicity, edges are described only by abbreviated intent types ( $I$  for insert,  $D$  for delete,  $R$  for repair and  $N$  for rename), supplemented by a repairing instruction parameter if relevant and, finally, the complete *cost* of assigned intent repair. Names of vertices are concatenations of a stratum number and an automaton state.

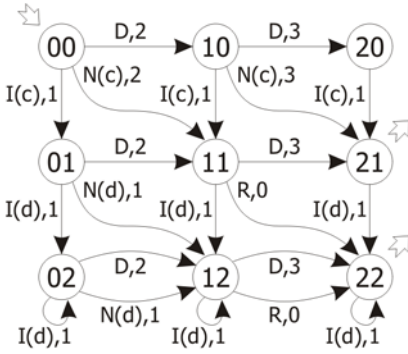


Fig. 4. Sample correction multigraph

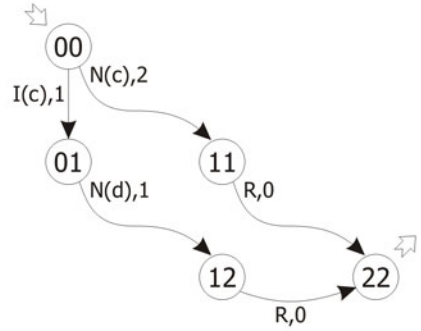


Fig. 5. Sample repairing multigraph

The problem of finding minimal repairs, i.e. the evaluation of correction intents, can now be easily converted to the problem of finding all shortest paths from the source vertex to any target vertex in correction multigraphs.

**Definition 22 (Correction Paths).** *Let  $C(\mathcal{I}) = (V, E, v_S, V_T)$  be a correction multigraph. Given  $x, y \in V$ , we define a correction path from  $x$  to  $y$  to be a sequence  $p_{x,y} = \langle e_1, \dots, e_n \rangle$  of correction edges, where  $n \in \mathbb{N}_0$  is a length and:*

- Let  $\forall k \in \mathbb{N}$ ,  $1 \leq k \leq n$ ,  $e_k = (v_1^k, v_2^k, \mathcal{I}^k, \mathcal{R}_{\mathcal{I}^k}^k, c^k)$ .
- If  $n > 0$ , then  $v_1^1 = x$  and  $v_2^n = y$ . Next,  $\forall k \in \mathbb{N}$ ,  $1 \leq k < n$ :  $v_2^k = v_1^{k+1}$ .
- $\neg \exists j, k \in \mathbb{N}$ ,  $1 \leq j < k \leq n$ :  $v_1^j = v_1^k$  or  $v_2^j = v_2^k$  or  $v_1^j = v_2^j$ .

If  $x = y$ , then  $p_{x,y} = \langle \rangle$ . Next,  $P_{x,y}$  is a set of all correction paths from  $x$  to  $y$ . Path cost for  $p_{x,y}$  is defined as  $\text{cost}(p_{x,y}) = \sum_{k=1}^n c^k$ . We say that  $p_{x,y}$  is the shortest path from  $x$  to  $y$ , if and only if  $\neg \exists p'_{x,y}$  such that  $\text{cost}(p'_{x,y}) < \text{cost}(p_{x,y})$ . By  $P_{x,y}^{\min}$  we denote a set of all shortest paths from  $x$  to  $y$ . Given nonempty  $Z \subseteq V$ , let  $m = \min_{z \in Z} \text{cost}(p_{x,z})$ . Then  $P_{x,Z}^{\min} = \{p \mid \exists z \in Z, p \in P_{x,z}^{\min} \text{ and } \text{cost}(p) = m\}$  is a set of all shortest paths from  $x$  to any  $z \in Z$ .

Finally, given a vertex  $v \in V$ , we say that  $v \in p_{x,y}$ , if  $\exists k \in \mathbb{N}$ ,  $1 \leq k \leq n$  such that  $v = v_1^k$  or  $v = v_2^k$ . Analogously, given an edge  $e \in E$ , we say that  $e \in p_{x,y}$ , if  $\exists k \in \mathbb{N}$ ,  $1 \leq k \leq n$  such that  $e = e_k$ .

Once we have found all required shortest paths, we can forget not involved parts of the correction multigraph. And moreover, these shortest paths themselves constitute the compact structure of the intent repair.

**Definition 23 (Repairing Multigraph).** Given a correction intent  $\mathcal{I}$  and its correction multigraph  $C(\mathcal{I}) = (V, E, v_S, V_T)$ , we define a repairing multigraph for  $\mathcal{I}$  to be a tuple  $R(\mathcal{I}) = (V', E', v_S, V_T, c)$  as a subgraph of  $C(\mathcal{I})$ , where:

- $V' = \{v \mid \exists p \in P_{v_S, V_T}^{\min}, v \in p\}$  and  $E' = \{e \mid \exists p \in P_{v_S, V_T}^{\min}, e \in p\}$ .
- $c = \text{cost}(p_{\min})$  for some (any)  $p_{\min} \in P_{v_S, V_T}^{\min}$ .

*Example 12.* A repairing multigraph  $R(\mathcal{I})$  for correction intent  $\mathcal{I}$  from Example 10 is derived from correction multigraph  $C(\mathcal{I})$  and is depicted in Figure 5.

**Definition 24 (Intent Repair).** Assume that  $R(\mathcal{I}) = (V, E, v_S, V_T, c)$  is a repairing multigraph for  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$ . We define an intent repair for  $\mathcal{I}$  to be a tuple  $\mathcal{R}_{\mathcal{I}} = (R_N, R_S, \text{cost})$ , where  $R_N = e$  is a repairing instruction,  $R_S = R(\mathcal{I})$  a repairing multigraph and  $\text{cost} = \text{cost}(e) + c$ .

At the bottom of the recursive intents nesting, the intent repair contains only one shortest path – a path on one vertex, without edges and with zero cost.

### 3.4 Repairs Translation

Assume that we have processed the entire data tree and thus we have computed all required nested intent repairs. The intent repair for the starting intent stands for all minimal corrections of the given XML document. Our goal is to prompt the user to choose the best suitable edit sequence, however, we first need to gain all these sequences from nested shortest paths, which involves also the translation of repairing instructions to edit operations along these paths.

**Definition 25 (Repairing Instructions Translation).** Given a repairing instruction  $e$ , we define a translation of  $e$  to the associated edit operation as  $\text{fix}(e)$ :

- If  $e = (\text{addLeaf}, a)$ , then  $\text{fix}(e) = \text{addLeaf}(0, a)$ .
- If  $e = (\text{removeLeaf})$ , then  $\text{fix}(e) = \text{removeLeaf}(0)$ .
- If  $e = (\text{renameLabel}, a)$ , then  $\text{fix}(e) = \text{renameLabel}(0, a)$ .

For the purpose of sequences translation, we need three auxiliary functions.

**Definition 26 (Auxiliary Translation Functions).** *Given a node  $u \in \mathbb{N}_0^*$  and a constant  $c \in \mathbb{N}_0$ , we define  $\text{modPre}(u, c) = c.u$ . If  $u \neq \epsilon$ ,  $u = i.v$ ,  $i \in \mathbb{N}_0$ ,  $v \in \mathbb{N}_0^*$ , then we define  $\text{modAlt}(i.v, c) = (i + c).v$  and  $\text{modCut}(i.v) = v$ .*

Once we have defined these functions on nodes, we can extend them on edit operations, edit sequences and, finally, sets of edit sequences. We just straightforwardly transform the node parameter of each particular edit operation, e.g.  $\text{modPre}(\text{addLeaf}(u, a), c) = \text{addLeaf}(\text{modPre}(u, c), a)$ .

**Definition 27 (Repairing Multigraph Translation).** *Let  $R(\mathcal{I}) = (V, E, v_S, V_T, c)$  be a repairing multigraph for  $\mathcal{I}$ . For each path  $p \in P_{v_S, V_T}^{\text{min}}$ ,  $p = \langle e_1, \dots, e_m \rangle$ ,  $m \in \mathbb{N}_0$ , we define  $S_p = \{s_p^1.s_p^2 \dots s_p^m \mid \forall i \in \mathbb{N}, 1 \leq i \leq m, s_p^i \in S_p^i\}$ , where all particular  $S_p^i$  are derived via the successive processing of edges from  $e_1$  to  $e_m$ . Thus let  $\forall i \in \mathbb{N}, 1 \leq i \leq m$ ,  $e_i = (v_1^i, v_2^i, \mathcal{I}^i, \mathcal{R}_{\mathcal{I}^i}, c^i)$ . Starting with  $a_0 = 0$  and  $i = 1$ , we put  $S_p^i = \text{modAlt}(\text{fix}(\mathcal{R}_{\mathcal{I}^i}), a_{i-1})$  and  $a_i = a_{i-1} + x_i$ , where:  $x_i = 1$  for  $t^i \in \{\text{insert}, \text{repair}, \text{rename}\}$  and  $x_i = 0$  for  $t^i = \text{delete}$ . Finally, we define a repairing multigraph translation  $\text{fix}(R(\mathcal{I})) = \bigcup_{p \in P_{v_S, V_T}^{\text{min}}} S_p$ .*

The intent repair translation idea is based on the traversal of all shortest paths stored in the repairing multigraph and the successive processing of their edges leading to the combination of already generated sequences from nested intents and the proper numbering of position parameters in edit operations.

*Example 13.* Suppose we have paths  $p_1$  and  $p_2$  from  $(0, 0)$  to  $(2, 2)$  via  $(0, 1)$  and  $(1, 1)$  respectively in  $R(\mathcal{I})$  from Example 12. For  $p_1$  we successively derive  $a_0 = 0$ ,  $S_p^1 = \{\langle \text{addLeaf}(0, c) \rangle\}$ ,  $a_1 = 1$ ,  $S_p^2 = \{\langle \text{renameLabel}(1, d) \rangle\}$ ,  $a_2 = 2$ ,  $S_p^3 = \{\langle \rangle\}$  and  $a_3 = 3$ . Analogously for  $p_2$ :  $a_0 = 0$ ,  $S_p^1 = \{\langle \text{renameLabel}(0, c), \text{removeLeaf}(0.0) \rangle\}$ ,  $a_1 = 1$ ,  $S_p^2 = \{\langle \rangle\}$  and  $a_2 = 2$ . Then  $S_{p_1} = \{\mathcal{X}_1\}$  and  $S_{p_2} = \{\mathcal{X}_2\}$  for  $\mathcal{X}_1$  and  $\mathcal{X}_2$  from Example 8. Finally,  $\text{fix}(R(\mathcal{I})) = \{\mathcal{X}_1, \mathcal{X}_2\}$ .

**Definition 28 (Intent Repair Translation).** *We define  $\text{fix}(\mathcal{R}_{\mathcal{I}})$  to be an intent repair translation for  $\mathcal{R}_{\mathcal{I}} = (R_N, R_S, \text{cost})$  of intent with type  $t$ , where:*

- If  $t = \text{correct}$ , then  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{\text{modCut}(r_S) \mid r_S \in \text{fix}(R_S)\}$ .
- If  $t = \text{insert}$ , then  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{(\text{fix}(R_N)).\text{modPre}(r_S, 0) \mid r_S \in \text{fix}(R_S)\}$ .
- If  $t = \text{delete}$ , then  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{\text{modPre}(r_S, 0).\langle \text{fix}(R_N) \rangle \mid r_S \in \text{fix}(R_S)\}$ .
- If  $t = \text{repair}$ , then  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{\text{modPre}(r_S, 0) \mid r_S \in \text{fix}(R_S)\}$ .
- If  $t = \text{rename}$ , then  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{(\text{fix}(R_N)).\text{modPre}(r_S, 0) \mid r_S \in \text{fix}(R_S)\}$ .

*Example 14.* The correction of a data tree in Figure 11 against a local tree grammar from Example 8 leads to  $\text{fix}(\mathcal{R}_{\mathcal{I}}) = \{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3\}$ , and thus all data trees in Figure 8 represent corrections with cost 2 using  $\mathcal{X}_1$ ,  $\mathcal{X}_2$  and  $\mathcal{X}_3$  respectively.



### 3.5 Correction Algorithms

Finally, we need to propose an algorithm for recursive intent repairs computation. A naive algorithm would first initiate the starting intent with the root node and at each level of nesting, it would construct the entire exploration multigraph, then evaluate its edges to acquire nested intent repairs to find shortest paths over them. However, such algorithm would be extremely inefficient.

---

**Algorithm 1.** cachingCorrectionRoutine
 

---

**Input** : Data tree  $\mathcal{T}$ , grammar  $\mathcal{G}$ , intent  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$ .

**Output**: Intent repair  $\mathcal{R}_{\mathcal{I}}$  for  $\mathcal{I}$ .

```

1  $\mathcal{R}_{\mathcal{I}} \leftarrow \text{getCachedRepair}(\mathcal{S}(\mathcal{I}))$ ; if  $\mathcal{R}_{\mathcal{I}} \neq \perp$  then return  $\mathcal{R}_{\mathcal{I}}$ ;
2 Let  $u \leftarrow \langle u_1, \dots, u_k \rangle$ ,  $\mathcal{C} \leftarrow (a, n, N_R, P_R, \text{map}, r)$  and  $\mathcal{A}_r \leftarrow (Q, N_R, \delta, q_0, F)$ ;
3  $C(\mathcal{I}) \leftarrow (V \leftarrow \{(0, q_0)\}, E \leftarrow \emptyset, v_S \leftarrow (0, q_0), V_T \leftarrow \{(k, q_T) \mid q_T \in Q_T\})$ ;
4  $pCost(v_S) \leftarrow 0$ ;  $pPrev(v_S) \leftarrow \emptyset$ ;  $reachedVertices \leftarrow \{v_S\}$ ;  $finalCost \leftarrow \perp$ ;
5 while  $reachedVertices \neq \emptyset$  do
6    $v \leftarrow \text{fetchMinimalVertex}(reachedVertices)$ ;
7   if  $v \in V_T$  and  $finalCost = \perp$  then  $finalCost \leftarrow pCost(v)$ ;
8   if  $finalCost \neq \perp$  and  $finalCost < pCost(v)$  then break;
9   foreach  $\mathcal{I}' = (t', p', e', v'_I = v, v'_E, u', \mathcal{C}', Q'_T, Y') \in \text{NestedIntents}(\mathcal{I})$  do
10     $\mathcal{R}'_{\mathcal{I}} = (R_N, R_S, cost) \leftarrow \text{cachingCorrectionRoutine}(\mathcal{T}, \mathcal{G}, \mathcal{I}')$ ;
11    if  $v'_E \notin V$  then Add correction vertex  $v'_E$  into  $V$  and  $reachedVertices$ ;
12    Add correction edge  $(v, v'_E, \mathcal{I}', \mathcal{R}'_{\mathcal{I}}, cost)$  into  $E$ ;
13     $c \leftarrow pCost(v) + cost$ ;
14    if  $pCost(v'_E) \neq \perp$  and  $pCost(v'_E) = c$  then  $p(v'_E) \leftarrow pPrev(v'_E) \cup \{v\}$ ;
15    else if  $pCost(v'_E) > c$  then  $pCost(v'_E) \leftarrow c$ ;  $pPrev(v'_E) \leftarrow \{v\}$ ;
16  $R(\mathcal{I}) \leftarrow \text{createRepairingGraph}(C(\mathcal{I}), finalCost, pPrev)$ ;
17  $\mathcal{R}_{\mathcal{I}} \leftarrow \text{createIntentRepair}(\mathcal{I}, R(\mathcal{I}))$ ;  $\text{setCachedRepair}(\mathcal{S}(\mathcal{I}), \mathcal{R}_{\mathcal{I}})$ ;
18 return  $\mathcal{R}_{\mathcal{I}}$ ;

```

---

**Definition 29 (Intent Signature).** Assume that  $\mathcal{T} = (D, \text{lab}, \text{val})$  is a data tree and  $\mathcal{I} = (t, p, e, v_I, v_E, u, \mathcal{C}, Q_T, Y)$  is a correction intent with grammar context  $\mathcal{C} = (a, n, N_R, P_R, \text{map}, r)$ . We define a signature  $\mathcal{S}(\mathcal{I})$  to be a tuple:

- If  $t = \text{correct}$ , then  $\mathcal{S}(\mathcal{I}) = (\text{correct})$ .
- If  $t = \text{insert}$ , then  $\mathcal{S}(\mathcal{I}) = (\text{insert}, n, a, \text{ContextChain}(\mathcal{I}))$ .
- If  $t = \text{delete}$ , then  $\mathcal{S}(\mathcal{I}) = (\text{delete}, p)$ .
- If  $t = \text{repair}$ , then  $\mathcal{S}(\mathcal{I}) = (\text{repair}, p, n)$ .
- If  $t = \text{rename}$ , then  $\mathcal{S}(\mathcal{I}) = (\text{rename}, p, n, a)$ .

First, we in fact do not need to construct and evaluate the entire correction multigraph, we can use the idea of Dijkstra algorithm and directly find shortest paths in a continuously constructed multigraph. Next, using the concept of intent

signatures, we can avoid repeated computations of the same repairs. Although two different intents are always different, the resulting intent repair may be the same, e.g. the deletion depends only on a subtree, but not on a particular context.

The algorithm first detects, whether we have already computed the repair with the same signature (line 1). If not, we initialise the correction multigraph (lines 2-3) and start (line 4) the traversal for finding all shortest paths to any of target vertices (lines 5-15). Finally, we compose the repair structure and store it in the cache under its signature (lines 16-18).

## 4 Conclusion

We have proposed and formally described a correction framework based on existing approaches and dealing with invalid nesting of elements in XML documents using the top-down recursive processing of potentially invalid data trees and the state space traversal of automata for recognising regular expression languages with the connection to regular tree grammars model of XML schemata.

Contrary to all existing approaches we have considered the class of single type tree grammars instead only local tree grammars. Under any circumstances we are able to find all minimal repairs using the efficient caching algorithm, which follows only the perspective ways of the correction and prevents repeated computations of the same correction intents. This efficiency is supported by performed experiments using the prototype implementation. A direct experimental comparison to other approaches cannot be presented, since these approaches result to different correction qualities and have different presumptions.

However, we do not support neither local transpositions, nor global moves of entire subtrees. In [7] we have considered wider set of edit operations and also corrections of attributes. The framework can also be extended to find not only minimal repairs and the algorithm can be improved to the parallel one.

## References

1. Bouchou, B., Cheriati, A., Ferrari Alves, M.H., Savary, A.: Integrating Correction into Incremental Validation. In: BDA (2006)
2. Allauzen, C., Mohri, M.: A Unified Construction of the Glushkov, Follow, and Antimirov Automata. In: Kráľovič, R., Urzyczyn, P. (eds.) MFCS 2006. LNCS, vol. 4162, pp. 110–121. Springer, Heidelberg (2006)
3. Corrector Prototype Implementation, <http://www.ksi.mff.cuni.cz/~svoboda/>
4. Thompson, H.S., Beech, D., Maloney, M., Mendelsohn, N.: XML Schema Part 1: Structures, 2nd edn. (2004), <http://www.w3.org/TR/xmlschema-1/>
5. Mlynkova, I., Toman, K., Pokorný, J.: Statistical Analysis of Real XML Data Collections. In: Proceedings of the 13th International Conference on Management of Data (2006)
6. Murata, M., Lee, D., Mani, M., Kawaguchi, K.: Taxonomy of XML Schema Languages using Formal Language Theory. ACM Trans. Internet Technol. 5(4), 660–704 (2005)

7. Svoboda, M.: Processing of Incorrect XML Data. Master's thesis, Department of Software Engineering, Charles University in Prague, Czech Republic, Malostranske namesti 25, 118 00 Praha 1, Czech Republic (July 2010)
8. Flesca, S., Furfaro, F., Greco, S., Zumpano, E.: Querying and Repairing Inconsistent XML Data. In: Ngu, A.H.H., Kitsuregawa, M., Neuhold, E.J., Chung, J.-Y., Sheng, Q.Z. (eds.) WISE 2005. LNCS, vol. 3806, pp. 175–188. Springer, Heidelberg (2005)
9. Staworko, S., Chomicky, J.: Validity-Sensitive Querying of XML Databases. In: Freund, Y., Györfi, L., Turán, G., Zeugmann, T. (eds.) ALT 2008. LNCS (LNAI), vol. 5254. Springer, Heidelberg (2008)
10. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F., Cowan, J.: Extensible Markup Language (XML) 1.1, 2nd edn. (2006), <http://www.w3.org/XML/>
11. Boobna, U., de Rougemont, M.: Correctors for XML data. In: Bellahsène, Z., Milo, T., Rys, M., Suci, D., Unland, R. (eds.) XSym 2004. LNCS, vol. 3186, pp. 97–111. Springer, Heidelberg (2004)
12. Tan, Z., Zhang, Z., Wang, W., Shi, B.-L.: Computing Repairs for Inconsistent XML Document Using Chase. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (eds.) APWeb/WAIM 2007. LNCS, vol. 4505, pp. 293–304. Springer, Heidelberg (2007)

# Identification of Scholarly Papers and Authors

Kensuke Baba, Masao Mori, and Eisuke Ito

Kyushu University, Japan

{baba@lib,mori@ir,itou@cc}.kyushu-u.ac.jp

**Abstract.** Repositories are being popular as places for publication of research outputs. To make more efficient use of scholarly information on the internet, repositories are required to cooperate with other databases. One of the essential processes of the cooperation is identification of scholarly papers and their authors. The straightforward approach is string matching of the title and authors' name, however this approach cannot always solve the difficulties by basic clerical errors and same names. This paper proposes a method to compensate for the inaccuracy of the identification by connecting different databases. The main idea of the method is that different metadata of a scholarly paper is linked by the authors themselves, therefore the correspondence is guaranteed by the authors. The authors of this paper are developing a system based on the idea on the repository and the researcher database in their university.

**Keywords:** Web database, repository, scholarly paper, identification.

## 1 Introduction

The number of digital contents on the internet is rapidly increasing. Especially, for scholarly information, electronic journals and repositories [10] are being popular as places for publication of research outputs. The metadata of a scholarly paper (that is, the information about the title, the author(s), and so on) is usually registered in plural databases severally, and therefore the metadata has some variations. For some papers, in addition to the metadata, the full-text is archived in plural databases and it has some versions such as the pre-/post-print. The scholarly papers should be organized to make more efficient use for the users of the information.

In order to organize scholarly papers, it is not practical the ideal solution that an authority should manage all the papers. A feasible solution is cooperation of databases and advanced search functions thereby. For the solution, we have to make clear the relation on scholarly papers. The first step is "identification of scholarly papers", that is, to link the variations of the metadata of each paper. The versions of each paper can be managed by processing this step in detail. As the second step, one of the simplest organizations is classification with respect to the authors. The classification requires "identification of authors". As the result of these identifications, the metadata should have IDs which correspond to real papers and authors, respectively.

The straightforward approach of the identification of scholarly papers and authors is string matching [8] by the title and authors' name. Some variations

of the title can be identified by approximate string matching [9]. As for authors, the accuracy can be improved by matching of extra information such as the affiliation. However, this approach cannot always solve the difficulties by basic clerical errors and same names. If we have enough data for the identification, machine learning and rule-based approach such as [4] are possible solutions. Another approach of a different quality is confirmation by the authors themselves. For example, the problems are solved by adding unified IDs for scholarly papers (such as DOI) and authors (such as the ID for membership of an association) to the metadata when the paper is registered. However, it is difficult to popularize unified IDs in advance, and moreover this solution cannot be applied to the papers which are already archived. The main idea of our solution is that the confirmation by the authors is realized by a cooperation of databases.

In this paper, we are trying to solve two problems in practical systems as a case study. Kyushu University has the researcher database DHJS (Kyushu University Academic Staff Educational and Research Activities Database, “Daigaku Hyoka Joho System” in Japanese) [1] and the repository QIR (Kyu(Q)shu University Institutional Repository) [2]. One of the problems we tackle is about identification of scholarly papers. DHJS has the metadata of scholarly papers which are produced by the researchers in the university. The number of the registered metadata is about 70,000, however it is estimated that at most about 20% is duplicate data. The other problem is about identification of authors. In QIR, a search of an author is operated by the naive string matching on the metadata, therefore the search cannot recognize any same name. The previous problems are solved by the following cooperation of the systems. By connecting the metadata in DHJS to the full-text in QIR,

- the first problem is solved since the identification of any paper is operated in QIR by handwork,
- the second problem is solved since a user authentication is required in DHJS for registration of metadata.

The number of the institutions who have own repository in the world is about 2,000 as of January 2011 [3], and most of the institutions are considered to have the same problem. In this paper, the situation of the practical systems in Kyushu University are shown in detail, and the problem and solution are described formally. Therefore, the proposed idea is applicable to other institutions.

## 2 Problem

This section describes the current situation of two databases, DHJS and QIR, and then formalize the problems we tackle.

### 2.1 DHJS

DHJS is the researcher database of Kyushu University. DHJS has various kinds of data about the researchers in the university, for example, the posts, their research interests, and the scholarly papers they produced. The number of the researchers in the university is about 3,000 as of October 2010, and any researcher has a duty

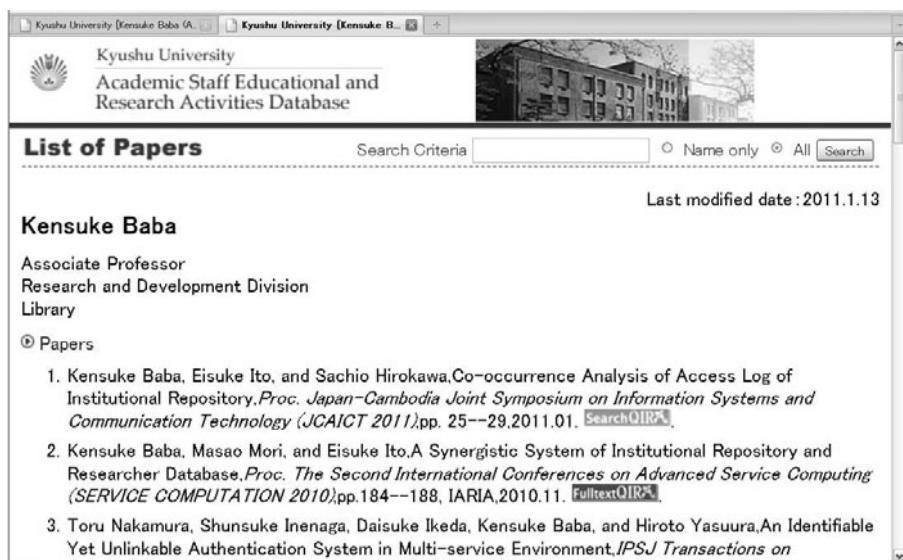
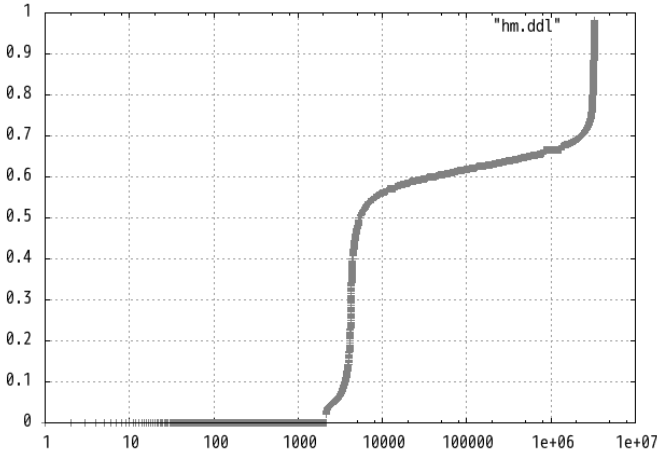


Fig. 1. The Web image of the list of scholarly papers in DHJS

to register their research activities includes the metadata of scholarly papers into DHJS. DHJS consists of the two subsystems, the data-entry system and the viewer system. The data-entry system supports researchers to register their research activities to DHJS and equips a user (that is, a researcher) identification by a password. The viewer system shows the research activities registered in DHJS by the data-entry system. Fig. 1 is an example of the list of the metadata of scholarly papers shown on DHJS. The icons in the figure are mentioned in the following section.

The number of the metadata of papers registered in DHJS is about 70,000 as of January 2011. If a paper was written by plural authors in Kyushu University, the metadata of the paper might be registered by each authors severally. We practically estimated the ratio of the duplicate data in DHJS by calculating the edit distance [11] between the titles of the papers. Fig. 2 is the result of the calculation for a department with about 15,000 pieces of metadata. The horizontal axis shows the number of pairs and the vertical axis the edit distance which is formalized by the length of the longer title. The number of the pairs whose edit distance is less than 0.1 is about 3,000, that is, the number of the duplicate data is at most about 3,000 (and 1,500 if we assume that 4 pieces of metadata are registered for a single paper on average [1]). Therefore, at most about 20% of the metadata are estimated to be duplicate. There was no significant

<sup>1</sup> The number of the duplicate data is defined to be the gap between the number of the metadata and the number of the distinct papers. If we assume that the duplicate data is made by  $n$  authors for each paper, then  ${}_n C_2$  pairs are counted for each paper and the number of the duplicate data is  $n - 1$  for each paper. Therefore, the number of the duplicate data is  $2m/n$  for the number  $m$  of the counted pairs.



**Fig. 2.** The edit distances between the all possible pairs of the 14,599 titles in DHJS for a department

difference of the ratio for every departments. By identification of the duplicate data, we can make more efficient use of the database, for example, we should be able to refer co-authors' site in DHJS from the metadata.

## 2.2 QIR

QIR is the institutional repository operated by Kyushu University Library. In general, institutional repository archives the full-text of each paper in addition to its metadata. The total number of the items (papers, slides, and so on) in QIR is about 16,000 as of January 2011. The registration of items to QIR are operated by staff in Kyushu University Library, and therefore the confliction of items are checked by handwork at the time. Fig. 3 is an example of the metadata of an item in QIR. The name of each author is linked to the profile page of the author, however the page is just the result of the naive string matching of the name for the items in QIR.

The problem of same name cannot be ignored. Actually, in 2,136 researchers of Kyushu University, there exist

- 10 pairs (20 persons) of the same given name and family name,
- 186 groups (488 persons) of the same initial of the given name and family name (for example, there exist 5 researchers of name “M. Tanaka”),
- 352 groups (1,255 persons) of the same family name (for example, there exist 22 researchers of name “Tanaka”).

Moreover, in addition to the researchers in Kyushu University, a lot of students and researchers in other institutions are included as co-authors of the papers in QIR.

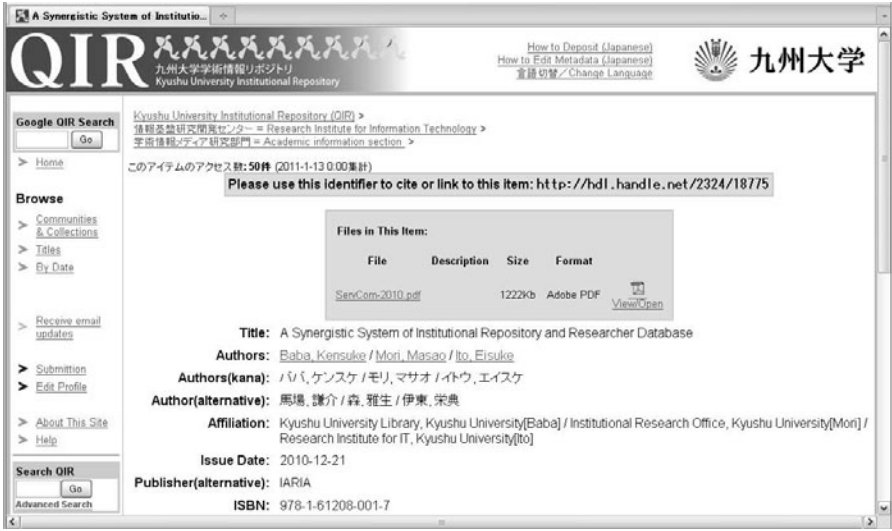


Fig. 3. The Web image of the metadata of an item in QIR

## 2.3 Formalization

The problems in the previous subsections are formalized. Since we are focusing on scholarly papers and their authors, we consider the set  $P$  of the real scholarly papers and the set  $A$  of the real authors. We define the *metadata* of a scholarly paper to be a pair of a string (the *title*) and a non-empty set of strings (the *author*). The problems are defined to be, for a given set  $M$  of metadata, to find the functions  $f : M \rightarrow P$  and  $g : M \rightarrow 2^A$  which represent the correspondence of the metadata to real papers and authors, respectively. In other words, it is to put indexes in  $P$  and  $A$  on the title and the author of each metadata.

In this paper, the input sets of metadata are  $M_D$  and  $M_Q$  for the metadata in DHJS and QIR, respectively. Then, the following assumptions are given for the problem by the current situations in the previous subsections. As mentioned in Subsection 2.1, the metadata in DHJS is registered by one of the authors and the registration requires a user authentication. Therefore, it can be regarded that we have the function  $g_1 : M_D \rightarrow A$  such that  $g_1(d) \in g(d)$  for any  $d \in M_D$ . As to Subsection 2.2, the correspondence between the metadata in QIR and the full-text is guaranteed by the check of the staff in the Library. Therefore, we have the function  $f_1 : M_Q \rightarrow P$  such that  $f_1(q) = f(q)$  for any  $q \in M_Q$ .

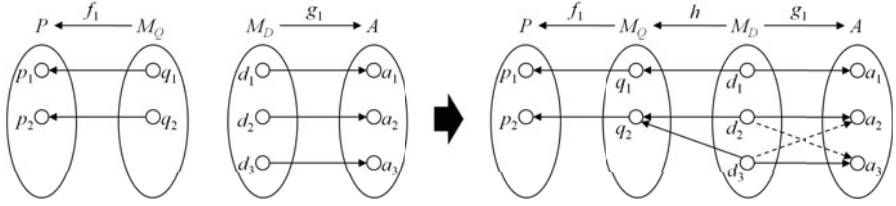
## 3 Solution

### 3.1 Main Idea

The main idea of our solution is, in terms of the formalization in Subsection 2.3, to find the function  $h : M_D \rightarrow M_Q$  which satisfies the following conditions:

1.  $f_1(h(d)) = f(d)$  for any  $d \in M_D$ ,





**Fig. 4.** The relation between  $f_1$ ,  $g_1$ , and  $h$  for an example  $(P, A, M_Q, M_D)$ . By  $h$ , we can compensate the relation of the dotted arrows

2.  $g_1(h^{-1}(q)) = g(q)$  for any  $q \in M_Q$ , where  $h^{-1}(q) = \{d \in M_D \mid h(d) = q\}$  and  $g_1(S) = \{g_1(s) \mid s \in S\}$  for a set  $S \subseteq M_D$ , and
3.  $g(q) = g(d)$  for any  $d \in h^{-1}(q)$ .

This situation is illustrated in Fig. 4. In this example,  $P = \{p_1, p_2\}$ ,  $A = \{a_1, a_2, a_3\}$ ,  $M_Q = \{q_1, q_2\}$ , and  $M_D = \{d_1, d_2, d_3\}$ , and then the functions  $f$  and  $g$  are as the following table.

$x$	$q_1$	$q_2$	$d_1$	$d_2$	$d_3$
$f(x)$	$p_1$	$p_2$	$p_1$	$p_2$	$p_2$
$g(x)$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_2, a_3\}$

As the assumptions, we have the functions  $f_1$  and  $g_1$  such that  $f_1(q_i) = p_i$  for  $i = 1, 2$  and  $g_1(d_i) = a_i$  for  $i = 1, 2, 3$  (the left-hand in Fig. 4). Then,  $h$  should be  $h(d_1) = q_1$ ,  $h(d_2) = q_2$ , and  $h(d_3) = q_2$ . For the conditions 1 and 2, by the  $h$  we have  $h \circ f_1$  and  $h^{-1} \circ g_1$  such that  $h \circ f_1(d_1) = p_1$ ,  $h \circ f_1(d_2) = p_2$ ,  $h \circ f_1(d_3) = p_2$ ,  $h^{-1} \circ g_1(q_1) = \{a_1\}$ , and  $h^{-1} \circ g_1(q_2) = \{a_2, a_3\}$ . Additionally, for the condition 3, we have  $h \circ h^{-1} \circ g_1$  such that  $h \circ h^{-1} \circ g_1(d_1) = \{a_1\}$ ,  $h \circ h^{-1} \circ g_1(d_2) = \{a_2, a_3\}$ , and  $h \circ h^{-1} \circ g_1(d_3) = \{a_2, a_3\}$ .

In the sense of the practical system, the condition 1 is clearly satisfied by linking the metadata of a paper in DHJS to the metadata of the paper in QIR. The condition 2 is satisfied by the previous link if any author of the papers in QIR register the metadata of the papers. It is also clear that the condition 3 is satisfied by the link since the author(s) of a paper in DHJS is same as the paper in QIR.

### 3.2 Implementation

We have already developed a system which links the metadata in DHJS to the full-text in QIR [6]. In Fig. 1, the dark-colored icon “fulltextQIR” is connected to the corresponding full-text in QIR. Researchers put icons on the list in the data-entry system of DHJS, and link them to the full-text by themselves. The other light-colored icon “SearchQIR” means that the metadata is not linked yet. Therefore, the correspondence between the metadata in DHJS and the metadata in QIR is guaranteed by a check of the author instead of string matching. Namely, the function  $h$  can be realized by this link system.

The following is the outline of the implementation. The ID of any paper in QIR is attached to the metadata in DHJS by this link, which realizes the condition 1. Additionally, since the metadata in DHJS has the ID of the author who registered the metadata, we can put the ID to the corresponding metadata in QIR by the link, which realizes the condition 2. By returning the author IDs from QIR to DHJS after the IDs for all the authors are attached, also the metadata in DHJS can have the IDs of the authors, which is for the condition 3.

One of the problems in the implementation is that the number of the metadata (full-text) in QIR is small compared with the number of metadata in DHJS. As mentioned in Section 2, the number of the metadata in QIR is 16,000 while the number in DHJS is at least about 56,000. To verify the effectiveness of our solution, the number of metadata in QIR is required to be large to complete the correspondence with the metadata in DHJS. For this problem, we are developing a system to encourage researchers to register their papers to QIR by showing the result of access log in QIR [5,7]. Another problem is about the dataflow between the databases. At the second phase in the outline, the ID of the author is sent from DHJS to QIR. However, the ID has to be returned from QIR to DHJS at the third phase. In general, such circulation of data is not suitable for managing databases. A precise policy should be defined for this dataflow.

## 4 Conclusion and Future Work

A method to compensate for the inaccuracy of identification of scholarly papers and authors on the metadata in separated databases was proposed. We formalized the problem in the practical systems and proposed the solution in terms of the formalization. Moreover, we showed the outline of the implementation based on the idea of the proposed solution.

One of our future work is the implementation of our solution along with the outline. Then, we are going to examine the accuracy of the identification, and observe the number of the access and the registered papers of QIR.

## References

1. Kyushu University Academic Staff Educational and Research Activities Database, [http://hyoka.ofc.kyushu-u.ac.jp/search/index\\_e.html](http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html) (accessed March 11, 2011)
2. QIR: Kyushu University Institutional Repository, <https://qir.kyushu-u.ac.jp/dspace/> (accessed March 11, 2011)
3. ROAR: Registry of Open Access Repositories, <http://roar.eprints.org/> (accessed March 11, 2011)
4. Afzal, M.T., Maurer, H., Balke, W.-T., Kulathuramaiyer, N.: Rule based autonomous citation mining with TIERL. *Journal of Digital Information Management* 8(3), 196–204 (2010)
5. Baba, K., Lto, E., Hirokawa, S.: Co-occurrence analysis of access log of institutional repository. In: *Proc. Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011)*, pp. 25–29 (2011)

6. Baba, K., Mori, M., Lto, E.: A synergistic system of institutional repository and researcher database. In: Proc. the Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010), IARIA, pp. 184–188 (2010)
7. Baba, K., Mori, M., Ito, E., Hirokawa, S.: A feedback system on institutional repository. In: Proc. the Third International Conference on Resource Intensive Applications and Services (INTENSIVE 2011), IARIA (2011)
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)
9. Crochemore, M., Rytter, W.: Text Algorithms. Oxford University Press, Oxford (1994)
10. Suber, P.: Open access overview. Open Access News (2007), <http://www.earlham.edu/~peters/fos/overview.htm> (accessed March 11, 2011)
11. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. Journal of the ACM 21(1), 168–173 (1974)

# A Social Network Model for Academic Collaboration

Sreedhar Bhukya

Department of Computer and Information Sciences  
University of Hyderabad, Hyderabad- 500046, India  
Email: sr2naik@gmail.com:

**Abstract.** A number of recent studies on social networks are based on a characteristic which includes assortative mixing, high clustering, short average path lengths, broad degree distributions and the existence of community structure. Here, a model has been developed in the domain of 'Academic collaboration' which satisfies all the above characteristics, based on some existing social network models. In addition, this model facilitates interaction between various communities (academic/research groups). This model gives very high clustering coefficient by retaining the asymptotically scale-free degree distribution. Here the community structure is raised from a mixture of random attachment and implicit preferential attachment. In addition to earlier works which only considered Neighbor of Initial Contact (NIC) as implicit preferential contact, we have considered Neighbor of Neighbor of Initial Contact (NNIC) also. This model supports the occurrence of a contact between two Initial contacts if the new vertex chooses more than one initial contacts. This ultimately will develop a complex social network rather than the one that was taken as basic reference.

**Keywords:** Social networks. Novel model. Random initial contact. Neighbor of neighbor initial contact. Tertiary contact. Academic collaboration.

## 1 Introduction

Now a days research in collaborations becoming domain independent. For example stock market analyst is taking the help of physics simulator for future predictions. Thus there is a necessity of collaboration between people in different domains (different communities, in the language of social networking.) Here we develop a novel model for collaborations in academic communities which gives a possibility of interacting with a person in a different community, yet retaining the community structure. Social networks are made of nodes that are tied by one or more specific types of relationships. The vertex represents individuals or organizations. Social networks have been intensively studied by Social scientists [3-5], for several decades in order to understand local phenomena such as local formation and their dynamics, as well as network wide process, like transmission of information, spreading disease, spreading rumor, sharing ideas etc. Various types of social networks, such as those related to professional collaboration [6-8], Internet dating [9], and opinion formation among people have been studied. Social networks involve Financial, Cultural,

Educational, Families, Relations and so on. Social networks create relationship between vertices; Social networks include Sociology, basic Mathematics and graph theory. The basic mathematics structure for a social network is a graph. The main social network properties includes hierarchical community structure [10], small world property [11], power law distribution of nodes degree [19] and the most basic is Barabasi Albert model of scale free networks [12]. The more online social network gains popularity, the more scientific community is attracted by the research opportunities that these new fields give. Most popular online social networks is Facebook, where user can add friends, send them messages, and update their personal profiles to notify friends about themselves. Essential characteristics for social networks are believed to include assortative mixing [13, 14], high clustering, short average path lengths, broad degree distributions [15, 16].

The existence of community structure, growing community can be roughly speaking set of vertices with dense internal connection, such that the inter community connection are relatively sparse [2].

Sousa et al. [20] developed a project for social networking system for educational professionals, this paper consider what kind of technologies could be used to create a web application that provided, type of interaction, behavior needed, requirements, technologies and the system implementation .

A report [21] also has been submitted which is examining three specific types of collaborative behavior and assessing their impacts on knowledge creation, Drawing on the toolkits of Social and Dynamic Network Analysis and a dataset of computer science department tenure and research track of faculty members of a major U.S. university.

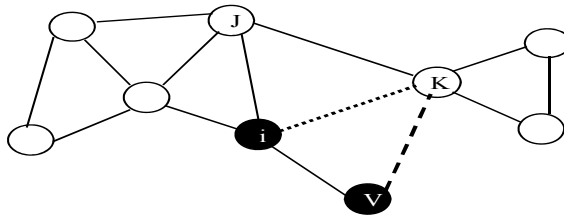
The advantage of our model can be understood from the following example. Let us consider a person contacting a person in a research group for his own purpose and suppose that he/she didn't get adequate support from that person or from his neighbors, but he may get required support from some friend of friend for his/her initial contact. Then the only way a new person could get help is that his primary contact has to be updated or create a contact with his friend of friend for supporting his new contact and introduce his new contact to his friend of friend. The same thing will happen in our day to day life also. If a person contacts us for some purpose and we are unable to help him, we will try to help him by some contacts of our friends. The extreme case of this nature is that we may try to contact our friend of friend for this purpose. We have implemented the same thing in our new model. In the old model [2], information about friends only used to be updated, where as in our model information about friend of friend also has been updated. Of course this model creates a complex social network model but, sharing of information or data will be very fast. This fulfills the actual purpose of social networking in an efficient way with a faster growth rate by keeping the community structure as it is.

## 2 Network Growth Algorithm

The algorithm includes three processes: (1) Random attachment (2) Implicit preferential contact with the neighbors of initial contact (3) In addition to the above

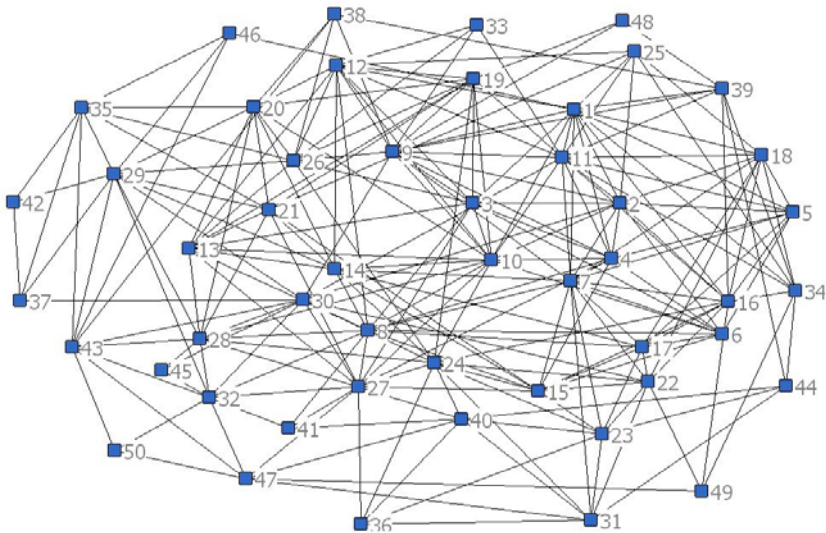
we are proposing a contact between the initial contact to its Neighbor of Neighbor contact (tertiary ). The algorithm of the model is as follows [1]:

- 1) Start with a seed network of N vertices
- 2) Pick on average  $m_r \geq 1$  random vertex as initial contacts
- 3) Pick on average  $m_s \geq 0$  neighbors of each initial contact as secondary contact
- 4) Pick on average  $m_t \geq 1$  neighbors of each secondary contact as tertiary contact
- 5) Connect the new vertex to the initial, secondary and tertiary contacts
- 6) Repeat the steps 2-5 until the network has grown to desired size.



**Fig. 1.** Growing process of community network, the new vertex ‘V’ initially connects through some one as initial contact (say *i*). Now *i*, updates its neighbor of neighbor contact list and hence connects to *k*. ‘V’ connects to  $m_s$  number of neighbors (say *k*) and  $m_t$  number of neighbor of neighbors of *i* (say *k*).

In this model we tried 50 sample vertices and prepared a growing network.



**Fig. 2.** showing social network graph with 50 vertices

### 3 Vertex Degree Distribution

We derive approximate value for the vertex degree distribution for growing network model mixing random initial contact, neighbor of neighbor initial contact and neighbor of initial contacts. Power law degree distribution with  $p(k) \sim k^{-\gamma}$  with exponent  $2 < \gamma < \infty$  have derived [17, 19]. In this model also the lower bound to the degree exponent  $\gamma$  is found to be 3, which is same as in the earlier model.

The rate equation which describes how the degree of a vertex changes on average during one time step of the network growth is constructed. The degree of vertex  $v_i$  grows in 3 processes:

- 1) When a new vertex directly links to  $v_i$  at any time  $t$ , there will be on average  $\sim t$  vertices. Here we are selecting  $m_r$  out of them with a probability  $m_r/t$ .
- 2) When a vertex links to  $v_i$  as secondary contact, the selection will give rise to preferential attachment. These will be  $m_r \cdot m_s$  in number.
- 3) When a vertex links to  $v_i$  as tertiary contact, this will also be a random preferential attachment. These will be  $2m_r \cdot m_s \cdot m_t$  in number.

These three processes lead to following rate equation for the degree of vertex  $v_i$  [1]

$$\frac{\partial k_i}{\partial t} = \frac{1}{t} \left( m_r + \frac{m_r m_s + 2m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)} k_i \right) \tag{1}$$

Based on the average initial degree of a vertex is

$$k_{init} = m_r + m_r m_s + 2m_r m_s m_t$$

Separating and integrating from  $t_i$  to  $t$ , and from  $k_{init}$  to  $k_i$ , we will get the following time evaluation for the vertex degrees

$$k_i(t) = B \left( \frac{1}{t_i} \right)^{1/A} - C \tag{2}$$

Where

$$A = 2 \left( \frac{m_r + m_r m_s + 2m_r m_s m_t}{m_r m_s + 2m_r m_s m_t} \right)$$

$$B=A \left( m_r + \frac{1}{2} m_r m_s + m_r m_s m_t \right)$$

$$C=Am_r$$

From time evolution of vertex  $k_i(t)$ , we can calculate the degrees of distribution  $p(k)$  by forming cumulative distribution  $F(k)$  and differentiating with respect to  $k$ . Since the mean field approximation[1] the degree  $k_i(t)$  of a vertex  $v_i$  increases monotonously from the time  $t_i$  the vertex initially added to the network, the fraction of vertices whose degree is less than  $k_i(t)$  at  $t$  is equivalent to the fraction of vertices that introduced after time  $t_i$ . Since  $t$  is evenly distributed, this fraction is  $(t-t_i)/2$ . These facts lead to the cumulative distribution [1]

$$F(k_i) = P(\tilde{k} \leq k_i) = P(\tilde{t} \geq t_i) = \frac{1}{t} (t-t_i) \tag{3}$$

Solving for  $t_i = t_i(k_i, t) = B^A (k_i + C)^{-A} t$  from (2) and inserting it into (3), differentiating  $F(k_i)$  with respect to  $k_i$ , and replacing the notation  $k_i$  by  $k$  in the equation, we get the probability density distribution for the degree  $k$  as

$$P(k) = AB^A (k+C)^{-2} / m_s + 2m_s m_t^{-3} \tag{4}$$

Here  $A$ ,  $B$  and  $C$  are as above. In the limit of large  $k$ , the distribution becomes a power law  $p(k) \sim k^{-\gamma}$  with  $\gamma = 3 + 2/m_s$ ,  $m_s > 0$ , leading to  $3 < \gamma < \infty$ . Hence the lower bound to the degree exponent is 3. Although the lower bound for degree exponent is same as earlier model. The probability density distribution is larger compared to earlier model, where the denominator of the first term of degree exponent is larger compared to the earlier model.

### 4 Clustering

The clustering coefficient on vertex degree can also be found by the rate equation method [18]. Let us examine how the number of triangles  $E_i$  changes with time. The triangle around  $v_i$  are mainly generated by three processes

1. Vertex  $v_i$  is chosen as one of the initial contact with probability  $m_r/t$  and new vertex links to some of its neighbors as secondary contact, giving raise to a triangle.
2. The vertex  $v_i$  is chosen as secondary contact and the new vertex links to it as its primary or tertiary contact giving raise to a triangles.
3. The vertex  $v_i$  is chosen as tertiary contact and the new vertex links to it as its primary or secondary contact, giving raise to a triangles.

These three process are described by the rate equation [1]



$$\frac{\partial E_i}{\partial t} = \frac{k_i}{t} - \frac{1}{t} \left( m_r - m_r m_s - 3m_r m_s m_t - \frac{5m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)t} k_i \right) \tag{5}$$

where second right-hand side obtained by applying Eq. (1) integrating both sides with respect to t, and using initial condition  $E_i(k_{init}, t_i) = m_r m_s(1+3m_t)$ , we get the time evaluation of triangle around a vertex  $v_i$  as

$$E_i(t) = (a + bk_i) \ln \left( \frac{t}{t_i} \right) + \left( \frac{a + bk_i}{b} \right) \ln \left( \frac{a + bk_i}{a + bk_{init}} \right) + E_{init} \tag{6}$$

Now making use of the previously found dependent of  $k_i$  on  $t_i$  for finding  $c_i(k)$ . solving for  $\ln(t/t_i)$  in terms of  $k_i$  from (2), inserting into it into (6) to get  $E_i(k_i)$ , and dividing  $E_i(k_i)$  by the maximum possible number of triangles,  $k_i(k_i-1)/2$ , we arrive the clustering the coefficient

$$c_i(k_i) = \frac{2 E_i(k_i)}{k_i(k_i - 1)} \tag{7}$$

where

$$\begin{aligned} E_i(k_i) &= Abk_i \ln(k_i + C) + k_i \ln \left( \frac{a + bk_i}{a + bk_{init}} \right) - bAk_i \ln B + aA \ln(k_i + C) + \\ &\frac{a}{b} \ln \left( \frac{a + bk_i}{a + bk_{init}} \right) + E_{init} - aA \ln B \\ &= Dk_i \ln(k_i + C) + k_i(F + Gk_i) - Dk_i \ln B + H \ln(k_i + C) + I \ln(F + Gk_i) + J \end{aligned}$$

Where

$$\begin{aligned} a &= m_r m_s + 3m_r m_s m_t - m_r, \quad b = \frac{5m_r m_s m_t}{2(m_r + m_r m_s + 2m_r m_s m_t)} \\ F &= \frac{a}{a + bk_{init}}, \quad D = Ab, \quad G = \frac{b}{a + bk_{init}}, \quad H = aA, \quad I = \frac{a}{b}, \quad J = E_{init} - H \ln B \end{aligned}$$

For large values of degree k, the clustering coefficient thus depend on k as  $c(k) \sim \ln k/k$ . This has very large clustering coefficient compared to the earlier work where it was  $c(k) \sim 1/k$ .

## 5 Results

In this model we tried 50 sample vertices and prepared a growing network, where edge to vertex ratio and triangle to vertex ratio for 50 nodes has been prepared. The results are given in Table: 1 .here one can see an enormous increase in secondary contacts. In addition tertiary contacts also have been added in our model, which leads to a faster and complex growth of network.

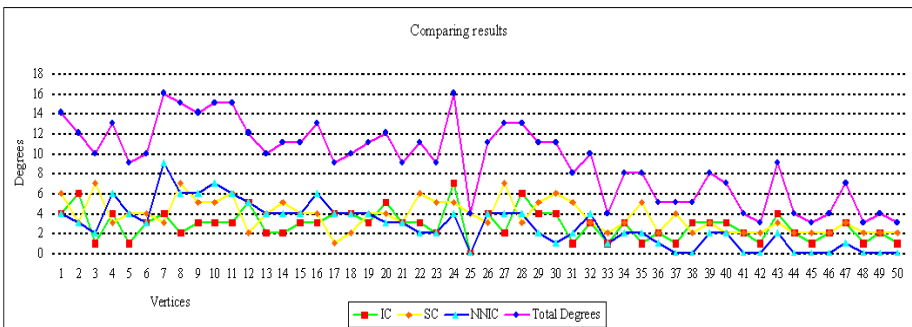
**Table 1.**

Data on our proposed model	Initial Contact (IC)	Secondary Contacts (SC)	Neighbor of Neighbor IC (NNIC)
Vertices	2.8	5.56	2.78
Triangles	0.8	6.0	6.44

### 5.1 Simulation Results

The below results have been represented graphically by calculating the degree (number of contacts) of a node. This also is shows an enormous growth in degree of nodes.

Simulation results



**Fig. 3.** Comparison results of growing network community: initial contacts are growing very slow rate compared to secondary contact i.e. ■ indicates initial contact, ◆ indicates secondary contacts, and ▲ indicates neighbor of neighbor of initial contact connects to the vertex  $v_i$ , Finally ● indicates degree of each vertices, when initial, secondary and tertiary contact connect to a vertex  $v_i$ . Our network community is growing very fast and complex when compared to existing model, vertices simulation results based on Table: 1.

## 6 Conclusion

In this paper, a model which reproduces very efficient networks compared to real social networks has been developed. And also here, the lower bound to the degree exponent is the same. The probability distribution for the degree  $k$  is in agreement with the earlier result for  $m_t = 0$ . The clustering coefficient got an enormous raise in growth rate of  $\ln(k_i)/k_i$  compared to the earlier result  $1/k_i$  for large values of the degree  $k$ . This is very useful in the case of academic groups, which helps in faster information flow and an enormous growth in research. Thus here an efficient but complex model of social network has been developed which gives an enormous growth in probability distribution and clustering coefficient and edge to vertex ratio by retaining the community structure. This model can be used to develop a new kind of social networking among various research groups.

### Tool

We have used C language, Ucinet, NetDraw and Excel for creating graph and simulation.

### Notations

Notation	Description
$m_t$	Initial Contact
$m_s$	Secondary Contact
$k_i$	Degree of vertex $i$
$E_i$	Number of triangles at vertex $i$
$P(k)$	Probability density distribution of degree $k$

### References

1. Bhukya, S.: A novel model for social networks. In: BCFIC, February 16-18, pp. 21–24. IEEE, Los Alamitos (2011)
2. Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., OKaski, K.: A model for social networks. *Physica A* 371, 851–860 (2006)
3. Milgram, S.: *Psychology Today* 2, 60–67 (1967)
4. Granovetter, M.: The Strength of Weak Ties. *Am. J. Soc.* 78, 1360–1380 (1973)
5. Wasserman, S., Faust, K.: *Social Network Analysis*. Cambridge University Press, Cambridge (1994)
6. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small -world’ networks. *Nature* 393, 440 (1998)
7. Newman, M.: The structure of scientific collaboration networks. *PNAS* 98, 404–409 (2001)
8. Newman, M.: Coauthorship networks and patterns of scientific collaboration. *PNAS* 101, 5200–5205 (2004)
9. Holme, P., Edling, C.R., Liljeros, F.: Structure and Time-Evolution of an Internet Dating Community. *Soc. Networks* 26, 155–174 (2004)

10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 7821–7826 (2002)
11. Newman, M.E.J.: The structure and function of complex networks. *SIAM Review* 45, 167–256 (2003)
12. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* 286, 509–512 (1999)
13. Newman, M.E.J.: Assortative Mixing in Networks. *Phys. Rev. Lett.* 89, 208–701 (2002)
14. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. *Phys. Rev. E* 68, 036–122 (2003)
15. Amaral, L.A.N., Scala, A., Barth, M., Stanley, H.E.: Classes of small-world networks. *PNAS* 97, 11149–11152 (2000)
16. Boguna, M., Pastor-Satorras, R., Díaz-Guilera, A., Arenas, A.: Models of social networks based on social distance attachment. *Phys. Rev. E* 70, 056122 (2004)
17. Evans, T., Saramäki, J.: Scale-free networks from self-organization. *Phys. Rev. E* 72, 026138 (2005)
18. Szabo, G., Alava, M., Kertesz, J.: Structural transitions in scale-free networks. *Phys. Rev. E* 67, 056102 (2003)
19. Krapivsky, P.L., Redner, S.: Organization of growing random networks. *Phys. Rev. E* 63, 066123 (2001)
20. Sousa, C., Martins, P., Fonseca, B., Paredes, H., Meehan, A., Devine, T.: Social networking system for academic collaboration. In: Luo, Y. (ed.) *CDVE 2008. LNCS*, vol. 5220, pp. 295–298. Springer, Heidelberg (2008)
21. Hill, V.A.: Collaboration in an Academic Setting: Does the Network Structure Matter? *NSF 07-576 (CASOS)*

# Performance Evaluation of Preference Evaluation Techniques

Alwan A. Ali, Ibrahim Hamidah, Tan Chik Yip,  
Sidi Fatimah, and Udzir Nur Izura

Department of Computer Science, Faculty of Computer Science and Information Technology,  
Universiti Putra Malaysia, 43400 Serdang, Selangor D. E., Malaysia  
ali83\_upm@yahoo.com, hamidah@fsktm.upm.edu.my,  
chikyiptan@gmail.com, {fatimacd, izura}@fsktm.upm.edu.my

**Abstract.** In recent years, there has been much focus on the design and development of database management systems that incorporate and provide more flexible query operators that return data items which are dominating other data items in all attributes (dimensions). This type of query operations is named preference queries as they prefer one data item over the other data item if and only if it is better in all dimensions and not worse in at least one dimension. Several preference evaluation techniques for preference queries have been proposed including top- $k$ , skyline, top- $k$  dominating,  $k$ -dominance, and  $k$ -frequency. All of these preference evaluation techniques aimed to find the “best” answer that meet the user preferences. This paper aims to evaluate these five preference evaluation techniques on real application when huge number of dimensions is the main concern. To achieve this, a recipe searching application with maximum number of 60 dimensions has been developed which assists users to identify the most desired recipes that meet their preferences. Two analyses have been performed, where execution time is the measurement used.

**Keywords:** Preference queries, preference evaluation techniques, query processing.

## 1 Introduction

In the recent years, there has been much interest to design and develop database management systems that incorporate and provide more flexible query operators that best meet the user preference and limit the result sets. The preference queries are used in many application domains, like multi-criteria decision making applications [5-7, 19], where many criteria are involved to select the most suitable answer to the user query. Decision support system helps to combine various interests to recommend a strategic decision. Other domains include e-commerce environments like trade off between the price, quality, and efficiency of the products to be assessed; personal preferences of users who request a web service such as hotel recommender [23], and restaurant finder [8, 21], and peer-to-peer network [15]. In this regards, there are many preference evaluation methods that have been proposed such as top- $k$  [25],

skyline [24],  $k$ -dominance [7], top- $k$  dominating [19], and  $k$ -frequency [6]. The ultimate goal of these preference evaluation methods is to reduce the search space and improve the quality of the given answer by providing the best possible relevant answer with respect to the given conditions (preferences).

In this paper we attempt to examine the most recent techniques of preference evaluation of query processing in the database systems, namely: top- $k$ , skyline, top- $k$  dominating,  $k$ -dominance, and  $k$ -frequency when huge number of dimensions is to be considered. The evaluation should be performed on real application. Thus, we have purposely developed a recipe searching application which offers a variety of recipes that best meet the ever-changing demands of user. We focus on a hybrid consumer as every user whenever attempts to find the most suitable recipe will consider several sources of information before deciding which recipe to be chosen.

The reasons for choosing the recipe domain to evaluate the performance of the preference evaluation techniques are mainly due to: (i) each recipe normally consists of several components like ingredients, course types, cuisine types, cooking method, occasions, diet and others while the requirements of the end-user are multi-dimensional and cannot be easily expressed on discrete scales. In this paper 60 dimensions have been identified. (ii) The main critical issue is a recipe component ratio which is defined by what is known, as the “best” recipe for user. To tackle this, the preference evaluation techniques that consider the ratio and rank results accordingly to the user requirements are the best techniques to be used and evaluated.

This paper is structured as follows. In Section 2, the previous works related to this work is presented. In Section 3, the recipe searching application is introduced. Performance analysis is presented and discussed in Section 4. Conclusions are presented in the final section, 5.

## 2 Related Works

Many types and variations of preference evaluation techniques of preference queries have been described in the database literature. These techniques include but not limited to: top- $k$ , skyline,  $k$ -dominant skyline, skyline frequency, top- $k$  dominating, ranked skyline, skycube, sort and limit skyline algorithm (SaLSa), SUBSKY, sort-filter-skyline (SFS), and linear elimination sort for skyline (LESS). Most of these preference evaluation techniques aim to improve the search performance by terminating the process of searching the data items as early as possible in obtaining the “best” answer that satisfies the conditions as indicated in the submitted query. In the following we present the most important types of preference evaluation techniques in preference queries.

**Top- $K$ :** Given a set of data items with  $d$  dimensions (attributes) and a monotonic preference ranking function  $F$ , top- $k$  technique retrieves a selected set of data items ( $k$ ) that dominates the data items according to the best scoring value based on  $F$ .

The basic concept of this technique is to give score (weight) to each data item in the database. Thus, in order to produce the scoring results a preference ranking function (monotone function) is involved to accumulate the values of dimensions for

each particular data item. Then depends on the final results of the preference ranking function, the  $k$ -data items with the best scores are considered the preferred data items [5, 14, 18, 20, 25].

Several algorithms have been proposed based on the top- $k$  preference evaluation technique such as Onion [26], PREFER [3], Mpro [16], Top- $k$  Monitoring Algorithm (TMA) [18], and Skyband Top- $k$  Monitoring Algorithm (SMA) [18] but these algorithms are being evaluated on small scale of dimensions within the range 2-7.

**Skyline:** The skyline preference evaluation method produces the set of data items in a way such that the set of data items  $S$  are the superior among the other data items in the dataset. In other words, a data item  $p$  is preferred over another data item  $q$  if and only if  $p$  is as good as  $q$  *strictly* in at least one possible dimension (attribute) and in all possible dimensions (attributes) at the same dataset [5-6, 13-14, 19, 21, 23-24].

Several algorithms have been proposed based on the skyline preference evaluation technique such as Block-Nested-Loop (BNL) and Divide-and-Conquer (DC) [24], Sort-Filter-Skyline (SFS) [12], Linear Elimination Sort for Skyline (LESS) [22], Nearest Neighbor (NN) [10], Branch-Bound-Skyline (BBS) [9], Bitmap and Index [17], SkyCube [28] and Sort and Limit Skyline algorithm (SaLSa) [11] but these algorithms are being evaluated on small scale of dimensions within the range 2-10.

**Top- $K$  Dominating:** Top- $k$  dominating technique retrieves the set of data items  $k$  which are dominating the largest number of data items in the dataset. That means data item  $p$  is preferred over another data item  $q$  if and only if the domination power of  $p$  is greater than the domination power of  $q$ . The value of domination power of data item  $p$  comes from the total number of data items in the dataset that are dominated by  $p$ . Top- $k$  dominating technique is a very significant tool for multi-criteria application such as decision making system and decision support system, since it identifies the most significant data items in an intuitive way [1, 14, 19].

**$K$ -Dominance:**  $K$ -dominance skyline technique prefers one data item  $p$  over another data item  $q$  in the dataset  $D$  if and only if  $p$  is as good as  $q$  *strictly* in at least one possible  $k$ -dimension (attribute) and in all possible  $k$ -dimensions (attributes) at the same dataset.

$K$ -dominance exhibits some characteristics over the traditional skyline. The size of  $k$ -dominance skyline answer is less than the size of conventional skyline answer, particularly when the considered dimensions are few. Furthermore,  $k$ -dominance has some similar characteristics with skyline technique especially when  $k = d$  ( $d$  is the total number of dimensions in the dataset). However,  $k$ -dominance skyline suffers from a significant problem which is *circular dominance* that leads to loss the transitivity property [7, 14, 27, 4].

**$K$ -Frequency Skyline:**  $K$ -frequency skyline technique retrieves a set of skyline data items  $D'$  from the given dataset  $D$  in a space  $S$ , where a data item  $p$  in  $D'$  has the lowest dominating score, denoted as  $S(p)$ , which represents the number of available sub-dimensions where  $p$  is not a skyline.

$K$ -frequency has many common characteristics with skyline technique such as transitivity property is preserved, and the  $k$ -frequency queries' answers are obtained by merely comparing the actual values for each identical dimension in two different data items. Further, this technique can be applied in the full space and subspace

dataset. However,  $k$ -frequency, need a powerful data structure that saves the dominated sub-dimensions for every data item  $p$  in order to precisely determine the score of every data item  $p$  [6, 14].

### 3 The Recipe Searching Application

The proposed recipe searching application has been successfully implemented using PHP web programming language and SQL server. Each preference technique has been developed and tested with respect to different type of recipes. We have identified six elements which are important in searching and later selecting a particular recipe. These elements are type of ingredients, courses, cooking methods, occasions, diet restrictions, and type of cuisines. Each element has its own set of dimensions (attributes) that can be selected. All together there are 60 dimensions. A range of 0-5 has been prepared for each dimension which indicates the degree of interest by a user towards a particular dimension. The smallest scale, 0, denotes no interest at all while the scale 5 denotes the highest preferences. Table 1 summarizes these dimensions. We use the notation  $d_i$  to denote the  $i$ th dimension.

**Table 1.** Dimensions of the Recipe Searching Application

Element	Number of dimensions
Main Ingredient	16 ( $d_1 - d_{16}$ )
Course	12 ( $d_{17} - d_{28}$ )
Cooking Method	8 ( $d_{29} - d_{36}$ )
Occasion	8 ( $d_{37} - d_{44}$ )
Diet	8 ( $d_{45} - d_{52}$ )
Cuisine	8 ( $d_{53} - d_{60}$ )

Figure 1 illustrates the main design interface of the proposed recipe searching application. The application provides several features for the user before a particular recipe is selected. These features include (i) users can select the preference evaluation technique they prefer; (ii) users are free to ignore any dimensions that are not interest to them. By default all dimensions are assigned the value 0; and (iii) users may rank the dimensions according to their needs by manipulating the scale to be assigned to the needed dimensions. For example, the following table represents a query submitted by a user.

**Table 2.** Example of dimensions selected in a User Query

Element	Dimensions selected
Main Ingredient	$d_1 = 5; d_2 = 3$
Course	$d_{18} = 4$
Cooking Method	$d_{29} = 4$
Occasion	$d_{43} = 5$
Diet	$d_{46} = 4$
Cuisine	$d_{54} = 5$

Note:  $d_1$  (chicken);  $d_2$  (rice);  $d_{18}$  (dinner);  $d_{29}$  (baking);  $d_{43}$  (Christmas);  $d_{46}$  (healthy);  $d_{54}$  (Italian)



After selecting the appropriate dimensions by giving the suitable preference value, then user is required to determine the type of preference technique before the application find the recipes. The default preference evaluation is the skyline. The best 20 first results will be shown based on the preference feature scoring and the preference evaluation technique that has been chosen. For the aim of this paper, 150 recipes have been collected and saved in a database called the Recipe Database (*RDb*).

Several steps are initially performed before the preference evaluation techniques are being applied. These steps mainly aim at removing the irrelevant data items (records) from the Recipe Database from being considered in the evaluation process as they will not contribute to the final result. The steps are discussed below:

1. Each recipe from the *RDb* is mapped into a two dimensional array, *RA*, with the following format:

Structure of *RA*

Index	0	1	2	3	...	60
Element	<i>Id</i>	<i>d1</i>	<i>d2</i>	<i>d3</i>	...	<i>d60</i>

Where *Id* is the identifier of the recipe and *di* is a score given to the *i*th dimension. We use the notation  $r_k.d_i$  to denote the *i*th dimension of the *k*th recipe. An example of a recipe stored in the array is as follow:

An instance of *RA*

Index	0	1	2	3	...	60
Element	101	5	0	2	...	5

The above is an information about the recipe 101 which uses chicken (*d1*) as the main ingredient, vegetable (*d3*), ..., and Southwestern (*d60*) is the main cuisine.

2. Given a query, *Q*, with a set of *n* selected dimensions,  $dq = \{dq1, dq2, \dots, dqn\}$  only those recipes in the *RA* that matched with these dimensions are selected and stored in a temporary array, *TRA*. The following definition defined the match criteria.

*Definition 1:* A recipe  $r_k$  is said to be matched to a query *Q* if  $\exists dq_i \in dq, \exists dj \in r_k$  and  $r_k.d_j > 0$  where *j* is the equivalent dimension as *i*.

This gives the following definition which defined the unmatched criteria.

*Definition 2:* A recipe  $r_k$  is said to be unmatched to a query *Q* if  $\forall dq_i \in dq, \exists dj \in r_k$  and  $r_k.d_j = 0$  where *j* is the equivalent dimension as *i*.

The following example clarifies this point. Consider the query given in Table 2 and the following instances of *RA*.

The screenshot shows a recipe search application interface. At the top, there is a navigation bar with links for 'home', 'myaccount', 'about', 'contact', and 'recipes search by keywords'. Below this is a 'USER PREFERENCE' section with various filters for ingredients, courses, cooking methods, occasions, diets, and cuisines. To the right, a 'TOP 20 RECIPE RESULT' table lists recipes with their IDs, names, URLs, and scores.

Recipe Id	Recipe Name	Url	Score
id1	Chicken Scampi	http://www.kitchendaily.com/recipe/chicken-scampi-78155	19
id2	Cheesy Chicken Pizza	http://www.kitchendaily.com/recipe/cheesy-chicken-pizza-77516	19
id22	Turkey-Sweet Potato Hash	http://www.kitchendaily.com/recipe/turkey-sweet-potato-hash-299	19
id39	Sichuan-Style Chicken with Peanuts	http://www.kitchendaily.com/recipe/sichuan-style-chicken-with-peanuts-74491	19
id56	Asian Peanut Chicken with Cucumber Salad	http://www.kitchendaily.com/recipe/asian-peanut-chicken-with-cucumber-salad-82551	19
id57	Japanese Meatballs	http://www.kitchendaily.com/recipe/japanese-meatballs-14219	19
id58	Grilled Chicken Caesar Salad	http://www.kitchendaily.com/recipe/grilled-chicken-caesar-salad-74222	19
id59	Chicken Tenders with Lemon Spinach Rice	http://www.kitchendaily.com/recipe/chicken-tenders-with-lemon-spinach-rice-96182	19
id60	Chicken Pasta & Vegetable Casserole	http://www.kitchendaily.com/recipe/chicken-pasta-and-vegetable-casserole-77232	19
id61	Mustard Tarragon Chicken Cutlets	http://www.kitchendaily.com/recipe/mustard-tarragon-chicken-cutlets-142453	19
id62	Chipotle Chicken Quesadilla with Avocado	http://www.kitchendaily.com/recipe/chipotle-chicken-quesadilla-with-avocado-142270	19
id63	Seasoned Chicken & Wild Rice	http://www.kitchendaily.com/recipe/seasoned-chicken-and-wild-rice-143275	19
id64	Seasoned Chicken & Rice (Less Sodium)	http://www.kitchendaily.com/recipe/seasoned-chicken-and-rice-less-sodium-143295	19
id65	Parmesan & Romano Cheese Chicken Primavera	http://www.kitchendaily.com/recipe/parmesan-and-romano-cheese-chicken-primavera-143396	19
id66	Citrus Curried Chicken & Wild Rice Salad	http://www.kitchendaily.com/recipe/citrus-curried-chicken-and-wild-rice-salad-143283	19
id67	Chicken Fried Rice with Snow Peas	http://www.kitchendaily.com/recipe/chicken-fried-rice-with-snow-peas-143376	19
id68	Greek-Style Chicken & Rice	http://www.kitchendaily.com/recipe/greek-style-chicken-and-rice-143287	19

Fig. 1. The main interface design of the recipe searching application

User query

Index	$d1$	$d2$	...	$d18$	...	$d29$	...	$d43$	...	$d46$	...	$d54$	...
Element	5	3	...	4	...	4	...	5	...	4	...	5	...

Note: The other dimensions have the value 0.

Instances of *RA*

Index	<i>Id</i>	<i>d1</i>	<i>d2</i>	...	<i>d18</i>	...	<i>d29</i>	...	<i>d43</i>	...	<i>d46</i>	...	<i>d54</i>	...
Element	102	5	5	...	5	...	5	...	5	...	5	...	5	...
	103	0	0	...	0	...	0	...	0	...	0	...	0	...
	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	110	0	5	...	0	...	0	...	0	...	0	...	0	...

Note: The other dimensions that are not listed in the table might have value 0, 1, 2, 3, 4, or 5.

From the above instances of *RA*, recipe *r.102* and *r.110* satisfied the Definition 1 and are selected while *r.103* is ignored as for all the dimensions requested by the user have the value = 0 (satisfied the Definition 2).

- Those dimensions in the temporary array, *TRA*, which are not considered in the query, *Q*, are then eliminated to reduce the size of dimensions to be considered. Based on the example given in Step 2 above, the following is the result of Step 3.

Instances of *TRA*

Index	<i>Id</i>	<i>d1</i>	<i>d2</i>	<i>d18</i>	<i>d29</i>	<i>d43</i>	<i>d46</i>	<i>d54</i>
Element	102	5	5	5	5	5	5	5
	...	...	...	...	...	...	...	...
	110	0	5	0	0	0	0	0

- The preference evaluation techniques are then applied towards the recipes that have been saved in the *TRA*. We will not give the detail algorithm for each of the preference evaluation technique as readers may refer to the references as provided in Section 2.

## 4 Performance Evaluation

We have conducted two analyses. The first analysis aims at analyzing the performance of the preference evaluation techniques with respect to the total number of dimensions that represents the user preference involving in query process. In this paper we vary the number of dimensions from 10 – 60 dimensions, while the size of the recipe database is fixed.

The second analysis aims at comparing the preference evaluation techniques with respect to the size of recipe database while the number of dimensions is fixed during the process of searching the best recipes that meet the user’s request. In this paper we focused exclusively on the number of dimensions and the size of databases as the most critical factors influence the process of finding preference answer.

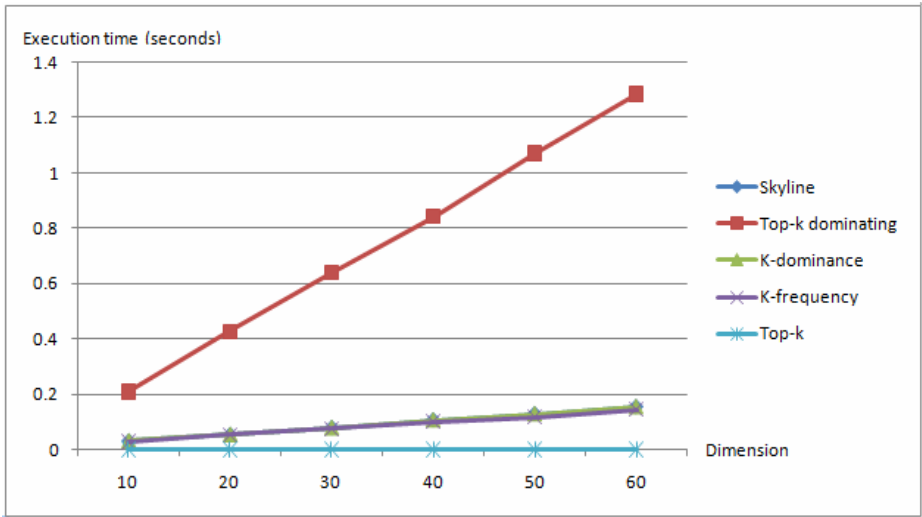
### 4.1 Results of Analysis 1

Figure 2 shows the results of applying different number of dimensions with fixed number of data items (recipe), which is 100. The initial number of dimensions is 10

and it is incrementally increased by 10, until the number of dimensions reached 60, which is the maximum number of dimensions considered in this paper. All together there are 6 experiments that have been conducted whereby in each experiment the number of dimensions considered is different. For each experiment 10 queries have been randomly generated where each query selects the appropriate number of dimensions (see Step 2 of Section 3). The execution time of each query is measured when Step 4 as described in Section 3 is performed. Averaging the execution time of these 10 queries gives the final execution time of the experiment. Thus, six different sets of queries have been designed for this analysis. The following table summarizes our experiment set up for this analysis.

**Table 3.** Experiments for the analysis 1

Experiment	Query	Number of dimensions	Number of Recipes
Experiment 1	$Q_1, Q_2, \dots, Q_{10}$	10	100
Experiment 2	$Q_{11}, Q_{12}, \dots, Q_{20}$	20	100
Experiment 3	$Q_{21}, Q_{22}, \dots, Q_{30}$	30	100
Experiment 4	$Q_{31}, Q_{32}, \dots, Q_{40}$	40	100
Experiment 5	$Q_{41}, Q_{42}, \dots, Q_{50}$	50	100
Experiment 6	$Q_{51}, Q_{52}, \dots, Q_{60}$	60	100



**Fig. 2.** The amount of execution time with respect to number of dimensions

From the above figure, the following can be concluded: in general the amount of execution time to retrieve the query answer increased for all the preference evaluation techniques when the number of dimensions increased. Top-k technique is the best as the increment rate of the execution time to obtain the query result is the lowest while skyline, k-dominance, k-frequency achieved almost the same execution time. However, top-k dominating technique performs the worst compared to the other

techniques as the execution time increased dramatically when the number of dimensions increased.

From this analysis, we can conclude that the number of dimensions involved in the process of preference queries has significant impact on the execution time in searching the “best” answer that meet the users’ preferences for most of the preference evaluation techniques. Moreover, this simple analysis shows that applying different type of preference evaluations give different impacts to the performance of the preference queries.

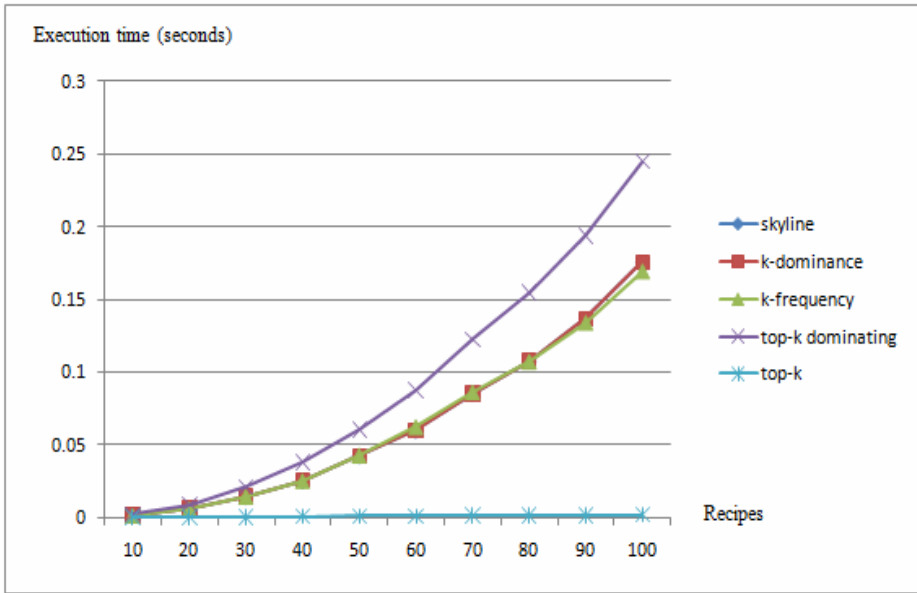
## 4.2 Results of Analysis 2

Figure 3 shows the results of applying different number of recipes which reflects the size of database with fixed number of dimensions, which is 10. The initial number of recipes is 10 and it is incrementally increased by 10, until the number of recipes reached 100, which is the maximum number of recipes considered in this analysis. All together there are 10 experiments that have been conducted whereby in each experiment the number of recipes considered is different. For each experiment 10 queries have been randomly generated where each query selects 10 dimensions (see Step 2 of Section 3). The execution time of each query is measured when Step 4 as described in Section 3 is performed. Averaging the execution time of these 10 queries gives the final execution time of the experiment. The following table summarizes our experiment set up for this analysis.

**Table 4.** Experiments for the analysis 2

Experiment	Query	Number of dimensions	Number of Recipes
Experiment 1	$Q_1, Q_2, \dots, Q_{10}$	10	10
Experiment 2	$Q_1, Q_2, \dots, Q_{10}$	10	20
Experiment 3	$Q_1, Q_2, \dots, Q_{10}$	10	30
Experiment 4	$Q_1, Q_2, \dots, Q_{10}$	10	40
Experiment 5	$Q_1, Q_2, \dots, Q_{10}$	10	50
Experiment 6	$Q_1, Q_2, \dots, Q_{10}$	10	60
Experiment 7	$Q_1, Q_2, \dots, Q_{10}$	10	70
Experiment 8	$Q_1, Q_2, \dots, Q_{10}$	10	80
Experiment 9	$Q_1, Q_2, \dots, Q_{10}$	10	90
Experiment 10	$Q_1, Q_2, \dots, Q_{10}$	10	100

From the above figure, it is obvious that the top- $k$  technique has the lowest amount of execution time compared to the other four techniques. This is due to the fact that most of the process in finding the best query answer is performed without needing to compare the individual dimensions at the data item level to determine the query results. i.e. accumulate the values of all dimensions as a single value. However,  $k$ -dominance,  $k$ -frequency and skyline techniques achieved almost the same amount of increment in the execution time when the number of recipes (the size of database) is increased. However, top- $k$  dominating has the worst performance compared to the other techniques.



**Fig. 3.** The amount of execution time with respect to number of recipes (database size)

## 5 Conclusion

In this paper we have presented and discussed a recipe searching application which has been developed with the aim to evaluate the various types of preference evaluation techniques for preference queries. Two analyses with different aims have been conducted by considering various numbers of dimensions and sizes of the databases. These are the most significant factors that impact on the execution time of the preference evaluation techniques in searching for the “best” query answer that meet the users’ preferences. We have also shown that the best preference technique in term of execution time is top- $k$ , while the worst is top- $k$  dominating.

## References

1. Jongwuk, L., Gae-won, Y., Seung-won, H.: Personalized top- $k$  Skyline Queries in High-Dimensional Space. *Information Systems* 34(1), 45–61 (2009)
2. Man, L.Y., Nikos, M.: Multi-Dimensional top- $k$  Dominating Queries. *The Very Large Data Bases Journal* 18(3), 695–718 (2009)
3. Vagelis, H., Yannis, P.: Algorithms and Applications for Answering Ranked Queries using Ranked Views. *The Very Large Data Bases Journal* 13(1), 49–70 (2004)
4. Zhenhua, H., Shengli, S., Wei, W.: Efficient Mining of Skyline Objects in Subspaces over Data Streams. *Knowledge and Information Systems* 22(2), 159–183 (2010)
5. Kontaki, M., Papadopoulos, A.N., Manolopoulos, Y.: Continuous Processing of Preference Queries in Data Streams. In: van Leeuwen, J., Muscholl, A., Peleg, D., Pokorný, J., Rumpe, B. (eds.) *SOFSEM 2010*. LNCS, vol. 5901, pp. 47–60. Springer, Heidelberg (2010)

6. Chee-Yong, C., Jagadish, H.V., Kian-Lee, T., Anthony, K.H., Zhenjie, Z.: On High Dimensional Skylines. In: 10th International Conference on Extending Database Technology, Munich, Germany, pp. 478–495 (2006)
7. Chee-Yong, C., Jagadish, H.V., Kian-Lee, T., Anthony, K.H., Zhenjie, Z.: Finding *k*-dominant Skylines in High Dimensional Space. In: ACM SIGMOD International Conference on Management of Data, Chicago, IL, USA, pp. 503–514 (2006)
8. Dana, A., Bouchra, S., Erick, L., Florence, S.: LA-GPS: A Location-aware Geographical Pervasive System. In: 24th International Conference on Data Engineering Works, Cancun, Mexico, pp. 160–163 (2008)
9. Dimitris, P., Yufei, T., Greg, F., Bernhard, S.: An Optimal and Progressive Algorithm for Skyline Queries. In: The International Conference on Management of Data, San Diego, California, USA, pp. 467–478 (2003)
10. Donald, K., Frank, R., Steffen, R.: Shooting Stars in the Sky: An Online Algorithm for Skyline Queries. In: 28th International Conference on Very Large Data Bases, Hong Kong, China, pp. 275–286 (2002)
11. Ilaria, B., Paolo, C., Marco, P.: SaLSa: Computing the Skyline without Scanning the Whole Sky. In: 15th International Conference on Information and Knowledge Management, Arlington, Virginia, USA, pp. 405–414 (2006)
12. Jan, C., Parke, G., Jarek, G., Dongming, L.: Skyline with Presorting. In: 19th International Conference on Data Engineering, Bangalore, India, p. 717 (2003)
13. Jian, P., Wen, J., Martin, E., Yufei, T.: Catching the Best Views of Skyline: A Semantic Approach Based on Decisive Subspaces. In: 31st International Conference on Very Large Data Bases, Trondheim, Norway, pp. 253–264 (2005)
14. Justin, J.L., Mohamed, F.M., Mohamed, E.K.: FlexPref: A Framework for Extensible Preference Evaluation in Database Systems. In: 26th International Conference on Data Engineering, Long Beach, California, USA, pp. 828–839 (2010)
15. Katerina, F., Evaggelia, P.: BITPEER: Continuous Subspace Skyline Computation with Distributed Bitmap Indexes. In: International Workshop on Data Management in Peer-to-Peer Systems, Nantes, France, pp. 35–42 (2008)
16. Kevin, C.C., Seung-won, H.: Minimal Probing: Supporting Expensive Predicates for Top-*k* Queries. In: International Conference on Management of Data, Madison, Wisconsin, pp. 346–357 (2002)
17. Kian-Lee, T., Pin-Kwang, E., Beng, C.O.: Efficient Progressive Skyline Computation. In: 27th International Conference on Very Large Data Bases, Roma, Italy, pp. 301–310 (2001)
18. Kyriakos, M., Spiridon, B., Dimitris, P.: Continuous Monitoring of Top-*k* Queries over Sliding Windows. In: International Conference on Management of Data, Chicago, Illinois, USA, pp. 635–646 (2006)
19. Man, L.Y., Nikos, M.: Efficient Processing of top-*k* Dominating Queries on Multi-Dimensional Data. In: 33rd International Conference on Very Large Data Bases, Vienna, Austria, pp. 483–494 (2007)
20. Martin, T., Gerhard, W., Ralf, S.: Top-*k* Query Evaluation with Probabilistic Guarantees. In: 30th International Conference on Very Large Data Bases, Toronto, Canada, pp. 648–659 (2004)
21. Mohamed, F.M., Justin, J.L.: Toward Context and Preference-aware Location-based Services. In: 8th International Workshop on Data Engineering for Wireless and Mobile Access, Providence, Rhode Island, pp. 25–32 (2009)
22. Parke, G., Ryan, S., Jarek, G.: Maximal Vector Computation in Large Data Sets. In: 31st International Conference on Very Large Data Bases, Trondheim, Norway, pp. 229–240 (2005)

23. Raymond, C.W., Ada, W.F., Jian, P., Yip, S.H., Tai, W., Yubao, L.: Efficient Skyline Querying With Variable User Preferences on Nominal Attributes. In: 34th International Conference on Very Large Data Bases, Auckland, New Zealand, pp. 1032–1043 (2008)
24. Stephan, B., Donald, K., Konrad, S.: The Skyline Operator. In: 17th International Conference on Data Engineering, Heidelberg, Germany, pp. 421–430 (2001)
25. Surajit, C., Luis, G.: Evaluating Top-k Selection Queries. In: 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, pp. 397–410 (1999)
26. Yuan-Chi, C., Lawrence, B., Vittorio, C., Chung-Sheng, L., Ming-Ling, L., John, R.S.: The Onion Technique: Indexing for Linear Optimization Queries. In: International Conference on Management of Data, Dallas, Texas, USA, pp. 391–402 (2000)
27. Yufei, T., Xiaokui, X., Jian, P.: SUBSKY: Efficient Computation of Skylines in Subspaces. In: 22nd International Conference on Data Engineering, Atlanta, Georgia, USA, pp. 65–74 (2006)
28. Zhenhua, H., Wei, W.: A Novel Incremental Maintenance Algorithm of SkyCube. In: 17th International Conference of Database and Expert Systems Applications, Kraków, Poland, pp. 781–790 (2006)



# Collective Action Theory Meets the Blogosphere: A New Methodology

Nitin Agarwal<sup>1,\*</sup>, Merlyna Lim<sup>2</sup>, and Rolf T. Wigand<sup>1</sup>

<sup>1</sup> Department of Information Science, University of Arkansas at Little Rock,  
2801 South University Avenue, Little Rock, AR 72204-1099, USA

<sup>2</sup> School of Social Transformation – Justice and Social Inquiry, Arizona State University,  
1711 South Rural Road, Tempe, AZ 85287-4902, USA

{nagarwal,rtwigand}@ualr.edu, Merlyna.Lim@asu.edu

**Abstract.** With the advent of advanced yet exoteric ICTs, especially the social media, new forms of collective actions have emerged to illuminate several fundamental yet theoretically obscure aspects of collective actions. Existing computational studies focusing on capturing and mapping the interactions and issues prevailing in social media manage to identify the manifestations of collective actions. They, however, lack modeling and predictive capabilities. In this paper, we propose a new methodology to gain deeper insights into cyber-collective actions by analyzing issue propagation, influential community members' roles, and transcending nature of collective actions through individual, community, and transnational perspectives. The efficacy of the proposed model is demonstrated by a case-study on Al-Huwaider's campaigns consisting of 150 blogs from 17 countries tracked between 2003 and 2010. To the best of our knowledge, the proposed methodology is the first to address the lacking fundamental research shedding light on re-framing Collective Action Theory in online environments.

**Keywords:** collective action, methodology, blogosphere, social computing, social network analysis, community, influence, transnational, issue crawler.

## 1 Introduction

Social media have played a major role in the formation of collective actions. They are often described as important tools for activists seeking to replace authoritarian regimes and to promote freedom and democracy, and they have been lauded for their democratizing potential. However, despite the prominence of “Iranian Twitter revolutions” [1] and the “Egyptian Facebook protests” [2], there is very little research on cyber-collective actions. Mere journalistic accounts on such actions are inevitably based on anecdotes rather than rigorously designed research.

The study of collective action has a long established history. Collective Action Theory was developed, however, in the pre-Internet era. As a result of emerging

---

\* Corresponding author.

information technologies, communication is not necessarily as costly, difficult, time consuming, or limited by the cognitive constraints of individuals as it once was. The availability of advanced systems of communication and information has prompted a reassessment of collective action theory, shedding light on the benefits and costs for successful contemporary collective action efforts. Simply put, new forms of collective actions reliant on information technology illuminate several fundamental aspects of collective actions that have remained theoretically obscure. This shows a clear need for new and innovative approaches as well as methods to re-frame Collective Action Theory in online environments.

On the other hand, computational studies on social media, that have increasingly become popular, such as mapping the blogosphere, tend to focus on capturing the connections between social media users. They predominantly do not study processes involved in collective acts in online environment. They also are mostly based on either link-analysis or content-analysis. They lack insights from social science, such as collective action theories, where issues (shared narratives/repertoires) are important in shaping collective action. This paper addresses the need for new methods by taking advantage of emerging new tools and combining link and content analysis as well as meme (shared issues) tracking. Moreover, the paper attempts to develop methods permitting explanatory and predictive powers that goes beyond the mere description of the studied phenomena as has been traditionally practiced so far.

With the emergence of cyber-collective action, there are several plausible research questions: Do social media reduce transaction costs for contentious political action? Do social media reduce/remove geographical barrier of collective-actions? Do social media speed up the shaping of collective action (reduce time)? Do social media create “flat” rather than hierarchical networks of collective action? Do social media more effectively disseminate and diffuse cause? Do social media change political opportunity? Do social media create different collective understandings of the distribution of societal opinion (i.e., change beliefs about what others believe)? Do social media uses enable participants of collective action to gain deeper understanding on the issue?

In seeking answers to these questions, in this paper we develop a novel methodology that combines insights from Collective Action Theory and advanced computational mapping methods. This new methodology will enable us not only to explain but also to predict the evolution of cyber-collective actions.

## **2 Related Work**

In this section, we present a brief review of collective action theory and an assessment of the various mapping efforts of the blogosphere. These reviews are necessary to gain an in-depth understanding on currently available methods and identify the need for a suitable methodology permitting the explanation and prediction of collective actions in the blogosphere.

## 2.1 Collective Action Theory

Theories of collective action are integral to explanation of human behavior. Perspectives on collective action have been useful in explaining diverse phenomena, including social movements be it in real world [3] or virtual worlds [37], membership in interest groups [4][5], the operation of the international alliance [6], establishment of electronic communities [7], formation of inter-organizational relationships [8], formation of standards-setting organizations [9][10], and even bidding behaviors [11]. This range of actions accounted by collective action perspectives illustrates the centrality of this body of theory to social science.

Collective action can be defined as all activity involving two or more individuals contributing to a collective effort on the basis of mutual interests and the possibility of benefits from coordinated action [12]. Traditional collective action theory dates back to 1937, when Ronald Coase sought to explain how some groups mobilize to address free market failures. Yet even when Mancur Olson began updating the theory in 1965 to explain “free-riding” the high-speed, low-cost communications now enjoyed were not imaginable [13]. New information and communication technologies (ICTs), especially the Internet, have completely transformed the landscape of collective action. In online environment, the burden of internal communication is no longer a hindrance to collective actions, so larger groups are no longer more successful than smaller ones (at least not by virtue of their size). E-mail, Web sites, chat rooms, blogs, and bulletin boards enable efficient communication, organization, and even deliberation within collective actions of any size [14][39].

However, some experts believe Internet-based collective action effects are overstated and may prove ephemeral. For example, [15] contends that easier international communication will not automatically translate into success for international collective actions because vital interpersonal networks cannot be adequately forged and maintained online. [16] agree that without face-to-face interaction, Internet communications cannot build the stable community a long-lasting movement requires. [17] argue that virtual demonstrations cannot satisfy the protester’s desire for the emotional rush and thrill of real, physical action. Yet there are many examples of successful Internet-based collective action such as the 1996 Zapatista rebellion in Mexico [18] or the 1998 Indonesian political revolution [19] [20]. The operation of groups such as these has recently been characterized as something beyond traditional collective action. These collective endeavors online have stirred debates about theories of collective action, raising questions of whether collective action, profoundly dependent on the Internet and other new technologies, is as effective or successful as collective action in more traditional modes [21] [22] [38]. Using both available successful and unsuccessful Internet-based collective actions, research has now begun identifying aspects of the collective action process that can succeed online as well as shortcomings and disadvantages of online collective action [23]. However, such research has not answered many other questions related to the emergence of various forms of collective actions in the online world. One of the major questions is to what extent the traditional collective action paradigm is even appropriate for explaining contemporary phenomena. As alluded above, this paper attempts to fill this gap by developing a methodology demonstrating how Collective

Action Theory can be reframed and applied, in combination of computational mapping, for explaining collective actions online.

## 2.2 Mapping the Blogosphere

Social media allows individuals to share their perspectives and opinions on various events in vast social network. The diffusion of opinions can lead to the formation of collective actions. One such form of collective actions, citizen journalism, has garnered interest from researchers and practitioners leading into many attempts to map issues in the blogosphere. However, a rigorous and fundamental analysis that explains cyber-collective action is not yet established. In this section, we assess some of these fundamental efforts to map the blogosphere that motivate the need for a more systematic and foundational analysis modeling collective action in the blogosphere or other forms of social media.

IssueCrawler [24] enables aggregating blogs and websites that mention issues of interest. Starting from a seed set of blog posts that contain the issue of interest, IssueCrawler crawls blogs and websites linked by at least two seed blog posts. IssueCrawler continues crawling the web resources that are three links away. For a good quality of crawl, it is imperative to start with a relevant seed set of blog posts, which are identified using Technorati search engine's results by using the issue as the search keyword.

Authors in [25] geocoded US blogs from LiveJournal and DiaryLand using city names and 3-digit ZIP codes specified by the bloggers in their postings. By identifying where people blog, local knowledge and culture can be gauged and certain behavior patterns could be identified, however, the geocoding mapping is not sufficient to explain why certain patterns exist and what they lead to. In essence, explanatory and predictive powers of such tools are missing. Authors in [40] mine blog content to identify local cuisine hotspots. In another study [26], authors mapped the US political blogosphere and observed the dichotomy between liberal and conservative blogs. Examining the link graph between and across these communities of blogs, authors observed certain interblog citation behavior patterns such as conservative bloggers tend to link more often than the liberal blogs, but there is no uniformity in the news or topics discussed by conservatives. However, the study fell short of suggesting a theory to explain these patterns.

Other studies such as [27] analyzed 60,000 blogs from the Iranian blogosphere using social network analysis and content analysis. A wide range of opinions representing religious conservative views, secular and reform-minded ones, and topics ranging from politics and human rights to poetry, religion, and pop culture were identified. In yet another study [28], authors analyzed the Arab Blogosphere consisting of 35,000 active blogs primarily from Egypt, Saudi Arabia, Kuwait, and other middle-east countries. The authors identified major clusters organized by countries, demographics, and discussion topics. The Arab Blogosphere primarily discussed domestic politics and religious issues with an occasional mention of US politics albeit in critical terms.

These and other similar studies clearly show that individuals discuss varied topics on various blogs, however, there is a lack of methodologies enabling the analysis of how the discussions converge to central themes. In lieu of addressing this gap, we

have specifically developed such a methodology (Section 3) enabling the use of Collective Action Theory and computational mapping in order to explain and predict the underlying processes involved in collective actions in the blogosphere.

### 3 Proposed Methodology

Scientific work typically aims at one (or both) of two things: (1) the precise, accurate and parsimonious description of some phenomena, and/or (2) the explanation of some phenomena, i.e. why does a phenomenon take place? These two questions need to be addressed carefully and well thought through in any piece of research. In this present area of research, it seems that considerable advancements may be made by taking the above two questions to heart. We argue that additional methodological rigor is needed to achieve the probably most important aim of theory, i.e. to explain and to predict. This is in addition to a theory being able to describe and to relate, prerequisites before explanation and prediction is possible. In general, a theory is designed to rationally and clearly explain a phenomenon. Moreover, a theory should be seen in light of the general nature of theory in that it should offer the following qualities and lend itself to be testable, falsifiable, generalizable, universal, and lasting over time.

In the following we will take a look at two broad methodological approaches permitting us to achieve the above stated aims of theory in relatively novel ways: (a) through relational or social network analytical approaches as well as (b) uniquely utilized tools enabling the capture and measurement of features within the blogosphere. Both approaches seem to be needed to achieve explanation and prediction in the context of Collective Action Theory.

The basic unit of analysis in social network analysis is a relationship between two system elements within the same system [29](p. 182). The term relationship deserves some specific attention: Generally, in social network analysis one is interested in dynamic, functional relationships, i. e. active interaction between the related elements. This kind of relationship, obviously, is of prime importance if one is to construct a network composed of relationships. Conceptually, the existence of a relationship between two elements is constituted by the recognition of some constraint, which restricts the behavior, at least minimally, of one or both of the elements [29](p. 182). Such a constraint suggests one other characteristic of a relationship, namely that of interdependence between the elements. Social scientists frequently have urged the need for relational analysis by emphasizing the importance to turn away from monadic and aggregate data [30][31][29]. The proponents of this approach to view 'reality' argue that the researcher not only manages to arrest data of two elements, A and B, as typically done in the monadic analysis, but that additional information is added to the recognition of constraints or, generally, a relationship between A and B. Four major properties of relational constraints can be identified: symmetry, strength, specificity and transitivity [29].

With regard to social network analytic purposes, a system is viewed as a set of elements embedded in a network of relationships. So far, the units of analysis, i.e. relationships, have been described and specified. Next, we offer a novel methodological approach how we may go beyond the mere mapping efforts typically done in issue tracking in the blogosphere. Our aim is to strive toward the

above-mentioned aims and features of theory, i.e. to develop predictive models, by combining social network analysis methods as well as focusing methodologically on information flows, issues and communities that, in turn, provide a deeper understanding of Collective Action Theory. In part this is accomplished via a mini-case study of the Weheja Al-Huwaider Campaign (in Section 4), illustrating the utility and strength of our novel methodological approach overcoming the previous limitations when looking at Collective Action Theory applied to research on the blogosphere and other social media or virtual world at large.

We have delved into emerging behavior patterns and their development into cyber collective movements from individual, community, and transnational perspectives, and in so doing delineate the challenges, propose an appropriate and fitting research methodology, evaluate various strategies, and analyze our findings. In order to cogently address the research questions posed in Section 1, we pursued a three-phased approach: phase 1, Individual Perspective; phase 2, Community Perspective; and phase 3, Transnational Perspective (see Fig. 1).

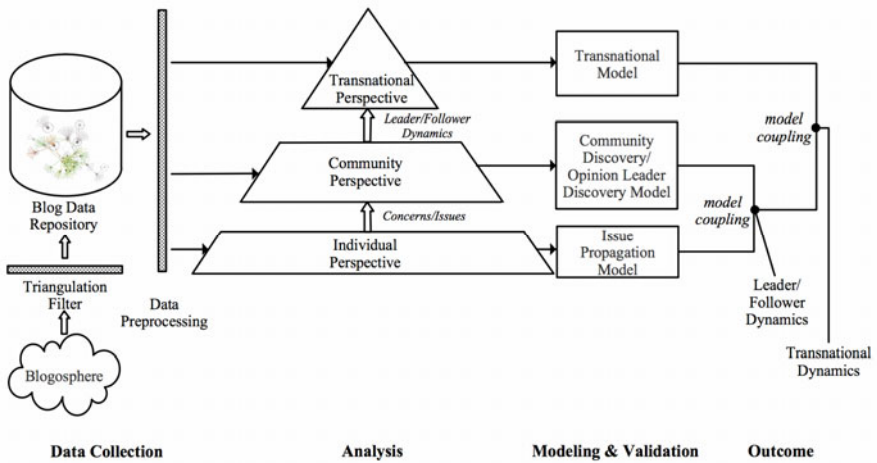


Fig. 1. Overall architecture of the research methodology.

### 3.1 Individual Perspective

It is observed that individual causes/issues can be transformed into collective cause. To understand and model this phenomenon, we need to study how personal issues and concerns evolve and propagate in social networks and how they converge and form collective concerns. We begin with preprocessing the blogs, identifying issues and concerns representing individual cause; and then modeling their diffusion in the network, and analyzing their convergence to collective cause. These steps are explained in further details next.

**Preprocessing and extracting cause:** For each new event that occurs, pre-event, during-event, and post-event blog reactions are analyzed. Probabilistic and statistical content analysis techniques such as Latent Semantic Analysis, Probabilistic Latent

Semantic Analysis, and Latent Dirichlet Allocation assist in identifying, segregating, and teasing out relevant topics. Blog posts containing relevant topics are summarized to reduce off-topic chatter narrowing in on the key information [32]. The summarized text is used to extract representative keywords using tag cloud generator (Wordle.net). Starting from the seed blog, the above process is repeated for all other blogs that are connected to the seed blog. Blogs connected to the *seed blogs* are termed *adjacent blogs*. This demonstrates whether the issues and concerns mentioned in the seed blog were diffused to the adjacent blogs.

**Modeling the diffusion of cause:** We analyze the extracted issues and concerns representing a certain cause and study their propagation. Specifically, we explore how network ties affect an individual's concerns. The proposed diffusion model extends the existing information diffusion models (linear threshold, independent cascade, etc.) by considering concerns as the information chunks that propagate over the social network of bloggers. Since the underlying social network remains the same, the structural properties of the concern diffusion are no different than information diffusion characteristics. In other words, leaders of the community who are responsible for the fastest information diffusion also tend to be the major influencing factors on the individual's issues and concerns and hence the collective concerns of the community.

The finding from the individual perspective leads us to think about possible trajectories for future research. Beyond extracting issues and concerns, a similar approach can be used to extract individual sentiment and track how it is diffused into collective sentiment. The exploration of existing sentiment analyzers in combination with the use of sentiment word thesaurus, sentiwordnet [42], will enable us to label the polarity and degree of the opinion word. For the future research agenda, we propose to longitudinally analyze the extracted issues, concerns and sentiment and to identify the factors involved in their propagation. We also propose to utilize existing cognitive and behavioral theories to gain deeper insights into the adaptation of individual behavior stemming from social interaction and cultural ties. These theories will form the basis of our exploration, aided by the development of novel statistical and stochastic diffusion models focusing on the transformation and propagation of sentiments along network ties over time. The model will help in advancing sociological as well as computational understanding of how collective sentiment shapes and will be improved upon in later phases of the analysis by incorporating community and transnational factors.

### 3.2 Community Perspective

Community leaders often exert significant influence over fellow members in transforming individual opinion and shaping into collective sentiment. To model this phenomenon, we analyze the community of bloggers, and identify the opinion leaders of the community. This enables us to address the following issues: how decisions travel across the network from leaders to followers?; do followers consistently follow the same leader(s) or is the influence of opinion leaders time-variant and/or topic-variant?; and is there a hierarchical structure in the rank of opinion leaders and can the

model identify it? Lastly, can we evaluate the model objectively? To address these questions, first, we extract and analyze the community of bloggers and then, identify the opinion leaders.

**Community identification:** Often in the blogosphere users do not explicitly specify their community affiliation. The discovery of communities through network-centric approaches has been extensively studied [33]; however, as pointed out in [34], blogs are extremely sparsely linked due to the casual environment that does not necessitate users to “cite” the sources that inspire them. Moreover, spam links generated by malicious users could connect completely unrelated and/or irrelevant blogs, affecting the performance of community discovery process. Further, spam may also adversely affect content-oriented community identification approaches [41]. We identify their implicit community affiliations and orientations leveraging the network structures (social ties, participation on other forms of social media) and issue/cause diffusion characteristics identified in the individual perspective phase.

Specifically, we explore both network and content-induced interactions between blogs to detect communities. The content-induced interactions approach, leveraging issues and concerns diffusion characteristics extracted from the individual perspective phase, not only guides the network-centric community extraction (while considering the relevant links and ignoring the spam/irrelevant links) but also complements it through revealing new potential links. Leveraging the insights from our prior study, the purpose of which is to identify communities from blog networks by examining the occurrence of shared concerns on particular events/causes, we unveil interactions through the observation of individual concerns. If the concerns of these blogs were similar we assume the blogs are themselves similar. Mathematically, the similarity between any two blogs can be computed using cosine similarity as follows,

$$Sim(B_m, B_n) = \frac{P_m \bullet P_n}{\|P_m\| \|P_n\|} \quad (1)$$

where,  $Sim(B_m, B_n)$  is the cosine similarity between blogs  $B_m$  and  $B_n$ . The concerns of  $B_m$  and  $B_n$  on an issue is represented by the column vectors  $P_m$  and  $P_n$ , respectively. The data mining clustering algorithm, *k-means*, is used to extract communities.

**Identifying Influentials:** After extracting the communities from blogs, we set out to identify community leaders. Given the sparse network structure of blogs, we leverage both network and content information to identify influentials. We examine how social gestures of “influentials” could be approximated by collectable statistics from the blogs. We gather network-based statistics from the blog graph (e.g., inlinks, outlinks, blogger social network, comments) and content-based statistics from blog text and comments to map the social gestures. Knowledge from prior work on identifying influential bloggers, iFinder [35], enables us to model community leaders factoring in socio-cultural traits of the community that bootstraps our understanding of opinion leaders. The model analyzes how issues and concerns travel across the blogger network from leaders to followers and identifies if there exists a hierarchical structure in the rank of opinion leaders. Due to lack of the ground-truth or benchmark datasets, an objective evaluation of the proposed model is extremely challenging. Here we propose an avant-garde evaluation framework that leverages social media sites such



as Digg ([www.digg.com](http://www.digg.com)) and blog search engines such as Technorati ([www.technorati.com](http://www.technorati.com)) as large-scale surveys to validate the model and identify opinion leaders. Details of this evaluation framework are given in [35].

The individual perspective phase provides an understanding of how issues and concerns propagate along the network. The outcome of the community perspective phase enlightens us with a deeper understanding of leader-followers dynamics. Together, outcomes from both phases lend insights into the emergence of cyber-collective movements in socio-culturally diverse environments. As a possible future direction, longitudinal analysis could be performed to address questions such as, whether followers consistently follow the same leader(s), or is the influence time-variant, offering deeper understanding of group dynamics.

### 3.3 Transnational Perspective

In this phase, we study and analyze whether collective concerns in communities transcend nation-state barriers and converge into transnational cyber-collective action or not. Analyzing the emergence of transnational actors and networks, structures relating to fluidity and boundless organizational architecture, is key to deeper understanding of the transnational underpinning of cyber-collective movements. Social networking platforms have undoubtedly intensified the degree of connectivity by building up capacity to circulate ideas and to transfer content very quickly across all barriers. An issue can be observed for a certain period of time and an issue-network can be constructed. The issue can be mapped periodically to detail the development of the issue-network. The mapping process can identify each blogger and classify her in one or more clusters (e.g., an Egyptian Canadian female blogger who resides in Arizona, United States belongs to three clusters: Egypt, Canada, and United States). The map of transnational collective movements then will show the overlap of various clusters and the expansion/evolution of networks.

This finding prompts us to seek answers for further questions such as the following: can transnational social movements be autonomous from national constraints in terms of discourses, strategies, and resources?; can the shifting scale (from local and national to global and transnational) also bring about a change of culture and identity of these movements?; with respect to outcomes and goals, can the transnational social movements deliver concrete strategies to overcome the unpredictability of their mobilizations?; with respect to their internal dynamics, can the transnational social movements encourage their perpetuation through mitigating the individual convictions of the collective actions/movements? Following we are presenting a brief case study illustrating how such research efforts can be accomplished methodologically.

## 4 Al-Huwaider Campaign Case Study

There are myriads of incidents and stories demonstrating the formation of collective causes and their manifestations in the form of cyber-collective movements. Among these stories we choose the Al-Huwaider campaign story that methodologically lends

itself to be captured through data and quite uniquely highlights how individual cause diffuses within the cyber-network of interactions and shapes into cyber-collective cause as time progresses.

The Al-Huwaider Campaign refers to the series of online campaigns for women's rights originally initiated by Saudi writer and journalist Waheja Al-Huwaider and later became a regional phenomenon [36]. Her YouTube campaign started in 2007. On International Women's Day 2007, Al-Huwaider drove a car in the Kingdom of Saudi Arabia (KSA), where it is forbidden for women to do so, while videotaping a plea to Saudi officials. She posted the video on YouTube that attracted international attention. Despite the obstacles placed by the Saudi government, Al-Huwaider continues to promote her ideas, through her writings online. Her articles analyze the Arab social situation, criticize the status of human rights, and vehemently protest discriminations and violence against women. Her online campaign has not only become an inspiration but also an influential voice for collective movements, calling for reform, among Middle Eastern women.

This story illustrates the potential of social media in facilitating cyber-collective actions. It shows how individual cause diffuses within the network, shapes into collective cause, and transforms into collective action. The overarching question anchored in this story is: How are decentralized online individual actions transformed into cyber-collective actions? Are the existing theories and methodologies capable of explaining cyber-collective action?

We present our analysis based on data from a real-world blog by collecting the blog posts of female Muslim bloggers from 17 countries. We handpicked a set of 150 blogs primarily written in English but also containing text in Arabic, Indonesian, and French. Bloggers were included based on three shared characteristics, also known as the 'triangulation' strategy, (1) explicit self-identification of gender and religious orientation – women over the age of 18, Muslim (verified through self-identification or Islamic references in their postings), and primarily blog in English; (2) evidence gathered from the blogger's friends and/or relatives; and (3) evidence gathered from the blogger's participation in other social media – we leveraged bloggers' registration on multiple blogs and multiple social media (such as MySpace, Twitter, Facebook, etc.) and cross-linking features.

#### **4.1 Individual Perspective**

We started with the original narrative of Wajeha Al-Huwaider's cause to lift the ban of driving for Saudi women as a source of issues and concerns. Representative keywords were then extracted using a tag cloud generator. We repeated the extraction for each blog within Al-Huwaider's network to seek whether Al-Huwaider's issues and concerns were diffused to these blogs. Our findings (Fig. 2) show the occurrence of similar keywords representing similar issues and concerns across these blogs (e.g., Saudi, women, cars, drive/driving, right/rights). This figure shows how an individual cause of Al-Huwaider was propagated in social networks and converged and, in turn, formed a collective cause.

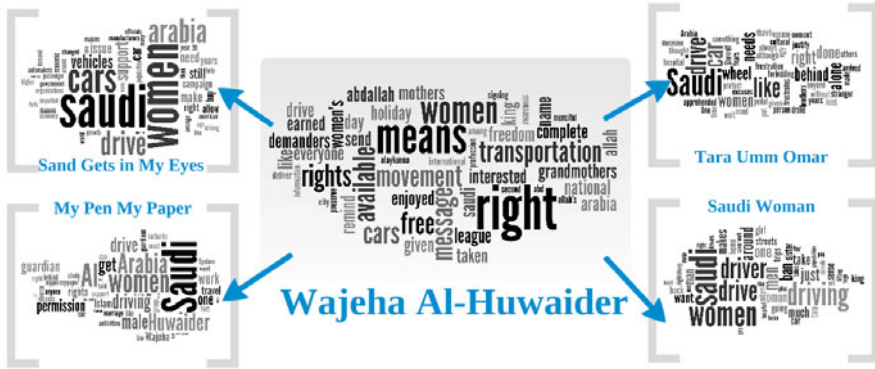


Fig. 2. Issue analysis of Al-Huwaider campaign.

### 4.2 Community Perspective

Al-Huwaider was a major factor in mobilizing individual bloggers with similar concerns (towards various issues) into a community and in leading the movement, i.e. transitioning individual cause to collective cause and ultimately manifesting into a cyber-collective movement. This also correlates with our findings in the individual phase, where the community leader was identified as the most significant influence over the individuals’ concerns. We followed the proposed methodology analyzing our data by extracting communities and opinion leaders and observing leader-follower dynamics.

Table 1. Occurrence of shared issues and concerns in each blog for different causes (overused words such as Saudi, Arabia and women are omitted)

Al-Huwaider’s Causes	Tara Umm Omar	Saudi Woman	Sand Gets in My Eyes
Women’s right to drive	drive, car, like, wheel, right, behind, alone,	driving, drive(r), want, around, make, men, ban,	cars, drive, vehicles, right, support, make, issue,
Black Ribbon Campaign to end the	guardianship, system, children, legal, denied,	black, ribbons, campaign, rights, women’s, November,	Al-Huwaider, actions, oppression, change,
Child Marriage	(-)	Marriages, change, allowing, child, justice, guardianship,	guardians, marry, father(s), ignorance, old,
...	...	...	...

Continuing with the example presented in Fig. 2, we identified the occurrence of various Al-Huwaider’s causes in three blogs, “Tara Umm Omar”, “Saudi Woman”, and “Sand Gets in My Eyes.” If the concerns of these blogs were similar we assume the blogs were themselves similar. We illustrate our analysis in Table 1, where we

aggregate the concerns from these three blogs (denoted in columns) for each cause/issue (denoted in rows).

Once communities of bloggers are extracted, our next step is to identify the influentials. We analyzed a community of 75 blogs that shared similar concerns for Al-Huwaider's campaigns and identified top ten influential blogs, as illustrated in Table 2. Due to space limitations we could not present the analysis of other blogs. However, all the 75 blogs had an average influence score of 198.306, a maximum influence score of 833, a minimum influence score of 1, and a standard deviation of 269.892. Representative tags extracted using Wordle are specified next to the blog posts to give contextual background.

**Table 2.** Top-10 influential blog posts discussing Wajeha Al-Huwaider's campaign along with their influence scores and representative tags extracted using Wordle.net

Blog	Representative Tags	Influence Score
<a href="http://hotair.com/archives/2009/07/12/saudi-feminist-blocked-from-leaving-country/">http://hotair.com/archives/2009/07/12/saudi-feminist-blocked-from-leaving-country/</a>	Saudi, Al-Huwaider, Arabia, border, male, passport, permission, activists, rights, guardian	833
<a href="http://jezebel.com/5552458/japan-likely-to-reject-ban-on-sexualization-of-minors-playboy-model-jailed-for-boob+grope">http://jezebel.com/5552458/japan-likely-to-reject-ban-on-sexualization-of-minors-playboy-model-jailed-for-boob+grope</a>	Women, minors, drinkers, Japan, Yousef, freedom, infected, prisoners, police, jail, charges, allegations	824
<a href="http://volokh.com/posts/1245159018.shtml">http://volokh.com/posts/1245159018.shtml</a>	Saudi, Arabia, HRW, Human, rights, links, mail, organization, government, Israel, workers	739
<a href="http://thelede.blogs.nytimes.com/2009/03/12/saudi-woman-drives-for-youtube-protest/">http://thelede.blogs.nytimes.com/2009/03/12/saudi-woman-drives-for-youtube-protest/</a>	Saudi, Huwaider, driving, BBC News, Arabia, Arab, women protest, video, Fattah, car, youtube	702
<a href="http://www.memeorandum.com/100418/p4">http://www.memeorandum.com/100418/p4</a>	Saudi, women, driving, Arabia, raped, reform, issues, populace	695
<a href="http://www.moonbattery.com/archives/2007/10/the_nobel_joke.html">http://www.moonbattery.com/archives/2007/10/the_nobel_joke.html</a>	Afghanistan, Navy, Murphy, bad, gore, Arafat, combat, killed, Marxist	690
<a href="http://latimesblogs.latimes.com/babylonbeyond/2010/06/saudi-women-use-fatwa-in-driving-bid.html">http://latimesblogs.latimes.com/babylonbeyond/2010/06/saudi-women-use-fatwa-in-driving-bid.html</a>	Women, Saudi, drive, Islamic, Wajeha, maternal, breastfeed, Obeikan, cars, ban, campaign	665
<a href="http://www.hrw.org/english/docs/2006/10/20/saudia14461.htm">http://www.hrw.org/english/docs/2006/10/20/saudia14461.htm</a>	Saudi, human, rights, police, detained, government, mabahith, Arabia, khobar, freedom	644
<a href="http://www.hrw.org/en/news/2006/10/30/saudi-arabia-lift-gag-order-rights-campaigner">http://www.hrw.org/en/news/2006/10/30/saudi-arabia-lift-gag-order-rights-campaigner</a>	Rights, al-Huwaider, Saudi, Arabia, human, september, mabahith, khobar, Abdullah, interrogated, police, officers,	644
<a href="http://globalvoicesonline.org/2008/08/12/saudi-arabia-bans-women-from-olympics/">http://globalvoicesonline.org/2008/08/12/saudi-arabia-bans-women-from-olympics/</a>	Feminist, Burundi, Olympics, Wajeha, Macha, Women, muharram	627

### 4.3 Transnational Perspective

Analyzing the emergence of transnational actors and networks, structures relating to fluidity, and boundless organizational architecture, is key to deeper understanding of transnational underpinning of cyber-collective movements. One such actor identified in our analysis was Wajeha Al-Huwaider. Despite the cultural, ethnic, political, social, and geographical diversity of Al-Huwaider's supporters as illustrated in Fig. 3

below, the sense of community superseded differences and converged individual concerns into collective action.

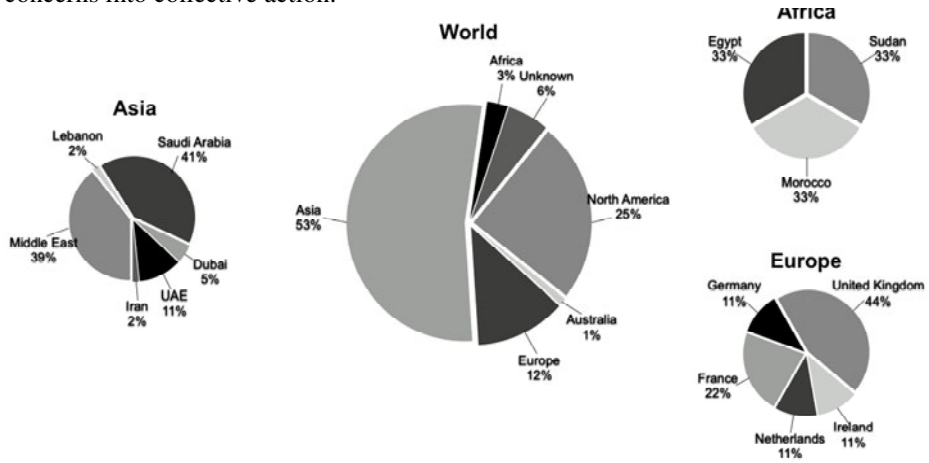


Fig. 3. Transnational support for Wajeha Al-Huwaider's campaign.

With access to more data, we can generate an issue network for Al-Huwaider's campaigns following our analysis in Fig. 3. Such issue networks can help decrypt the dimensions of: issues (on local, global, global-local levels), clusters (nation or content-based), political affiliations (conservative, liberal), time, and scale (network links, number of individuals, issue clusters) from actor and network perspectives.

This case study of the Wajeha Al-Huwaider campaign illustrates how our methodology enables the analysis of blogging behavior at the individual, community as well as transnational levels. In essence, our methodology illustrates how we can look at and explain collective actions in the blogosphere.

## 5 Conclusions

In this paper, we sought to understand the fundamentals, complexity, and dynamics of cyber-collective actions. By reaching out to existing social theories on collective action and computational social network analysis, we have proposed novel algorithms to model cyber-collective movements from individual, community, and transnational perspectives. The proposed methodology addresses the lacking fundamental research and re-framing Collective Action Theory in online environments making it the key contribution of this work. As utilized and illustrated in the mini-case study of the Al-Huwaider Campaign, our novel methodological approach convincingly overcomes the previous limitations when looking at Collective Action Theory applied to research on the blogosphere and other social media or virtual world at large. Further, our methodology goes beyond descriptive tendency of most computational studies on social media, such as mapping the blogosphere. By delving into the processes involved in collective acts in online environment and focusing on the formation of issues (shared narratives/repertoires), our approach offers the predictive power. We

also demonstrate that it is possible to develop predictive models of collective actions in blogosphere by combining social network analysis methods as well as focusing methodologically on information flows, issues and communities that, in turn, provide a deeper understanding of Collective Action Theory.

The findings in this paper also enable us to outline future research agenda that is geared towards the development of more advanced computational models. Such models would better our understanding of conventional social theories, assist in developing new ones, reinforcing the development of more accurate and efficient social interaction modeling algorithms for diverse environments allowing us to determine the trajectory of emerging cyber-collective movements.

**Acknowledgments.** This research was funded in part by the National Science Foundation's Digital Society and Technology Program (Award Number: IIS-0704978) and the US Office of Naval Research (Grant number: N000141010091). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

## References

1. Quirk, P.W.: Iran's Twitter Revolution. *Foreign Policy in Focus* (June 17, 2009), [http://www.fpif.org/articles/irans\\_twitter\\_revolution](http://www.fpif.org/articles/irans_twitter_revolution)
2. Masr, B.: Stop, Look, What's that Sound – The Death of Egyptian Activism (February 8, 2009), <http://bikyamasr.wordpress.com/2009/08/02/bm-opinion-stop-look-whats-that-sound-the-death-of-egyptian-activism/>
3. Tarrow, S.: *Power in Movement: Social Movements and Contentious Politics*. Cambridge University Press, Cambridge (1998)
4. Berry, J.: *The Interest Group Society*. Little Brown, Boston (1984)
5. Olson, M.: *The Logic of Collective Action*. Harvard University Press, Cambridge (1965)
6. Olson, M., Zeckhauser, R.: An economic theory of alliances. *Review of Economics and Statistics* 48, 266–279 (1966)
7. Rafaeli, S., Larose, R.: Electronic Bulletin Boards and 'Public Goods' Explanations of Collaborative Mass Media. *Communication Research* 20(2), 277–297 (1993)
8. Flanagan, A., Monge, P., Fulk, J.: The value of formative investment in organizational federations. *Human Communication Research* 27, 69–93 (2001)
9. Wigand, R., Steinfield, C., Markus, M.: IT Standards Choices and Industry Structure Outcomes: The Case of the United States Home Mortgage Industry. *Journal of Management Information Systems* 22(2), 165–191 (2005)
10. Markus, M., Steinfield, C., Wigand, R., Minton, G.: Industry-wide IS Standardization as Collective Action: The Case of the US Residential Mortgage Industry. *MIS Quarterly* 30, 439–465 (2006); Special Issue on Standard Making
11. Kollock, P.: The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. In: Smith, M., Kollock, P. (eds.) *Communities in Cyberspace*, pp. 220–239. Routledge, London (1999)
12. Marwell, G., Oliver, P.: *The Critical Mass in Collective Action*. Cambridge University Press, Cambridge (1993)

13. Lupia, A., Sin, G.: Which Public Goods Are Endangered? How Evolving Technologies Affect The Logic of Collective Action. *Public Choice* 117, 315–331 (2003)
14. Bimber, B., Flanagan, A., Stohl, C.: Reconceptualizing collective action in the contemporary media environment. *Communication Theory* 15(4), 365–388 (2005)
15. McAdam, D.: The framing function of movement tactics: strategic dramaturgy in the American civil rights movement. In: McAdam, D., McCarthy, J.D., Zald, M.N. (eds.) *Comparative Perspectives on Social Movements Opportunities, Mobilizing Structures, and Framing*, Cambridge, pp. 38–55. Cambridge University Press, Cambridge (1996)
16. Etzioni, A., Etzioni, O.: Face-to-face and computer-mediated communities, a comparative analysis. *The Information Society* 15, 241–248 (1999)
17. Van de Donk, W., Foederer, B.: E-movements or emotions? ICTs and social movements: Some preliminary results. In: Prins, J. (ed.) *Ambitions and limits on the crossroad of technological innovation and institutional change*, Boston (2001)
18. Bob, C.: *The Marketing of Rebellion*. Cambridge University Press, Cambridge (2005)
19. Lim, M.: CyberCivic Space in Indonesia: From Panopticon to Pandemonium. *International Development and Planning Review (Third World Planning Review)* 24(4), 383–400 (2002)
20. Lim, M.: From War-net to Net-War: The Internet and Resistance Identities in Indonesia. *The International Information & Library Review* 35(2-4), 233–248 (2003)
21. Bimber, B.: *Information and American Democracy: Technology in the Evolution of Political Power*. Cambridge University Press, Cambridge, MA (2003)
22. Norris, P.: *Democratic Phoenix: Reinventing Political Activism*. Cambridge University Press, New York (2002)
23. McCaughey, M., Ayers, M. (eds.): *Cyberactivism: Online activism in theory and practice*. Routledge, New York (2003)
24. Bruns, A.: Methodologies for mapping the political blogosphere: An exploration using the IssueCrawler research tool. University of Illinois, Chicago (2010)
25. Lin, J., Halavais, A.: Mapping the blogosphere in America. In: Workshop on the Weblogging Ecosystem in the 13th International World Wide Web Conference (2004)
26. Adamic, L., Glance, N.: The political blogosphere and the 2004 US election. In: *In the Proceedings of the Third International Workshop on Link Discovery*, pp. 36–43 (2005)
27. Kelly, J., Etling, B.: *Mapping Irans Online Public: Politics and Culture in the Persian Blogosphere*, Berkman Center for Internet and Society and Internet & Democracy Project, Harvard Law School (2008)
28. Etling, B., Kelly, J., Faris, R., Palfrey, J.: *Mapping the Arabic Blogosphere: Politics, Culture, and Dissent*, Berkman Center for Internet and Society and Internet & Democracy Project, Harvard Law School (2009)
29. Wigand, R.T.: Some Recent Developments in Organizational Communication: Network Analysis - A Systemic Representation of Communication Relationships. *Communications International Journal of Communication Research* 3(2), 181–200 (1977)
30. Coleman, J.S.: Relational Analysis: A Study of Social Organization with Survey Methods. In: Lazarsfeld, P.P., Pasanella, A.K., Rosenberg, M. (eds.) *Continuities in the Language of Social Research*, pp. 258–266 (1972)
31. Rosenberg, M.: Conditional Relationships. In: Lazarsfeld, P.P., Pasanella, A.K., Rosenberg, M. (eds.) *Continuities in the Language of Social Research*, pp. 133–147 (1972)
32. Coombs, M., Ulicny, B., Jaenisch, H., Handley, J., Faucheux, J.: Formal Analytic Modeling of Bridge Blogs as Personal Narrative: A Case Study in Grounding Interpretation. In: *Proceeding of the Workshop on Social Computing, Behavioral Modeling, and Prediction (SBP)*, Phoenix, pp. 207–217 (2008)

33. Fortunato, S.: Community detection in graphs. *Phys. Reports* 486(3-5), 75–174 (2009)
34. Kritikopoulos, A., Sideri, M., Varlamis, I.: Blogrank: ranking weblogs based on connectivity and similarity features. In: *International Workshop on Advanced Architectures and Algorithms for Internet DELivery and Applications, AAA-IDEA* (2006)
35. Agarwal, N., Liu, H., Tang, L., Yu, P.: Identifying Influential Bloggers in a Community. In: *Proceedings of the 1st International Conference on Web Search and Data Mining (WSDM)*, California, pp. 207–218 (February 10-12, 2008)
36. Jamjoom, M.: Saudi women raise their voices over male guardianship. In: *CNN World* (September 7, 2010),  
<http://bit.ly/fjcpNH>
37. Blodgett, B.M.: And the ringleaders were banned: an examination of protest in virtual worlds. In: *Proceedings of the fourth International Conference on Communities and Technologies (C&T 2009)*, pp. 135–144. ACM Press, New York (2009)
38. Clarks, J., Thernado, N.: Linking the Web and the Street: Internet-Based Dotcause and the Anti-Globalization Movement. *World Development* 34, 50–74 (2005)
39. Saeed, S., Rohde, M., Wulf, V.: ICTs, An alternative sphere for Social Movements in Pakistan: A Research Framework. Paper Presented at IADIS international conference on E-Society, Algarve, Portugal (April 9-12, 2008)
40. Shih, C.-C., Peng, T.-C., Lai, W.-S.: Mining the Blogosphere to Generate Cuisine Hotspot Maps. *Journal of Digital Information Management* 8(6), 396–401 (2010)
41. Ishida, K.: Spam Blog Filtering with Bipartite Graph Clustering and Mutual Detection between Spam Blogs and Words. *Journal of Digital Information Management* 8(2), 108–116 (2010)
42. Esuli, A., Sebastiani, F. S.: A publicly available lexical resource for opinion mining. In: *Proceedings of Language Resources and Evaluation (LREC)*, vol. 6, Genoa, Italy, pp. 417–422 (May 24-26, 2006)



# Evaluating $K$ -Best Site Query on Spatial Objects

Yuan-Ko Huang and Lien-Fa Lin

Department of Information Communication Kao-Yuan University,  
Kaohsiung Country, Taiwan, R.O.C.  
{huangyk,lienfa}@cc.kyu.edu.tw

**Abstract.** A novel query in spatial databases is the  $K$ -Best Site Query ( $KBSQ$  for short). Given a set of objects  $O$ , a set of sites  $S$ , and a user-given value  $K$ , a  $KBSQ$  retrieves the  $K$  sites  $s_1, s_2, \dots, s_K$  from  $S$  such that the total distance from each object to its closest site is minimized. The  $KBSQ$  is indeed an important type of spatial queries with many real applications. In this paper, we investigate how to efficiently process the  $KBSQ$ . We first propose a straightforward approach with a cost analysis, and then develop the  $K$  Best Site Query ( $KBSQ$ ) algorithm combined with the existing spatial indexes to improve the performance of processing  $KBSQ$ .

**Keywords:** spatial databases,  $K$ -Best Site Query,  $KBSQ$ , spatial queries.

## 1 Introduction

With the fast advances of positioning techniques in mobile systems, spatial databases that aim at efficiently managing spatial objects are becoming more powerful and hence attract more attention than ever. Many applications, such as mobile communication systems, traffic control systems, and geographical information systems, can benefit from efficient processing of spatial queries [1,2,3,4,5,6,7]. In this paper, we present a novel and important type of spatial queries, namely the  $K$ -Best Site Query ( $KBSQ$  for short). Given a set of objects  $O$ , a set of sites  $S$ , and a user-given value  $K$ , a  $KBSQ$  retrieves the  $K$  sites  $s_1, s_2, \dots, s_K$  from  $S$  such that

$$\sum_{o_i \in O} d(o_i, s_j)$$

is minimized, where  $d(o_i, s_j)$  refers to the distance between object  $o_i$  and its closest site  $s_j \in \{s_1, s_2, \dots, s_K\}$ . We term the sites retrieved by executing  $KBSQ$  the *best sites* (or *bs* for short).

The  $KBSQ$  problem arises in many fields and application domains.

- As an example of real-world scenario, consider a set of soldiers on the battlefields that is fighting the enemy. In order to immediately support the injured soldiers, we need to choose  $K$  sites to build the emergicenters. To achieve the fastest response time, the sum of distances from each battlefield to its closest emergicenter should be minimized.

- Another real-world example is that the McDonald’s Corporation may ask “what are the optimal locations in a city to open new McDonald’s stores.” In this case, the  $KBSQ$  can be used to find out the  $K$  best sites so that every customer can rapidly reach his/her closest store.

Let us use an example in Figure 1 to illustrate the  $KBSQ$  problem, where six objects  $o_1, o_2, \dots, o_6$  and four sites  $s_1, s_2, \dots, s_4$  are depicted as circles and rectangles, respectively. Assume that two best sites (i.e.,  $2bs$ ) are to be found in this example. There are six combinations  $(s_1, s_2), (s_1, s_3), \dots, (s_3, s_4)$ , and one combination would be the result of  $KBSQ$ . As we can see, the sum of distances from objects  $o_1, o_2, o_3$  to their closest site  $s_3$  is equal to 9, and the sum of distances between objects  $o_4, o_5, o_6$  and site  $s_1$  is equal to 12. Because combination  $(s_1, s_3)$  leads to the minimum total distance (i.e.,  $9 + 12 = 21$ ), the two sites  $s_1$  and  $s_3$  are the  $2bs$ .

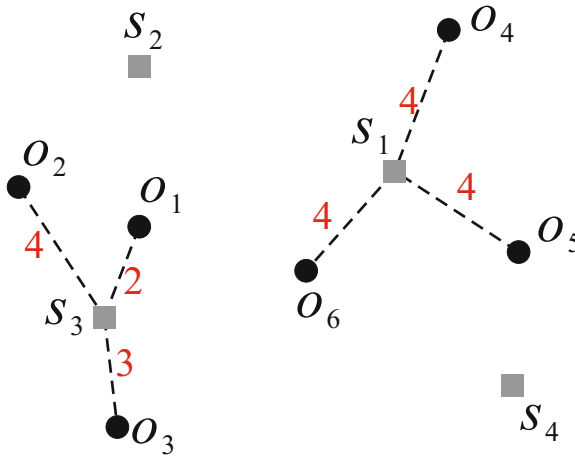


Fig. 1. An example of  $KBSQ$

To process the  $KBSQ$ , the closest site for each object needs to be first determined and then the distance between object and its closest site is computed so as to find the best combination of  $K$  sites. When a database is large, it is crucial to avoid reading the entire dataset in identifying the  $K$  best sites. For saving CPU and I/O costs, we develop efficient method combined with the existing spatial indexes to avoid unnecessary reading of the entire dataset. The major contributions of this paper are summarized as follows.

- We present a novel query, namely the  $K$  Best Site Query, which is indeed an important type of spatial queries with many real applications.
- We propose a straightforward approach to process the  $KBSQ$  and also analyze the processing cost required for this approach.

- An efficient algorithm, namely the *K Best Site Query (KBSQ)* algorithm, operates by the support of R-tree [8] and Voronoi diagram [9] to improve the performance of *KBSQ*.

The rest of this paper is organized as follows. In Section 2, we discuss some related works on processing spatial queries similar to *KBSQ*, and point out their differences. In Section 3, the straightforward approach and its cost analysis is presented. Section 4 describes the *KBSQ* algorithm with the used indexes. Section 5 concludes the paper with directions on future work.

## 2 Related Work

In recent years, some queries similar to the *KBSQ* are presented, including the Reverse Nearest Neighbor Query (*RNNQ*) [10], the Group Nearest Neighbor Query (*GNNQ*) [11], and the Min-Dist Optimal-Location Query (*MDOLQ*) [12]. Several methods have been designed to efficiently process these similar queries. However, the query results obtained by executing these queries are quite different from that of the *KBSQ*. Also, the proposed methods cannot be directly used to answer the *KBSQ*. In the following, we investigate why the existing methods for processing the similar queries cannot be applied to the *KBSQ* separately.

### 2.1 Methods for RNNQ

Given a set of object  $O$  and a site  $s$ , a *RNNQ* can be used to retrieve the object  $o \in O$  whose closest site is  $s$ . The object  $o$  is termed a *RNN* of  $s$ . An intuitive way for finding the query result of *KBSQ* is to utilize the *RNNQ* to find the *RNNs* for each site. Then, the  $K$  sites having the maximum number of *RNNs* (meaning that they are closer to most of the objects) are chosen to be the  $K$  best sites.

Taking Figure 2 as an example, the *RNNs* of site  $s_1$  can be determined by executing the *RNNQ* and its *RNNs* are objects  $o_4$  and  $o_6$ . Similarly, the *RNN* of sites  $s_2$ ,  $s_3$ , and  $s_4$  are determined as  $o_1$  and  $o_2$ ,  $o_3$ , and  $o_5$ , respectively. As sites  $s_1$  and  $s_2$  have the maximum number of *RNNs*, they can be the *2bs* for the *KBSQ*. However, sites  $s_1$  and  $s_2$  lead to the total distance 24 (i.e.,  $d(o_4, s_1) + d(o_5, s_1) + d(o_6, s_1) + d(o_1, s_2) + d(o_2, s_2) + d(o_3, s_2)$ ), which is greater than the total distance 22 as sites  $s_1$  and  $s_3$  are chosen to be the *2bs*. As a result, the intuition of using the *RNNQ* result to be the *KBSQ* result is infeasible.

### 2.2 Methods for GNNQ

A *GNNQ* retrieves a site  $s$  from a set of sites  $S$  such that the total distance from  $s$  to all objects is the minimum among all sites in  $S$ . Here, the result  $s$  of *GNNQ* is called a *GNN*. To find the  $K$  best sites, we can repeatedly evaluate the *GNNQ*  $K$  times so as to retrieve the first  $K$  *GNNs*. It means that the sum of distances between these  $K$  *GNNs* and all objects is minimum, and thus they can be the  $K$

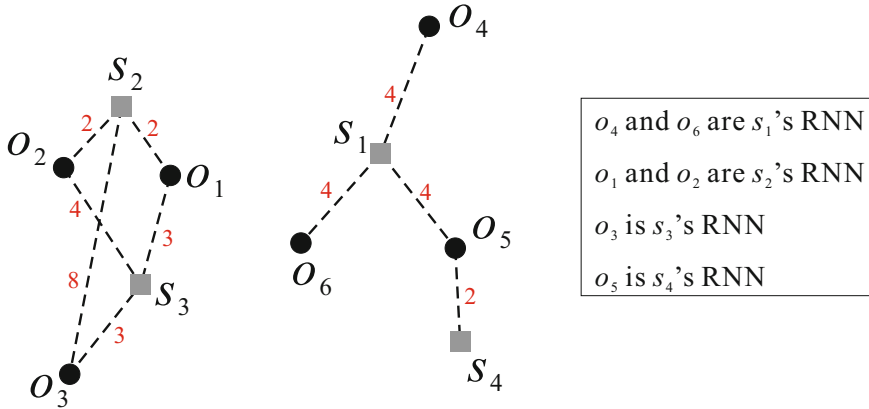


Fig. 2. An example of RNNQ

$bs$ . However, in some cases the result obtained by executing the  $GNNQ$   $K$  times is still different from the exact result of  $KBSQ$ .

Let us consider an example shown in Figure 3, where 2 $bs$  are required. As shown in Figure 3(a), the first and second  $GNNs$  are sites  $s_3$  and  $s_1$ , respectively. As such, the 2 $bs$  are  $s_3$  and  $s_1$ , and the total distance  $d(o_1, s_1) + d(o_2, s_1) + d(o_4, s_1) + d(o_3, s_3) + d(o_5, s_3) + d(o_6, s_3) = 23$ . However, another combination ( $s_2, s_4$ ) shown in Figure 3(b) can further reduce the total distance to 13. Therefore, using the way of executing  $GNNQ$   $K$  times to find the  $K$  best sites could return incorrect result.

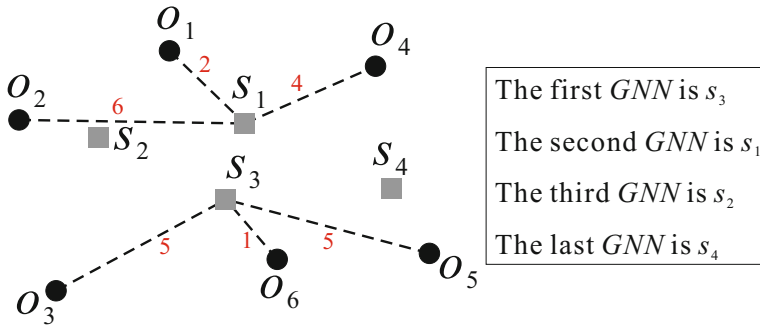
### 2.3 Methods for MDOLQ

Given a set of objects  $O$  and a set of sites  $S$ , a  $MDOLQ$  can be used to find out a new site  $s \notin S$  so as to minimize

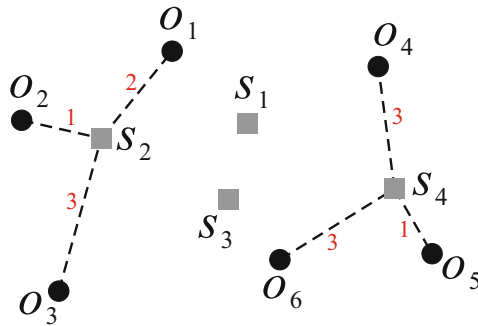
$$\sum_{o_i \in O} d(o_i, s_j),$$

where  $d(o_i, s_j)$  is the distance between object  $o_i$  and its closest site  $s_j \in S \cup \{s\}$ . At first glance, the  $MDOLQ$  is more similar to the  $KBSQ$  than the other queries mentioned above. However, using the  $MDOLQ$  to obtain the  $K$  best sites may still lead to incorrect result.

Consider an example of using  $MDOLQ$  to find the  $K$  best sites in Figure 4. As 2 $bs$  are to be found, we can evaluate the  $MDOLQ$  two times to obtain the result. In the first iteration (as shown in Figure 4(a)), the site  $s_1$  becomes the first  $bs$  because it has the minimum total distance to all objects. Then, the  $MDOLQ$  is executed again by taking into account the remaining sites  $s_2, s_3$ , and  $s_4$ . As the site  $s_2$  can reduce more distance compared to the other two sites, it becomes the second  $bs$  (shown in Figure 4(b)). Finally, 2 $bs$  are  $s_1$  and  $s_2$  and the total distance is computed as  $d(o_4, s_1) + d(o_5, s_1) + d(o_6, s_1) + d(o_1, s_2) + d(o_2, s_2) + d(o_3, s_2) = 20$ .



(a) incorrect result



(b) correct result

Fig. 3. An example of GNNQ

However, the computed distance is not minimum and can be further reduced. As we can see in Figure 4(c), if  $s_2$  and  $s_4$  are chosen to be the 2bs, the total distance can decrease to 16.

### 3 Straightforward Approach

In this section, we first propose a straightforward approach to solve the *KBSQ* problem, and then analyze the processing cost required for this approach. Assume that there are  $n$  objects and  $m$  sites, and the  $K$  bs would be chosen from the  $m$  sites. The straightforward approach consists of three steps. The first step is to compute the distance  $d(o_i, s_j)$  from each object  $o_i$  ( $1 \leq i \leq n$ ) to each site  $s_j$  ( $1 \leq j \leq m$ ). As the  $K$  best sites are needed to be retrieved, there are totally  $C_K^m$  possible combinations and each of the combinations comprises  $K$  sites. The second step is to consider all of the combinations. For each combination, the distance from each object to its closest site is determined so as to compute the total distance. In the last step, the combination of  $K$  sites having the minimum

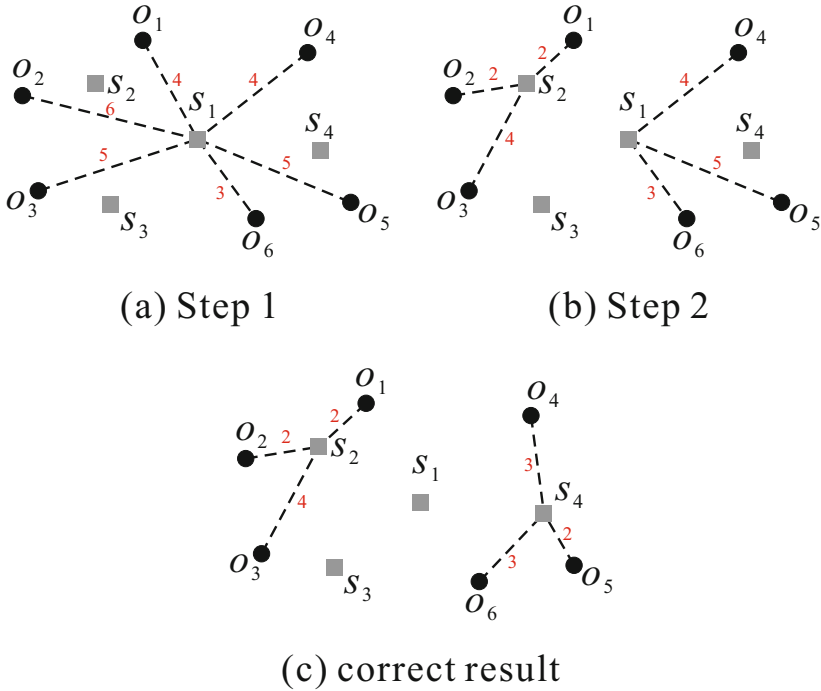


Fig. 4. An example of MDOLQ

total distance is chosen to be the query result of  $KBSQ$ . The procedure of the straightforward approach is detailed in Algorithm 1.

Figure 5 illustrates the three steps of the straightforward approach. As shown in Figure 5(a), the distances between objects and sites are computed and stored in a table, in which a tuple represents the distance from an object to all sites. Then, the  $C_K^m$  combinations of  $K$  sites are considered so that  $C_K^m$  tables are generated (shown in Figure 5(b)). For each table, the minimum attribute value of each tuple (depicted as gray box) refers to the distance between an object and its closest site. As such, the total distance for each combination can be computed by summing up the minimum attribute value of each tuple. Finally, in Figure 5(c) the combination 1 of  $K$  sites can be the  $K$  bs because its total distance is minimum among all combinations.

Since the straightforward approach includes three steps, we consider the three steps individually to analyze the processing cost. Let  $m$  and  $n$  be the numbers of sites and objects, respectively. Then, the time complexity of the first step is  $m \times n$  because the distances between all objects and sites have to be computed. In the second step,  $C_K^m$  combinations are considered and thus the complexity is  $C_K^m \times n \times K$ . Finally, the combination having the minimum total distance is

---

**Algorithm 1.** The straightforward approach

---

```

Input : A number  $K$ , a set of  $n$  objects, and a set of  $m$  sites
Output: The  $K$   $bs$ 
/* Step 1 */
1 foreach object  $o_i$  do
2   foreach site  $s_j$  do
3      $\lfloor$  compute the distance  $d(o_i, s_j)$  from  $o_i$  to  $s_j$ ;
/* Step 2 */
4 foreach combination  $c \in C_K^m$  do
5   foreach object  $o_i$  do
6      $\lfloor$  determine the distance  $d(o_i, s_j)$  from  $o_i$  to its closest site  $s_j$ ;
7     compute the total distance  $d_c$  for combination  $c$  as  $\sum_{o_i} d(o_i, s_j)$ ;
/* Step 3 */
8 return the combination  $c$  having the minimum total distance;

```

---

determined among all combinations so that the complexity of the last step is  $C_K^m$ . The processing cost of the straightforward approach is represented as

$$m \times n + C_K^m \times n \times K + C_K^m.$$

## 4 *KBSQ* Algorithm

The above approach is performed without any index support, which is a major weakness in dealing with large datasets. In this section, we propose the *KBSQ* algorithm combined with the existing indexes R-tree and Voronoi diagram to efficiently process the *KBSQ*.

Recall that, to process the *KBSQ*, we need to find the closest site  $s$  for each object  $o$  (that is, finding the *RNN*  $o$  of site  $s$ ). As the Voronoi diagram can be used to effectively determine the *RNN* of each site [13], we divide the data space so that each site has its own Voronoi cell. For example, in Figure 6(b), the four sites  $s_1, s_2, s_3,$  and  $s_4$  have their corresponding Voronoi cells  $V_1, V_2, V_3,$  and  $V_4$ , respectively. Taking the cell  $V_1$  as an example, if object  $o$  lies in  $V_1$ , then  $o$  must be the *RNN* of site  $s_1$ . Based on this characteristic, object  $o$  needs not be considered in finding the *RNNs* for the other sites. Then, we use the R-tree, which is a height-balanced indexing structure, to index the objects. In a R-tree, objects are recursively grouped in a bottom-up manner according to their locations. For instance, in Figure 6(a), eight objects  $o_1, o_2, \dots, o_8$  are grouped into four leaf nodes  $E_4$  to  $E_7$  (i.e., the minimum bounding rectangle MBR enclosing the objects). Then, nodes  $E_4$  to  $E_7$  are recursively grouped into nodes  $E_2$  and  $E_3$ , that become the entries of the root node  $E_1$ .

Combined with the R-tree and Voronoi diagram, we design the following pruning criteria to greatly reduce the number of objects considered in query processing.

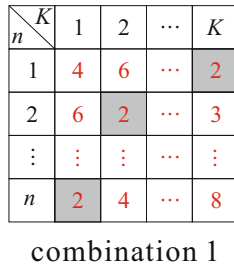
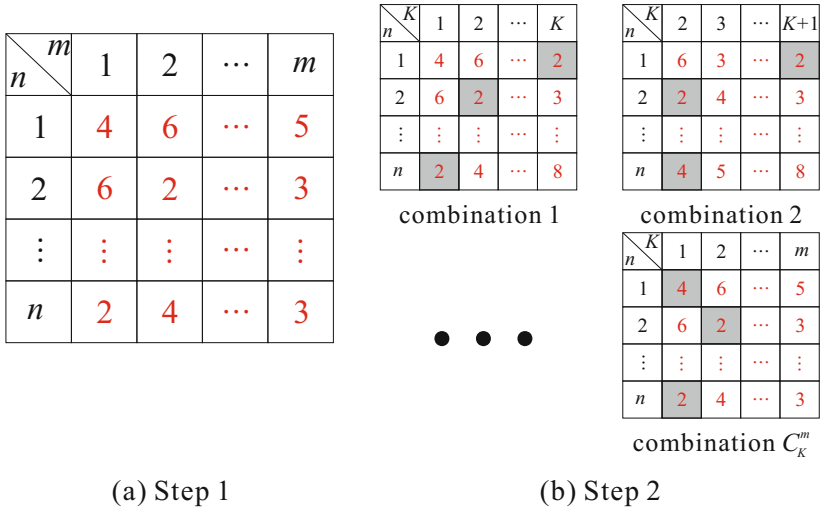


Fig. 5. Straightforward approach

- **Pruning objects.** Given an object  $o$  and the  $K$  sites  $s_1, s_2, \dots, s_K$ , if  $o$  lies in the Voronoi cell  $V_i$  of one site  $s_i \in \{s_1, s_2, \dots, s_K\}$ , then the distances between the object  $o$  and the other  $K - 1$  sites need not be computed so as to reduce the processing cost.
- **Pruning MBRs.** Given a MBR  $E$  enclosing a number of objects and the  $K$  sites  $s_1, s_2, \dots, s_K$ , if  $E$  is fully contained in the cell  $V_i$  of one site  $s_i \in \{s_1, s_2, \dots, s_K\}$ , then the distances from all objects enclosed in  $E$  to the other  $K - 1$  sites would not be computed.

To find the  $K$  bs for the  $KBSQ$ , we need to consider  $C_K^m$  combinations of  $K$  sites. For each combination of  $K$  sites  $s_1, s_2, \dots, s_K$  with their corresponding Voronoi cells  $V_1, V_2, \dots, V_K$ , the processing procedure begins with the R-tree root node and proceeds down the tree. When an internal node  $E$  (i.e., MBR  $E$ )



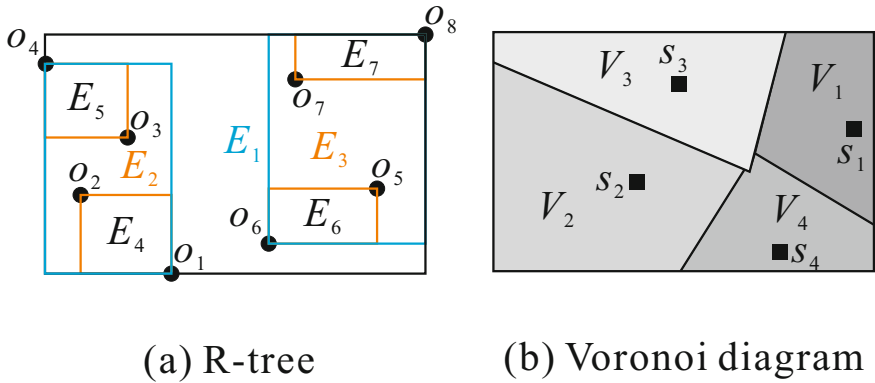


Fig. 6. R-tree and Voronoi diagram

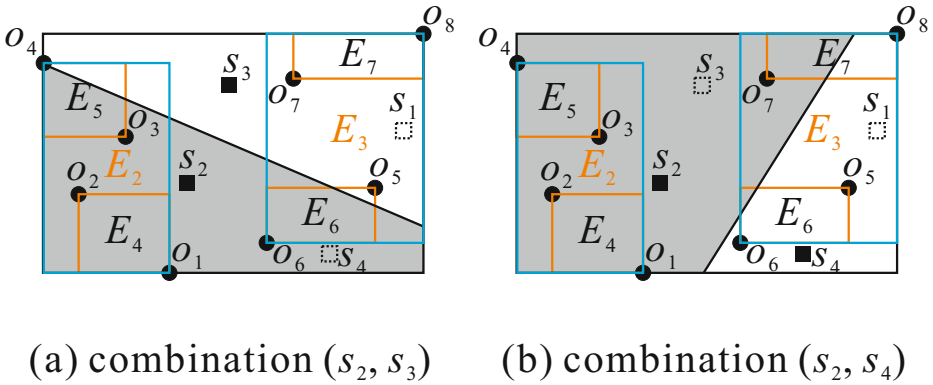


Fig. 7. *KBSQ* algorithm

of the R-tree is visited, the pruning criterion 2 is utilized to determine which site is the closest site of the objects enclosed in  $E$ . If the MBR  $E$  is not fully contained in any of the  $K$  Voronoi cells, then the child nodes of  $E$  need to be further visited. When a leaf node of the R-tree is checked, the pruning criterion 1 is imposed on the entries (i.e., objects) of this leaf node. After the traversal of the R-tree, the total distance for the combination of  $K$  sites  $s_1, s_2, \dots, s_K$  can be computed. By taking into account the total combinations, the combination of  $K$  sites whose total distance is minimum would be the query result of the *KBSQ*. Algorithm 2 gives the details for the *KBSQ* algorithm.

Figure 7 continues the previous example in Figure 6 to illustrate the processing procedure, where there are eight objects  $o_1$  to  $o_8$  and four sites  $s_1$  to  $s_4$  in data space. Assume that the combination  $(s_2, s_3)$  is considered and the Voronoi cells of sites  $s_2$  and  $s_3$  are shown in Figure 7(a). As the MBR  $E_2$  is not fully contained in the Voronoi cell  $V_2$  of site  $s_2$ , the MBRs  $E_4$  and  $E_5$  still need to be visited.

**Algorithm 2.** The  $KBSQ$  algorithm

---

**Input** : A number  $K$ , a set of  $n$  objects indexed by R-tree, and a set of  $m$  sites indexed by Voronoi diagram

**Output:** The  $K$   $bs$

```

1 create an empty queue  $Q$ ;
2 foreach combination  $c \in C_K^m$  do
3   insert the root node of R-tree into  $Q$ ;
4   while  $Q$  is not empty do
5     de-queue  $e$ ;
6     if  $e$  corresponds to an internal node  $E_i$  then
7       if  $E_i$  is fully contained in a voronoi cell  $V_j$  then
8         foreach object  $o_i$  enclosed in  $E_i$  do
9           compute the distance  $d(o_i, s_j)$  from  $o_i$  to site  $s_j$ ;
10          /*  $e$  is not fully contained in any of the  $K$  voronoi cells */
11          else
12            insert child nodes of  $E_i$  into  $Q$ ;
13          /*  $e$  corresponds to an object  $o_i$  */
14          else
15            if  $o_i$  is enclosed by a voronoi cell  $V_j$  then
16              compute the distance  $d(o_i, s_j)$  from  $o_i$  to site  $s_j$ ;
17          compute the total distance  $d_c$  for combination  $c$  as  $\sum_{o_i} d(o_i, s_j)$ ;
18 return the combination  $c$  having the minimum total distance;
```

---

When the MBR  $E_4$  is checked, based on the pruning criterion 2 the distances from objects  $o_1$  and  $o_2$  to site  $s_3$  would not be computed because their closest site is  $s_2$ . Similarly, the closest site of the objects  $o_7$  and  $o_8$  enclosed in MBR  $E_7$  is determined as site  $s_3$ . As for objects  $o_3$  to  $o_6$ , their closest sites can be found based on the pruning criterion 1. Having determined the closest site of each object, the total distance for combination  $(s_2, s_3)$  are obtained. Consider another combination  $(s_2, s_4)$  shown in Figure 7(b). The closest site  $s_2$  of four objects  $o_1$  to  $o_4$  enclosed in MBR  $E_2$  can be found when  $E_2$  is visited. Also, we can compute the total distance for the combination  $(s_2, s_4)$  after finding the closest sites for objects  $o_5$  to  $o_8$ . By comparing the distances for all combinations, the  $2bs$  are retrieved.

## 5 Conclusions

In this paper, we focused on processing the  $K$  Best Site Query ( $KBSQ$ ) which is a novel and important type of spatial queries. We highlighted the limitations of the previous approaches for the queries similar to the  $KBSQ$ , including the  $RNNQ$ , the  $GNNQ$ , and the  $MDOLQ$ . To solve the  $KBSQ$  problem, we first proposed a straightforward approach and then analyzed its processing cost. In order to improve the performance of processing the  $KBSQ$ , we further proposed

a *KBSQ* algorithm combined with the R-tree and Voronoi diagram to greatly reduce the CPU and I/O costs.

Our next step is to process the *KBSQ* for moving objects with fixed or uncertain velocity. More complicated issues will be introduced because of the movement of objects. Finally, we would like to extend the proposed approach to process the *KBSQ* in road network.

## Acknowledgment

This work was supported by National Science Council of Taiwan (R.O.C.) under Grants NSC 99-2221-E-244 -017.

## References

1. Benetis, R., Jensen, C.S., Karciuskas, G., Saltenis, S.: Nearest neighbor and reverse nearest neighbor queries for moving objects. *VLDB Journal* 15(3), 229–249 (2006)
2. Hakkoymaz, V.: A specification model for temporal and spatial relations of segments in multimedia presentations. *Journal of Digital Information Management* 8(2), 136–146 (2010)
3. Huang, Y.-K., Chen, C.-C., Lee, C.: Continuous k-nearest neighbor query for moving objects with uncertain velocity. *GeoInformatica* 13(1), 1–25 (2009)
4. Huang, Y.-K., Liao, S.-J., Lee, C.: Evaluating continuous k-nearest neighbor query on moving objects with uncertainty. *Information Systems* 34(4-5), 415–437 (2009)
5. Mokbel, M.F., Xiong, X., Aref, W.G.: Sina: Scalable incremental processing of continuous queries in spatio-temporal databases. In: *Proceedings of the ACM SIGMOD* (2004)
6. Pagel, B.-U., Six, H.-W., Winter, M.: Window query-optimal clustering of spatial objects. In: *ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (1995)
7. Papadias, D., Tao, Y., Mouratidis, K., Hui, C.K.: Aggregate nearest neighbor queries in spatial databases. *ACM Trans. Database Syst.* 30(2), 529–576 (2005)
8. Guttman, A.: R-trees: A dynamic index structure for spatial searching. In: *ACM SIGMOD Conf.*, pp. 47–57 (1984)
9. Samet, H.: *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading (1990)
10. Korn, F., Muthukrishnan, S.: Influence sets based on reverse nearest neighbor queries. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, May 16-18, pp. 201–212 (2000)
11. Papadias, D., Shen, Q., Tao, Y., Mouratidis, K.: Group nearest neighbor queries. In: *Proceedings of the International Conference on Data Engineering* (2004)
12. Zhang, D., Du, Y., Xia, T., Tao, Y.: Progressive computation of the min-dist optimal-location query. In: *Proceedings of the VLDB* (2006)
13. Zhang, J., Zhu, M., Papadias, D., Tao, Y., Lee, D.L.: Location-based spatial queries. In: *ACM SIGMOD*, San Diego, California, USA, (June 9-12, 2003)

# Performance of Annotation-Based Image Retrieval

Phani Kidambi, Mary Fendley, and S. Narayanan

College of Engineering & Computer Science, Wright State University,  
3640 Colonel Glenn Highway, Dayton, OHIO, USA – 45435  
{phani.kidambi, mary.fendley, s.narayanan}@wright.edu

**Abstract.** As the proliferation of available and useful images on the web grows, novel methods and effective techniques are needed to retrieve these images in an efficient manner. Currently major commercial search engines utilize a process known as Annotation Based Image Retrieval (ABIR) to execute search requests focused on image retrieval. The ABIR technique primarily relies on the textual information associated with an image to complete the search and retrieval process. Using the game of cricket as the domain, we describe a benchmarking study that evaluates the effectiveness of three popular search engines in executing image-based searches. Second, we present details of an empirical study aimed at quantifying the impact of inter-human variability of the annotations on the effectiveness of search engines. Both these efforts are aimed at better understanding the challenges with image search and retrieval methods that purely rely on ad hoc annotations provided by the humans.

## 1 Introduction

The Internet houses an inexhaustible number of images due to digital technologies that allow images to easily be uploaded to the web. These images are sought for recreation, education, and scientific purposes. As the sheer number of images increases, the user is met with the problem of image overload, where the user has access to more images than can be viewed, with only a portion of them being relevant [1]. Major search engines use a technique called Annotation Based Image Retrieval (ABIR) to perform queries focusing on image retrieval. The ABIR technique relies on text based information regarding the image in order to execute the search and retrieval process. Annotation involves the professional judgment of an individual to interpret material and its content.

To complete a search, an ABIR-driven engine employs a number of standard steps [2]. Images are retrieved by evaluating the vector of word frequencies in their annotations and returning the images with the closest vectors. A relevancy ranking is calculated by evaluating the degree of the match of the order and separation of the words that exists between the search terms and the annotation of each individual image [3]. Thus, even though the user is searching for images, the images that are retrieved are actually determined by the textual annotation. This annotation usually consists of the manually assigned keywords or the text associated with the images such as captions.

While a significant body of research exists to evaluate the effectiveness of textual information retrieval processes [4, 5], there has been very little focus on evaluating image retrieval on the Internet. The first part of this article describes a benchmarking study which evaluates the effectiveness of three popular search engines in executing image based searches. The domain selected to assess their performance is the game of cricket. The second phase of this study assesses the impact of variability in human annotation of web images on the image search and retrieval process.

## 2 Background

### 2.1 Image Annotation

Image search and retrieval through ABIR draws its capabilities by taking advantage of natural language text to represent semantic content of images and the search needs of the user [6]. Studies have shown that textual information is key to image retrieval using both video and photo image retrieval [7, 8].

To search and retrieve the images in a database (or the World Wide Web), the current commercial search engines provide some text descriptors to the images and retrieve them based on the text. Figure 1 presents an overview of the background in this area.

Chen et al. [9] defined two methods of providing text to images: categorization & annotation. Categorization is the association of a predefined category of text to an image while annotation provides the image with detailed text descriptors. Bimbo [10] stated that three different types of information can be associated with an image that include

- Content-independent metadata – the metadata is related to the image but cannot be extracted with computer vision algorithms (example: date when the image was taken)
- Content-dependent metadata – the metadata can be extracted using content based image retrieval algorithms (example: based on color, texture & shape)
- Content-descriptive metadata – the semantics in the image which cannot be extracted using computer vision algorithms and need the expertise of a domain expert to annotate the image.

Annotation based on content-dependent metadata for an image can be generated using computer vision algorithms while for content-descriptive metadata, human annotation is required. The computer vision algorithms captures only one aspect of the image (color, texture, shape) and even a combination of these low level image features do not capture the semantic meaning (high level textual features) of the image. Liu et al. [11] termed the gap between the low level image features and the high level textual features as the semantic gap. The computer vision algorithms are still at an early stage and hence most of the commercial search engines focus on retrieving the images based on the text descriptors rather than the content of the image. The human can associate some text descriptor for an image by i) free flowing text ii) keywords from restricted vocabularies iii) ontology classification [12]. Though manual annotation of

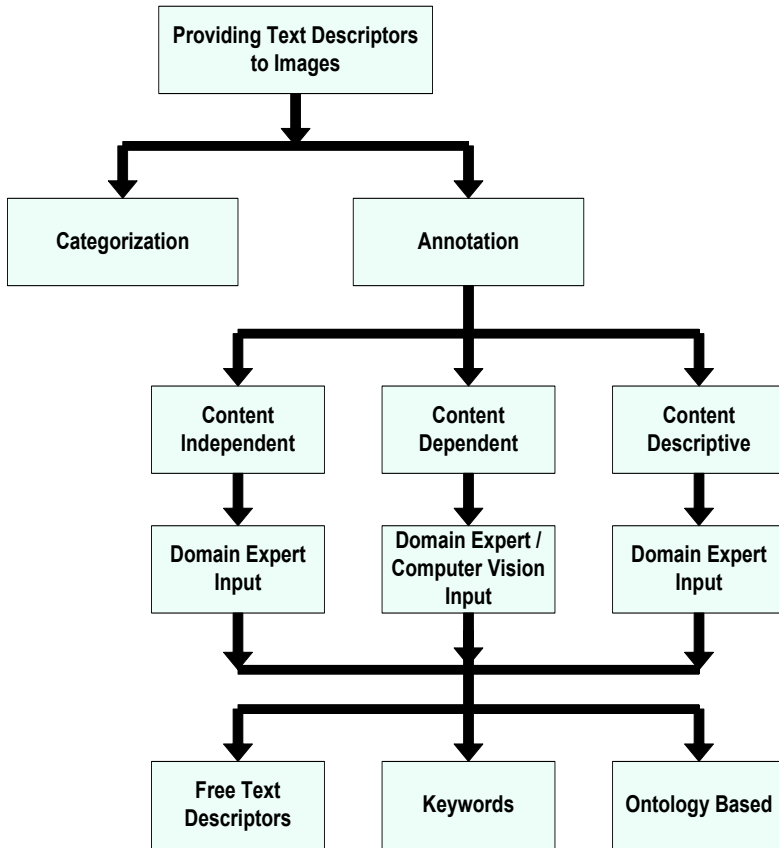


Fig. 1. Overview of the background

images is labor intensive and time consuming, it still is the preferred way of annotating images [13]. To reduce the effort of a human, an innovative way known as the ESP game was developed to collect the text descriptors by Ahn et al. [14]. In this Extra Sensory Perception (ESP) game, two players are paired randomly, and are asked to generate keywords for a particular image in a given time interval. Whenever the keywords suggested by the two players match, the image is annotated with that keyword.

## 2.2 Image Retrieval

Evaluating the effectiveness of information retrieval is important but challenging. Most researchers adopt the definition of evaluation posited by Hernon et al [15] which states that evaluation is

“ the process of identifying and collecting data about specific services or activities, establishing criteria by which their success can be assessed, and determining both the quality of the service or activity and the degree to which the service or activity accomplishes stated goals and objectives.”

Meadow et al [16] classified information retrieval measures into two categories: evaluation of performance (descriptive of what happens during the use of the information retrieval system) and evaluation of outcome (descriptive of the results obtained). Hersh [17] also classified evaluation into two categories, although different from those proposed by Meadow, Hersh identifies: macro evaluation (investigates information retrieval system as a whole and its overall benefit) and micro evaluation (investigates different components of the system and their impact on the performance in a controlled setting). Lancaster et al [18] defined three levels of evaluation. The first level evaluates the effectiveness of the system, the second level evaluates the cost effectiveness and the third level evaluates cost benefits of the system. While, Smith [19] proposed several measures for image retrieval evaluation including precision, recall, fallout and F-measure ( $F\text{-measure} = 2 (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$ ). Finally, Cooper [20] suggests Expected Search Length (ESL) as an alternative to recall and precision. ESL measures the number of unwanted documents the user can expect to examine before finding the desired number of relevant documents.

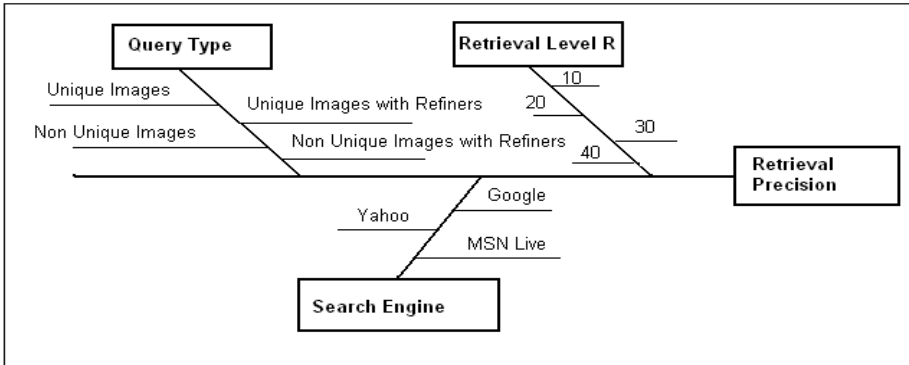
Though the research literature contains numerous studies on various ways of ascribing text to images by humans and computers, and the evaluation metrics for information and image retrieval, there is a dearth of literature benchmarking the performance of image search engines and investigating the human role in the annotation of images for search and retrieval engines.

This paper begins to fill this information gap, employing a systematic approach to evaluate search engines based on a number of independent factors including query types, number of images retrieved and the type of search engine. The research also conducts a systematic study to investigate the role of humans in the annotation of images. The remainder of this paper discusses the experiments that have been performed, the results, and their implications.

### 3 Benchmarking Commercial Image Search Engines

#### 3.1 Design of Experiment

**Methodology:** Research studies [19, 21], have showed that the quality of images retrieved are typically a function of query formulation, type of search engine and the retrieval level. These independent factors and their levels are illustrated in the Ishikawa diagram illustrated in Figure 2.



**Fig. 2.** Ishikawa diagram of independent factors used in this study

In this experiment three search engines: Google, Yahoo, and MSN Live are evaluated for varying query types and retrieval levels. Details on the query types and query levels are provided below.

Query Types: Broder [22], proposed a three-pronged approach of web searching types for text retrieval: navigational, informational and transactional. Navigational searches are those where the user intends to find a specific website. Informational searches intend to find some information assumed to be present on one or more web pages. Transactional searches perform some web mediated activity, i.e., the purpose is to reach a site where further interactions will happen. Unfortunately, Broder's query types cannot be easily extended to image retrieval solutions since the end goal of the user in these two scenarios varies significantly.

Ensor & McGregor [23] summarized that the user search requests for images fall into four different categories:

- Search for unique images – The property of uniqueness is a request for the visual representation of an entity where the desired entity (image) can be differentiated from every other occurrence of the same entity type. An example is – “find the image of Sachin Tendulkar”.
- Search for unique images with refiners – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – “find the image of Sachin Tendulkar in 2004”.
- Search for non – unique images – The property of non – uniqueness is a request for the visual representation of an entity where the desired entity (image) cannot be differentiated from every other occurrence of the same entity type. An example is – “find the images of Indian cricketers”.
- Search for non – unique images with identifiers – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – “find images of Indians waving the Indian flag”.



Search Engines: According to Nielsen’s May 2009 ratings [24], Google’s search engine accounts for 63.3% of the total searches on the internet, Yahoo’s accounts for 17.3% and MSN Live accounts for 9.4%. These three search engines execute 90% of the total searches conducted on the internet. It is for this reason that they have been selected for comparison purposes in this study.

**Experiment:** To carry out the evaluation, a user-centered interpretative approach, based on the actual information-seeking behavior of real users [19] was employed. Since this research is focused on the domain specific evaluation of the system, a subject matter expert in the domain of the game of Cricket, was used to evaluate the existing search engines. Five queries for each query type are chosen. The queries consist of multi-word queries, related to the game of cricket, as shown in Table 1.

**Table 1.** Query Formulations for various Query types

Query Types	Queries
Unique Images	MS Dhoni Vijay Bharadwaj Ricky Pointing Gary Sobers Abey Kuruville
Unique Images with refiners	Kapil Dev lifting World Cup Sreesanth + beamer + Pietersen Andy Flower + protest + black band Allan Donald + run out+ WC semifinal '99 Inzamam Ul Haq hitting a spectator + Canada
Non-Unique Images	Indian Cricket Players Surrey Cricket Team Ashes (Eng vs Aus) Cricket Players Huddle Rajasthan Royals + IPL
Non-Unique Images with refiners	Victorious Indian Team + 20-20 WC SA chasing 438 Aus players with World Cup 2007 SL protesting against Aus + walking out of the ground Eng vs SA + WC stalled by rain + 1992

The queries associated with “unique images” are all internationally known cricket players from different playing eras. The queries associated with “unique images with refiners” are related to a cricket player involved in a context such as winning a world cup. The queries associated with “non-unique images” are internationally well known cricket teams and finally the queries associated with “non-unique images with refiners” are internationally known cricket teams involved in a context similar to winning a world cup.

Each query is run on each of the three search engines. The first forty images retrieved in each search run are evaluated for relevance by the subject matter expert based on his knowledge. Relevance is determined in a binary manner. That is the image is either deemed relevant or not relevant. In instances when the same image appears on different websites, these are evaluated as different images and each is evaluated for relevance. In instances where the same images appear in multiple places on the same website, the first image is evaluated for relevance and the other images are considered not relevant. Additionally, if the image retrieved is not accessible due to technical difficulties in the site domain, the image is considered to be non-relevant. In order to obtain a stable performance measurement of image search engines, all the searches are performed within a short period of time (one hour) and the relevance of the images is decided by the subject matter expert.

### 3.2 Results

Traditionally evaluation for information retrieval has been based on the effectiveness ratios of precision (proportion of retrieved documents that are relevant) and recall (proportion of the relevant documents that are retrieved) [19]. Since the World Wide Web is growing constantly, obtaining an exact measure of recall requires knowledge of all relevant documents in the collection. Given the sheer volume of documents this is, for all practical purposes, impossible. Because of this, recall and any measures related to recall cannot be readily used for evaluation. This necessitates that the evaluation be based on the effectiveness ratios of precision. For purposes of this evaluation precision is defined as the number of relevant images retrieved to the total number of images retrieved. The search engines were evaluated based on the precision at a retrieval length  $R$  at  $R=10, 20, 30$  and  $40$ .

To check the adequacy of the factors thought, a factorial analysis was conducted and the results were analyzed using the analysis of variance (ANOVA) method. As previously discussed, the factors that were hypothesized to have a significant effect on the average precision of the retrieved results are Query Type, Search Engine and Retrieval Level.

The response variable is the average precision at retrieval length  $R$  which is defined as the ratio of the relevant retrievals to the overall number of images retrieved.

### Hypothesis

Null Hypothesis:  $H_0$ : There is no significant effect of Query Type, Search Engine or Retrieval level  $R$  on the precision of the retrieved results.

Alternate Hypothesis:  $H_1$ : There is significant effect of Query Type, Search Engine or Retrieval level  $R$  on the precision of the retrieved results.

### Statistical Analysis

Data were collected for the 48 experimental trials of the  $4 \times 3 \times 4$  full factorial design that was run five times.

At the 99 % confidence level, the ANOVA results show that there is a significant effect of the main effects,

- (A) Query Type,  $F_{(3,192)} = 55.70, p < 0.0001$
- (B) Search Engine,  $F_{(2,192)} = 14.02, p < 0.0001$  and
- (C) Retrieval Level R,  $F_{(3,192)} = 4.46, p < 0.0001$

and there are no effects due to interactions between the main effects. The ANOVA results of the overall model (taking all the main effects and interactions into consideration) are also significant at the 99% confidence level.

The results clearly show that all the main effects are significant; hence we can further analyze the response variable.

**Performance Evaluation**

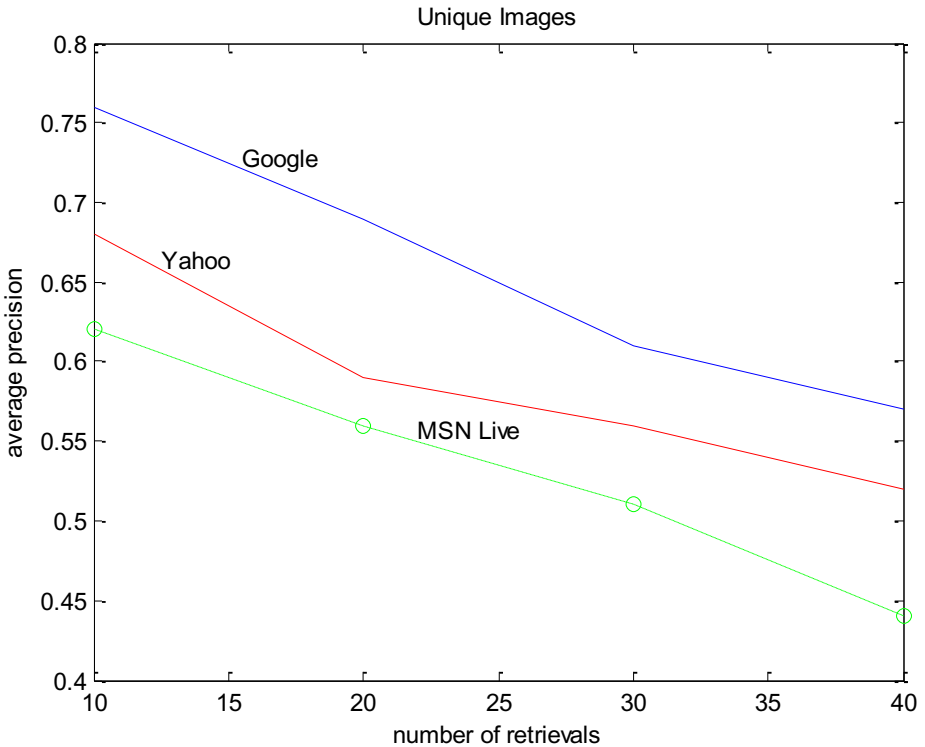
The performance of the search engines for various queries and retrieval levels is discussed in this section. The average precision of the retrieved images for Unique Images for different levels & search engines is tabulated in Table 2 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 3.

**Table 2.** Average Precision of retrieved images for Unique Images

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.76	0.69	0.61	0.57
Yahoo	0.68	0.59	0.56	0.52
Live	0.62	0.56	0.51	0.44

Figure 3 clearly illustrates that Google has the best average precision at any cut-off point for unique images, followed by Yahoo and MSN Live respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased.

The average precision of the retrieved images for Unique Images with refiners for different levels and search engines is tabulated in Table 3 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 4.



**Fig. 3.** Average Precision of retrieved images for Unique Images

**Table 3.** Average Precision of retrieved images for Unique Images with Refiners

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
<b>Google</b>	0.38	0.27	0.21	0.18
<b>Yahoo</b>	0.08	0.05	0.03	0.025
<b>Live</b>	0.2	0.14	0.09	0.07

For unique images with refiners, Google has the best average precision at any cut-off point, followed by MSN Live and Yahoo respectively. The precision has dropped off drastically as compared to the precision levels for unique images (without refiners).

The average precision of the retrieved images for Non-Unique Images for different levels & search engines is tabulated in Table 4 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 5.

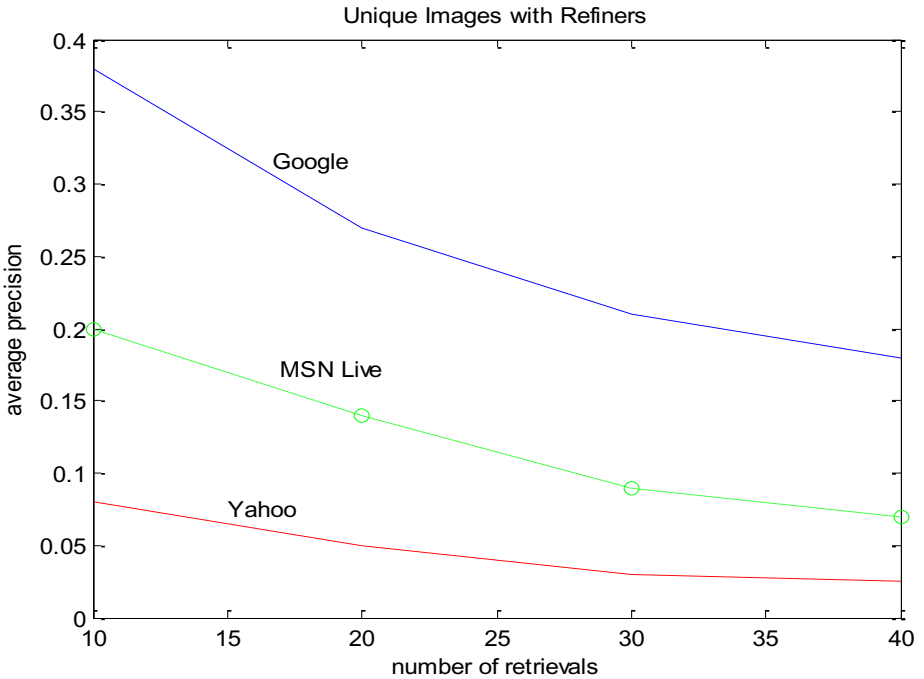


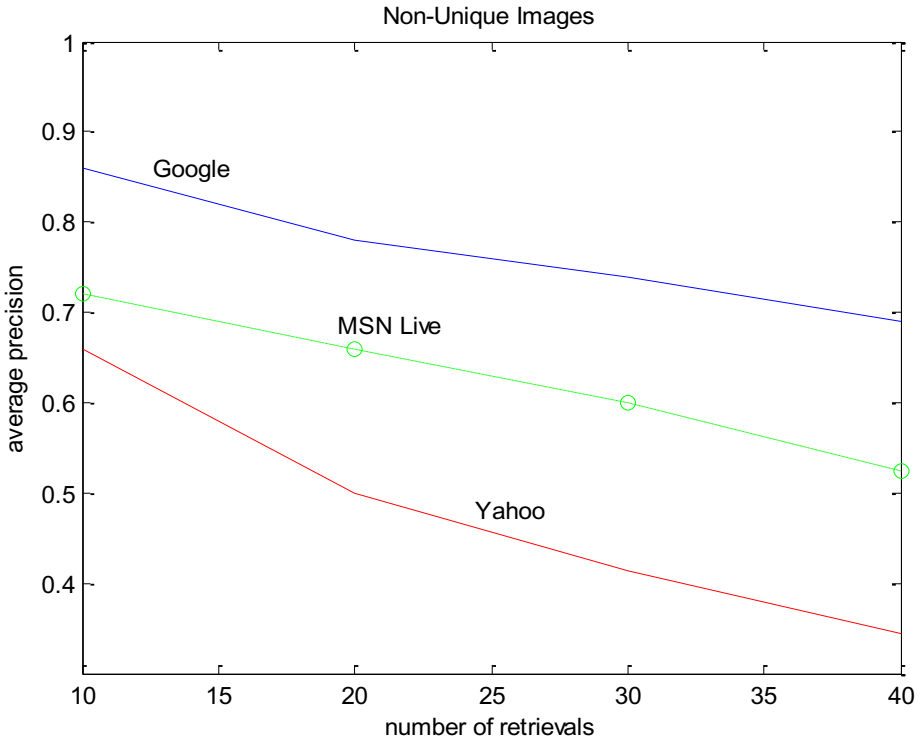
Fig. 4. Average Precision of retrieved images for Unique Images with Refiners

Table 4. Average Precision of retrieved images for Non-Unique Images

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.86	0.78	0.74	0.69
Yahoo	0.66	0.5	0.413	0.345
Live	0.72	0.66	0.6	0.525

Figure 5 illustrates that Google has the best average precision at any cut-off point for non-unique images, followed by MSN Live and Yahoo respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased.

The average precision of the retrieved images for Non-Unique Images with refiners for different levels & search engines is tabulated in Table 5 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 6.



**Fig. 5.** Average Precision of retrieved images for Non-Unique Images

**Table 5.** Average Precision of retrieved images for Non-Unique Images with Refiners

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
<b>Google</b>	0.42	0.41	0.39	0.355
<b>Yahoo</b>	0.24	0.17	0.113	0.085
<b>Live</b>	0.26	0.17	0.153	0.135

Tukey's Honest Significant Difference for Search Engines (Figure 7) clearly shows that Google Image Search Engine outperforms Yahoo and MSN Live. This analysis also indicates that the performance of Yahoo and MSN Live does not differ statistically. Tukey's Honest Significant Difference for Query Types (Figure 8) clearly shows that the performance of search engines is better whenever there is no additional refiner.

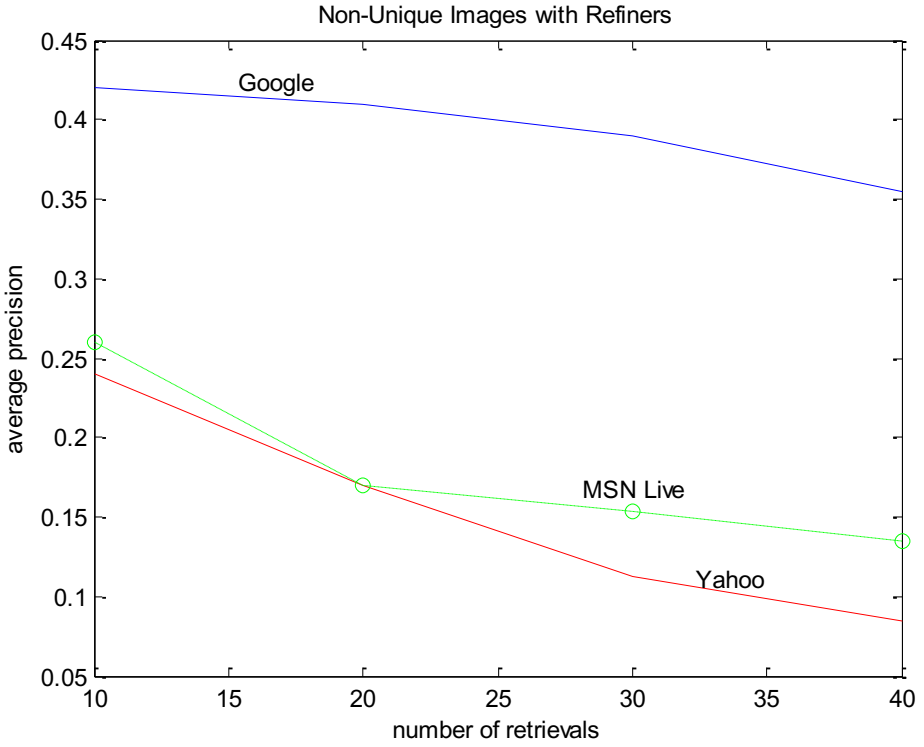


Fig. 6. Average Precision of retrieved images for Non-Unique Images with Refiners

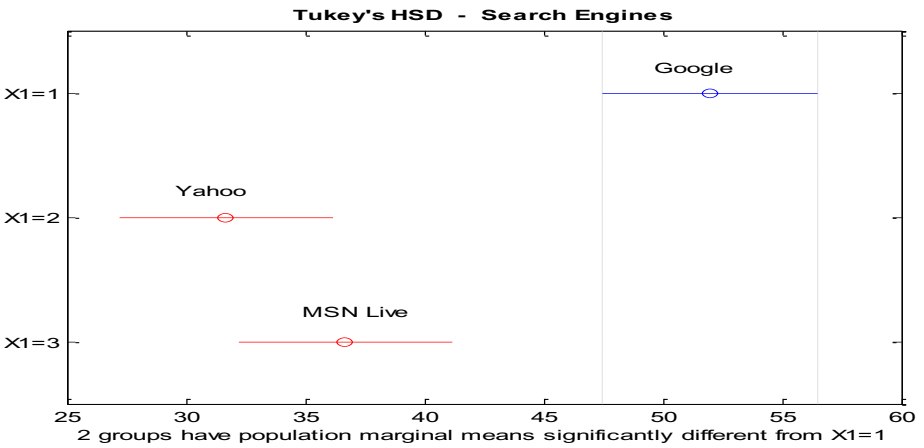
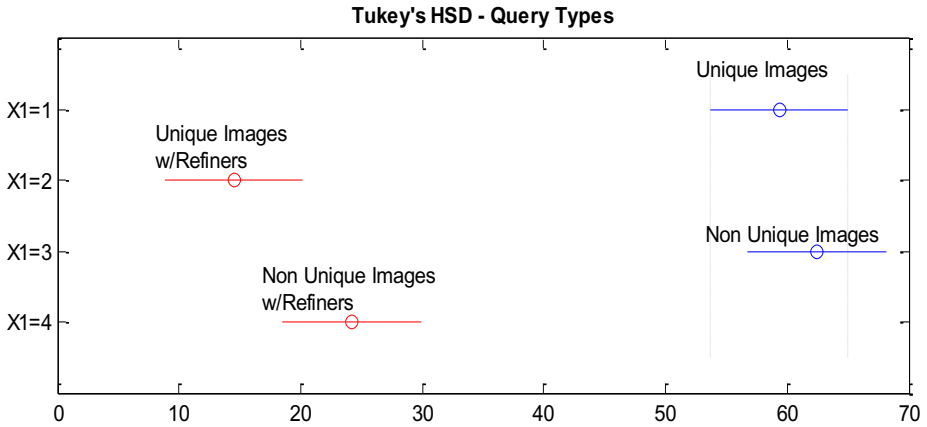


Fig. 7. Tukey's Honest Significant Difference for Search Engines



**Fig. 8.** Tukey's Honest Significant Difference for Query Types

## 4 Role of Humans in Image Annotation

### 4.1 Design of Experiment

**Participants:** For this study, eight participants were selected from a pool of candidates professing knowledge of the domain (the game of cricket). For the purposes of this experiment, a pre-annotation test was prepared to determine the level of expertise in the game of cricket. Some of the questions in the test are listed in Table 6. Those scoring higher than 90% were selected as the domain experts.

**Table 6.** Sample pre-annotation test questions

Questions	Multiple Choices
<b>Q. Which country won the Cricket World Cup in 1999</b>	Pakistan Srilanka Australia South Africa
<b>Q. Which of these is not a recognized fielding position</b>	Short Third Man Short Extra Cover Silly Point Long Twelfth Man

**Apparatus and Stimuli:** The participants viewed 40 images that were randomly selected from a database of 2,964 images taken over 25 years, all relating to the game of cricket. The domain experts were asked to label images on a personal computing system running at a minimum of 2.5GHz, Windows XP machine. A 17-inch LCD monitor was used to display the interface, with a mouse and keyboard used as the input devices. The experiment took place in an office type environment with ambient



lighting conditions. The participants sat in an adjustable office chair, and the mouse and keyboard were placed at a comfortable position as determined by each participant.

**Experimental Design:** This experiment was an 8x4 full factorial design that was run twice. There were two independent variables, including search database (DE) and query type (QUERY). The eight levels of DE were DE 1, DE 2, etc., and the levels of QUERY were UNIQUE, UNIQUE with IDENTIFIER, NON UNIQUE, NON UNIQUE with IDENTIFIER. Each level of QUERY was tested with each database.

Traditionally, evaluation for information retrieval has been based on the effectiveness ratios of precision (proportion of retrieved documents that are relevant) and recall (proportion of relevant documents that are retrieved) [19]. In this study, the two dependent variables that were examined, precision and recall are calculated as the knowledge of all relevant documents in the collection.

**Procedure:** The eight participants were asked to view 40 images and annotate each of these individually. The images were presented to the participants in random order and they were only allowed to see one image at a time. The participants were told that the annotations should be a descriptor of the image from a domain expert point of view and should be comprehensive. They were also told that annotations should be similar to the keywords that they use when they upload an image or a video online. The annotations were to be filled in the text box below the image. If a participant was unable to come up with a label for an image they were asked to fill "N/A" in the text area box. Finally they were told that the annotations will be used for information retrieval (image retrieval) purposes. After the annotation was completed, eight different search databases were built to test the queries.

These annotations were used to develop eight search databases against which the queries seen in Table 7 were run.

**Table 7.** Queries Run on the Databases

Query Type	Queries
Unique Images	1. Henry Olunga 2. Kevin Pietersen
Unique Images with Identifiers	1. Sachin Tendulkar + Marriage 2. Kapil Dev + World Cup
Non Unique Images	1. Pakistan Cricket Team 2. Indian Cricket Team
Non Unique Images with Identifiers	1. England Team + Ashes 2. Australian Team + World Cup

The search engine used to run the queries was Google Desktop Search Engine. This was chosen as it has been shown in previous studies to perform better than other desktop search engines on many difference measures [25].

Data were collected for the 32 experimental trials to test two hypotheses:

Hypothesis 1: There is no significant effect of the Search Engine and Query Type on the Precision of the retrievals.

Hypothesis 2: There is no significant effect of the Search Engine and Query Type on the Recall of the retrievals.

Figure 9 shows an example of the annotations of images for each query type.

<p style="text-align: center;"><b>Unique</b></p> 	<p style="text-align: center;"><b>Unique with Refiner</b></p> 
<p style="text-align: center;">The Young Ravi Shastri</p>	<p style="text-align: center;">Kapil Dev with 1983 World Cup</p>
<p style="text-align: center;"><b>Non-Unique</b></p> 	<p style="text-align: center;"><b>Non-Unique with Refiner</b></p> 
<p style="text-align: center;">Pakistan's team on the field</p>	<p style="text-align: center;">Ricky Ponting with Ashes cup</p>

**Fig. 9.** Example of an image and its annotation for each query type

## 4.2 Results

The ANOVA results were obtained for the full factorial design experiment conducted for precision (proportion of retrieved images that are relevant) and recall (proportion of relevant images that are retrieved with a search query). At a 99% confidence level, the ANOVA results show a significant main effect for QUERY ( $F_{(3,32)} = 6.4, p < 0.0001$ ) for precision and QUERY ( $F_{(3,32)} = 6.72, p < 0.0001$ ) for recall. The other main effect Search Engine (A) or the interaction effect is not significant at 99% confidence, for either precision or recall.

These results reject our Null Hypothesis and validate that as long as the person annotating the images in a database is a domain expert, the performance of the search engine did not change significantly. Clearly, the human's cognitive abilities have to be better tuned for the person to annotate the images in a more systematic way which may improve the performance of the search engines.

## 5 Conclusion and Future Work

The results of the benchmarking research suggest that overall, commercial search engines continue to have significant difficulties effectively executing image retrieval tasks. The Google search engine performs significantly better than Yahoo or MSN Live in any query type. The results also indicate that the precision of the search engines tended to drop with the increase in the number of retrievals. This performance reduction was noted across-the-board, that is irrespective of the search engines and the query types. The performance of the search engines also dropped dramatically when the queries had refiners (unique or non-unique).

The role of human annotation results also show that there is no significant difference in the performance of the search engines when a domain expert annotates the images. This result is important as it supports that we do not need a panel of annotators to label an image, as long as the annotator is proficient in the image domain. The results also indicate that the performance of the current commercial search engines can only be improved by a disciplined annotation approach by domain experts. With the number of images available on the internet growing exponentially, the human is incapable of annotating all these images in a systematic manner. Computer Vision algorithms can potentially be used to alleviate the load of the human operator for annotating the images. These algorithms can annotate images for content dependent metadata, but as they are still in their infancy, they fail when annotations requiring content descriptive metadata are required.

Clearly, the human's cognitive abilities have to be better tuned for the person to annotate the images in a more systematic way which may improve the performance of the search engines, i.e., there needs to be a more systematic way to annotate images to improve the performance of the search engines. The results of this study can be used to develop a semi-automated annotation system that provides a systematic template to improve upon the negative aspects associated with manual annotation. Computer vision algorithms can be used to fill the content dependent metadata in the template. The Human can then be brought into the loop to validate the metadata already present and can either delete or add any other metadata deemed pertinent for the image description within the template requirements. This template will incorporate human expertise to capitalize on the strengths of manual annotation that we have shown with using domain experts while avoiding the out-of-the-loop performance problems that occur when such a system is completely automated [26, 27].

## References

1. Kidambi, P., Narayanan, S.: A human computer integrated approach for content based image retrieval. In: Proceedings of the 12th WSEAS International Conference on Computers, Recent Advances in Computer Engineering, pp. 691–696 (2008)
2. Yates, B., Neto, R.: Modern Information Retrieval. ACM Press, New York (1999)
3. Witten, I.H., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing documents and images. Morgan Kaufmann Publishers, San Francisco (1999)
4. Kuralenok, I.E., Nekrestyanov, I.S.: Evaluation of Text Retrieval Systems. *Programming and Computer Software* 28(4), 226–242 (2002)
5. Text Retrieval Conference (TREC) National Institute of Standards and Technology (NIST) and S. Department of Defense (1992), <http://trec.nist.gov/>
6. Inoue, M.: On the need for annotation-based information retrieval. *Information Retrieval in Context*. In: SIGIR IRiX Workshop, pp. 44–49 (2004)
7. Choi, Y., Rasmussen, E.M.: Users' relevance criteria in image retrieval in American history. *Information Processing & Management* 38(5), 695–726 (2002)
8. Hughes, A., Wilkens, T., Wildemuth, B., Marchionini, G.: Text or pictures? An eyetracking study of how people view digital video surrogates. In: Proceedings of the International Conference on Image and Video Retrieval, pp. 271–280 (2003)
9. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research* 5, 913–939 (2004)
10. del Bimbo, A.: Visual Information Retrieval. Morgan Kaufmann, Los Altos (1999)
11. Liu, Y., Zhang, D., Lu, G., Ma, W.: A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40(1), 262–282 (2007)
12. Hyvönen, E., Styrman, A., Saarela, S.: Ontology-based image retrieval. In: Proceedings of XML Finland Conference, pp. 27–51 (2002)
13. Hanbury, A.: A survey of methods for image annotation. *Journal of Visual Languages & Computing* (19), 617–627 (2008)
14. Ahn, L.V., Dabbish, L.: Labeling images with a computer game. In: Proceedings of ACM CHI, pp. 319–326 (2004)
15. Hernon, et al.: Evaluation and Library Decision Making. Alex Publishing (1990)
16. Meadow, et al.: Text Information Retrieval Systems. Library and Information Science series. Elsevier publications, Amsterdam (1999)
17. Hersh, W.: Information Retrieval – A Health Care perspective. Springer publications, Heidelberg (1995)
18. Lancaster, et al.: Information Retrieval Today. Information Resource Press (1993)
19. Smith, J.R.: Image Retrieval Evaluation. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, vol. 21, pp. 112–113 (1998)
20. Cooper, W.S.: Expected Search Length – A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science* 19, 30–41 (1968)
21. Cakir, E., Bahceci, H., Bitirim, Y.: An Evaluation of Major Image Search Engines on Various Query Topics. In: The Third International Conference on Internet Monitoring and Protection, pp. 161–165. IEEE Computer Society, Los Alamitos (2008)
22. Broder, A.: A Taxonomy of web search. *SIGIR Forum* 36(2), 3–10 (2002)
23. Enser, P.G.B., McGregor, C.: Analysis of Visual Information Retrieval Queries. British Library Research, and Development Report 6104 (1993)

24. Nielsen Search Rankings (2009),  
[http://www.nielsen-online.com/pr/pr\\_090616.pdf](http://www.nielsen-online.com/pr/pr_090616.pdf)
25. Lu, et al.: Performance Evaluation of Desktop Search Engines. In: IEEE International Conference on Information Reuse and Integration, pp. 110–115 (2007)
26. Endsley, M., Kiris, E.: The out-of-the-loop performance problem and level of control in automation. *Human Factors* 37, 381–394 (1995)
27. Thackray, R., Touchtone, R.: Detection efficiency on an air-traffic control monitoring task with and without computer aiding. *Aviation Space and Environmental Medicine* 60, 744–748 (1989)

# Multimedia Streams Retrieval in Distributed Systems Using Learning Automata

Safiye Ghasemi<sup>1</sup> and Amir Masoud Rahmani<sup>2</sup>

<sup>1</sup> Department of Computer, Sepidan Branch, Islamic Azad University, Sepidan,  
Islamic Republic of Iran  
ghasemi.ss@gmail.com

<sup>2</sup> Islamic Azad University, Science and Research  
Branch, Tehran  
rahmani@srbiau.ac.ir

**Abstract.** Current academic Internet environment has enabled fast transfers of huge amounts of data, and has made high quality multimedia and collaborative applications a reality. This article describes a model for distributed multimedia retrieval which performs the retrieval of different multimedia to a variety of clients using learning automata algorithm named LAGridMSS. LAGridMSS allocates a proportion of bandwidth of each node for sending a specified file, and then applies learning automata to allocate packets of files to each node that contains the context. The files' popularity and the remainder bandwidth of nodes are two main factors here for allocating packets to each node.

Simulation results show improvements in proposed model, LAGridMSS, in some QoS factors such as delay, jitter, and reliability compared to previous multimedia retrieval system such as GridMSS and GridMedia.

**Keywords:** Grid Computing Environment, Streaming Files, Learning Automata, Quality of Service.

## 1 Introduction

The term “multimedia” has been used in many different contexts and means different things to different people [2]. In this issue, multimedia mainly relates the individual or combined use of large-volume, high-quality continuous digital media such as audio and video.

Grid computing has been evolving over recent years towards the use of distributed and parallel computing and also it provides dynamic, secure and coordinated sharing of heterogeneous resources that may be distributed geographically as well as organizationally [11]. One of the major challenges for grid computing is to ensure “non-trivial qualities of service” to the users. A grid computing platform is exploited and content adaptation is performed through appropriate agents that are allocated as jobs and executed on grid systems. These systems can be used for sharing resources of heterogeneous computers. In such a newborn environment, there are so much researching area, one of them is streaming applications which a poor work has been done on it so far.

Resources and information sharing in grid environment should be done with a desired QoS (Quality of Service), especially when sharing is in the field of time dependant applications. Quality of Service of a specific system describes how good that system will function with respect to a set of parameters specified by client. In this paper QoS is considered as a set of quality and quantity specifications of a distributed multimedia that is important for an application to satisfy the clients.

The client specifies the desired quality of media. An application's QoS requirements are conveyed in terms of high-level parameters that specify what the user requires. The specification of requirements of an application occurs according to client's parameters and is known as QoS parameters. After QoS specification, a system has to provide the demanded services according to accessible bandwidth. If it was not available the system accesses the related agents to negotiate the requirements identified by the client and may remove, change or adopt some parameters. As this negotiation successfully finishes then the streaming application will be done. Adding agents to multimedia systems causes the model to be modular and behave more flexible and adoptable in grid environment. Agent's changing role adds the system flexibility in replying different queries.

Many multimedia streaming algorithms in grid environment have been introduced, but none of them support all QoS parameters. Some of them use much more time for coordinating resource nodes. In the GridMSS framework [1] many of important QoS parameters such as reliability, delay, jitter, accessibility and flexibility are supported and the overall framework performance is high.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work about improving QoS of media streaming in grid systems. In Section 3, we analyze the concepts of learning automata. Section 4 presents the specification of existed problem and in Section 5 the proposed model for streaming applications is presented and we discuss in detail our experimental result in Section 6, while in Section 7 we conclude the paper.

## 2 Related Works

GridMedia [18] is a peer-to-peer based multicast architecture designed especially for large scale video streaming with quality requirement in terms of real time, low latency and bandwidth demanding. This architecture mainly contains the overlay protocol and the transmitting algorithm. The overlay protocol MSOMP advocates mesh-based two layer structure, which makes the upper layer much more robust than traditional tree-based scheme and ensures the demanding bandwidth.

A framework for QoS-based service discovery in manufacturing grid is described in [15]. The main focus of this framework is to provide a means for the service requesters to search for services based on QoS criteria in MG, to provide QoS guarantees for service execution and to enforce these guarantees by establishing SLAs.

Adrizona in [2] has presented an approach to multimedia content retrieval in a distributed digital library implemented on a P2P network. The proposed system employs an adaptive technique for routing queries and addresses the issues arising from the presence huge amounts of data, their peculiar nature and, finally, the lack of a centralized index.

Our approach employs a decentralized architecture which fully exploits the storage and computation capability of computers in the Internet and forwards queries throughout the network using an adaptive routing strategy that dynamically performs local topology adaptations.

In [5], the heterogeneous asynchronous multi-source streaming (HAMS) model for transmitting continuous multimedia files from multiple peers to a leaf peer is discussed. Goudarzi[6] discusses how to support a receiver peer with enough QoS of the multimedia steaming service by multiple source peers in [6]. Not only a receiver peer but also source peer is moving in a network. Here, QoS supported by source peer is changing according to the movement of the receiver peer and source peer.

GridMSS [1] is a streaming contents retrieval framework which supports the QoS. The main characteristics of this framework are highlighted by its promising QoS supporting such as accessibility, low latency and reliability, its flexibility and packet allocation mechanisms of nodes. The approach is that the allocation is done according to available bandwidth of each node. GridMSS provides a high performance files streaming framework for a large population of users and ensures the QoS in terms of accessibility, low latency and reliability. To tide over the bandwidth bottleneck and the rate of losing packets some new relations between bitrates of nodes and services they can provide are determined and in the extended GridMSS a genetic algorithm is used to find the best nodes with suitable bit rate for sending demanded content to receiver.

CoopNet (Distributing Streaming Media Content Using Cooperative Networking) [20] divides streaming media content into multiple sub-streams using MDC and each sub-stream is delivered to the requesting client via a different peer. This improves robustness and also helps balance load amongst peers. The system has two cases. First is live streaming refers to the synchronized distribution of streaming media content to one or more clients, and second one is on-demand streaming, which refers to the distribution of pre-recorded streaming media content on demand (e.g., when a user clicks on the corresponding link).

Another streaming system is a QoS-Aware P2P streaming framework called Whirlpool which organizes peers into different levels based on their end-to-end latencies from the streaming source. The key design issues of Whirlpool can be mentioned as presenting a membership service, which selects potential good neighbors with higher probabilities, and analyzing the convergence property of the level calculation the key design issues of Whirlpool, presented a membership service, which selects potential good neighbors with higher probabilities, and analyzed the convergence property of the level calculation algorithm.

### 3 Learning Automata

Learning Automata (LA) are adaptive decision making units that can learn to choose the optimal action from a set of actions by interaction with an unknown random environment. At each instant,  $n$ , the LA chooses an action  $\alpha_n$  from its action probability distribution and applies it to the random environment. The random environment provides a stochastic response, which is called a reinforcement signal to the LA. Then the LA uses the reinforcement signal and learning algorithm to update the action probability distribution.



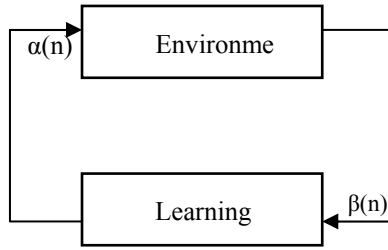


Fig. 1. Learning automata connection with environment

A learning automaton is an automaton that improves its performance by interacting with the random environment in which it operates. Its goal is to find among a set of  $\mu$  actions the optimal one, so that the average penalty received by the environment is minimized. This means that there exists a feedback mechanism that notifies the automaton about the environment’s response to a specific action.

The operation of a learning automaton constitutes a sequence of cycles that eventually lead to minimization of average penalty. The learning automaton uses a vector  $p(n) = \{p_1(n), p_2(n), \dots, p_\mu(n)\}$ , which represents the probability distribution for choosing one of the actions  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ . The core of the operation of the learning automaton is the probability updating algorithm, also known as the reinforcement scheme, which uses the environmental response  $\beta(n)$  triggered by the action  $\alpha_i$ , selected at cycle to update the probability distribution vector . After the updating is finished, the automaton selects the action to perform at cycle  $n+1$ , according to the updated probability distribution vector  $p(n+1)$ . A general reinforcement scheme has the form of equations (1, 2).

$$\begin{aligned}
 p_i(n+1) &= p_i(n) + a \cdot (1 - p_i(n)) \\
 p_j(n+1) &= p_j(n) - a \cdot p_j(n) \quad \forall j \ j \neq i
 \end{aligned}
 \tag{1}$$

$$\begin{aligned}
 p_i(n+1) &= (1 - b) \cdot p_i(n) \\
 p_j(n+1) &= \frac{b}{r - 1} + (1 - b) \cdot p_j(n) \quad \forall j \ j \neq i
 \end{aligned}
 \tag{2}$$

In these two equations,  $a$  and  $b$  are reward and penalty parameters respectively. For  $a=b$ , learning algorithm is called  $L_{R-P}^1$ , for  $a < b$ , it is called  $L_{R\epsilon P}^2$ , and for  $b=0$ , it is called  $L_{R-I}^3$ .

<sup>1</sup> Linear Reward-Penalty.

<sup>2</sup> Linear Reward epsilon Penalty.

<sup>3</sup> Linear Reward Inaction.

## 4 Problem Specification

There are a variety of nodes with different files in grid computing. Care is taken that some of the files may exist in several nodes. It means that the number of content all over the grid system is greater than one. In this research we suppose that the frequency of files in grid computing have been defined. Furthermore the nodes supposed to be on line during the process of sending. The considered system has two parts. First is the user part which gets requests of users, second one is information part which has the information of current active nodes (e.g. list of contained files and remainder bandwidth).

The process starts by getting requests of clients, and then these requests are sent to information part. After receiving requests, the information part allocates a quote of bandwidth to each file in each node. The allocation procedure is done according to remain bandwidth of nodes and the popularity of files which nodes contain.

As mentioned before, each node should send some parts of files to the buffer of receiver nodes. Actually nodes send some part of each file according to the quotas assigned to them. In the buffers after arriving enough parts user can use them.

Files are comprised of numbers of packets. Size of packets can be the same or not. In this paper we assumed different size of packets. But all copies of files have the same number and size of packets, therefore allocation can be done centrally, and then the results are sent to nodes to transfer the packets which they have to send.

## 5 Proposed Model: LAGridMSS

Consider a scenario which clients send their requests to the user part of system. Each request contains the name and some special specification of file such as bit rate. After gathering requests, the user part sends them to the information part. This part calculates the popularity of each file according to requests.  $Req$  is an array which consists of requests of the clients. It contains name of files and the bit rate which the user demanded before. Clients prefer to get the file with specified bit rate.

To find the popularity of a file such as  $i$  we used equation (3).

$$P_i = \frac{NReq_i}{NoF_i} \quad (3)$$

Where  $NReq_i$  shows the number of all requests for receiving content with the label of  $i$  in  $Req$ , and  $NoF_i$  is related to frequency of the content with label  $i$  in the system. To make it more clear this parameter is the number of content  $i$  in all over the considered grid system.

The user part which receives clients' requests can wait for a specified duration of time and then sends requests to information part or it can wait until a specified amount of requests. Selecting one of these situations depends on to the distribution of requests and to the number of requests that are gotten in a specified duration of time.

The bandwidth for sending files to receivers' buffer can be determined according to the popularity of files in each node. We used the relation shown in equation (4) for specifying a quote of remainder bandwidth of nodes to each of files it belongs.

$$BW_j^i = \frac{P_i}{N_j} * \beta_j \quad (4)$$

$$\sum_{k=1} P_k$$

Where  $BW_j^i$  shows the bandwidth which is assigned to file  $i$ , in node  $j$ ,  $N_j$  is the number of files in node  $j$  and  $\beta_j$  is remainder bandwidth of node  $j$ .

After this phase, the bandwidth of each node for sending any of included files will be determined. In other words the bit rate for sending requested files in each of nodes is calculated. Take care that popularity of files which no requests have received to get them is zero according to the equation (3). Therefore no bandwidth will be assigned to these files.

The files are comprised of some packets with different sizes. As mentioned before the files are distributed in the grid computing. It means that content can be in a variety of nodes in the system. According to our proposed approach, some simplification assumptions have been considered. We have assumed that each content has the same number and size of packets all over the system. This assumption eases the allocation of packets to nodes of grid system.

The packets in each content are shown by  $pkt$ . Therefore a streaming content can be considered as streaming content =  $\langle pkt_1, pkt_2, \dots, pkt_l \rangle$ .  $l$  is the number of packets in the content. Care is taken so that  $l$  has the same value for a specified content in all of the nodes which include that content.

Suppose that  $m$  requests have been received for getting  $n$  files. These requests are arranged in an array by user part of the considered system.

When the bandwidth allocated to demanded files in each node is determined, the approach should assign packets of files to each node to be sent. This phase can be performed randomly. Therefore the assignment will not be the optimal, but if a learning method is used the results may be optimal. Here we have used learning automata for selecting a node to send each packet of files.

As mentioned before, size of packets are not the same and there are different files in each node which may be requested. In previous phase the bandwidth for sending of each file in nodes was calculated. The allocated bandwidth for sending a file has different values in nodes which contains the file. According to this fact some packets are better to be sent by some special nodes as the assigned bandwidth of nodes for sending each file is different. We used learning automata to select the best nodes for sending each packet of content. The number of automata used for each file is related to the size of buffer in receiver node. We used automata as much as number of packets which fill the buffer.

According to previous parts of the paper, each automaton has a set of actions. Here actions of them are the nodes which have the requested content. In fact, a LA selects one of the nodes to send the associated packet to the buffer. As mentioned before the selected action should be evaluated according to the input from environment. For evaluating each action the equation (5) is used.

$$B_j^i(n) = \frac{T_{Ideal}^i}{T_j^i}$$

$$T_j^i = \frac{sizeof(pkt_i)}{BW_j^i} + delay(node_j) \quad (5)$$

$$T_{Ideal}^i = \frac{sizeof(Pkt_i)}{Req\_BR(File_c)}$$

Where  $B_j^i(n)$  related to evaluation factor of the LA assigned to packet  $i$ , in its  $n$ 'th repetition when node  $j$  is selected as the sender of  $pkt_i$ . This factor is related to  $T_{Ideal}^i$  and  $T_j^i$ . The first is the time needed for sending  $pkt_i$  by the requested bit rate which is known as ideal transfer time, and the former one is the time which is required for sending  $pkt_i$  when is transferred by node  $j$ . As can be concluded from these relations,  $B_j^i(n)$  has a value less or equal to one.

Each LA selects a node for sending its packet. In the first run of LA probability of all nodes are equal. Then the selected action is evaluated and according to the evaluation results, probability of selected node will be changed. The probability may be decreased or increased, and so the probabilities of other nodes will be increased or decreased.

The automata model will be iterated until the probability of a chosen action in each automaton exceeds 95% [2], or transfer time does not change for much iteration, or the number of iterations reaches a maximum limit. If first or second condition stop the automata, then we say that the model coverages.

## 6 Simulation

For simulating the mentioned approach, a set of predefined nodes with their specifications as the included files and properties, the remainder of their bandwidth and some requests to the system were used. The following section is about these simulations in detail.

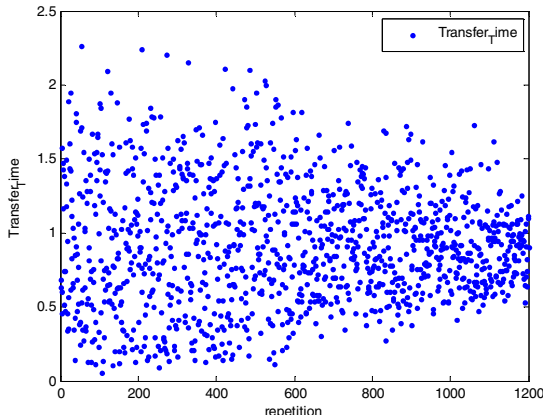
### 6.1 Simulation Environment Setup

We considered a grid system with predefined number of nodes and we used some assumptions in the model, such as no variability of nodes while sending process. Every node was assumed to have some files and each stream is initially set to have some different sized packets in the interval [20, ..., Maxpacket], with Maxpacket less or equal to 10000.

Number of nodes, the files in nodes, all requests of the system, frequency of files in grid system and the properties of demanded files are produced randomly. Transfer rate of each node is produced by uniform distribution with average of [10<sup>2</sup> kb/s, 10<sup>6</sup> kb/s].

To accurate the results of our simulation, we test it by using a variety of parameters of learning automata algorithm. These parameters are reward parameter and penalty parameter. To illustrate some results of the simulation, we have given maxpackets =

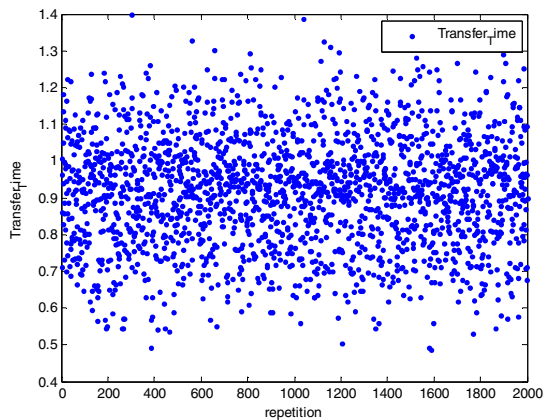
10000, reward parameter is randomly generated between 0.2 and 0.7 and penalty parameter is varied from 0.05 to 0.3. According to the results of simulation the best observed value of reward parameter can be 0.6 and for the penalty one is 0.1 (fig. 2).



**Fig. 2.** Convergence of learning automata algorithm with reward parameter of 0.6 and reward parameter of 0.1

In the following depicted figures the horizontal ax is the number of LA algorithm iterations before convergence. Care is taken that each repetition in ax can be a composition of several repetitions. As mentioned before the termination condition of automata as no change in the value of transfer time can be claimed as convergence of the algorithm. The vertical ax shows the average time needed for transferring demanded files to buffers.

In fig. 3, we suppose reward parameter with the value of 0.2 and the penalty as 0.3. The automata need more iteration to be converged in comparison with the previous run in fig. 2. In fig. 2 the convergence occurred much earlier than fig. 3.



**Fig. 3.** Automata convergence with reward parameter of 0.2 and reward parameter of 0.3

## 6.2 Detection Simulation Result

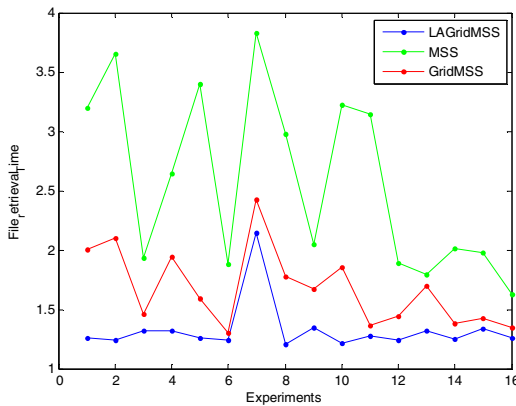
We performed our experiments with an event-driven simulator coded in C. The simulator models the grid environment, the nodes, communication channels and files in the system. After determining the accurate values of required parameters of proposed approach, we compare it with some other previous methods of streaming files. One of these algorithms is MSS which is introduced in previous sections. Another one is GridMSS which is the authors' previous investigation results on streams retrieval in grid system.

In the obtained results of our simulations, we use LAGridMSS for the proposed approach in this research. Table 1 shows the parameters of simulated experiments.

**Table 1.** Main parameters used in some simulated experiments

Experiment	1	2	3	4	5
Number of nodes	300	550	800	1250	1800
Average number of files in each node	20	25	30	30	30
Average bandwidth of each link(kbps)	40	40	50	60	60

In fig. 4 the comparison results of these methods is depicted. The vertical ax is the duration of time needed for transferri`ng the demanded files by entered requests to the system and the horizontal one shows the number of runs with some variations in specification of nodes or requested files.



**Fig. 4.** Retrieval time needed in MSS, GridMSS and LAGridMSS

In this comparison depicted in fig. 4, we run the simulation 16 times for different values of number of nodes, size and distribution of files, number of requests and bit rates. The performance of LA based approach was much better than others.

## 7 Conclusion

This research proposed an intelligent streams retrieval method in grid environments. In first phase the appropriate bandwidth for each file, in nodes is considered using the popularity of files in grid system. There is a LA for each packet of files. These LA selects a node for transmitting the packet and then tries to improve its next selection. Transfer time of packets in comparison to ideal transfer time is the evaluation factor of selections. According to the results of evaluations the probability for selecting each node is changed. Finally, when the automata converge, the probabilities of nodes are calculated. Simulations of the model showed the decreased delay in retrieving requested files in comparison to existing models such as MSS and GridMSS.

## References

1. Ghasemi, S., Rahmani, A.M., Mohsenzade, M.: An agent-based framework for supporting QoS in grid computing environment. In: ICCEE, 2008, Thialand. IEEE, Los Alamitos (2008)
2. Ardizzone, E., Gatani, L., Cascia, M.L., Re, G.L., Ortolani, M.: Distributed Multimedia Digital Libraries on Peer-to-Peer Networks. In: 14th International Conference of Image Analysis and Processing. IEEE, Los Alamitos (2007)
3. Itaya, S., Hayashibara, N., Enokido, T., Takizawa, M.: Scalable Peer-to-Peer Multimedia Streaming Model in Heterogeneous Networks. In: Proceedings of the Seventh IEEE International Symposium on Multimedia, IEEE, Los Alamitos (2005)
4. Itaya, S., Hayashibara, N., Enokido, T., Takizawa, M.: Distributed Multimedia Streaming Systems in Peer-to-Peer Overlay Networks. In: Proceedings of the Eighth IEEE International Symposium on Multimedia, IEEE, Los Alamitos (2006)
5. Itaya, S., Enokido, T., Takizawa, M., Yamada, A.: A Scalable Multimedia Streaming Model Based-on Multi-source Streaming Concept. In: Proceedings of the 2005 11th International Conference on Parallel and Distributed Systems. IEEE, Los Alamitos (2005)
6. Nemat, A.G., Takizawa, M.: Application Level QoS in Multimedia Peer-to-Peer (P2P) Networks. In: WAINA, pp. 319–324. IEEE, Los Alamitos (2008)
7. Itaya, S., Hayashibara, N., Enokido, T., Takizawa, M.: Distributed Coordination Protocols to Realize Scalable Multimedia Streaming in Peer-to-Peer Overlay Networks. In: Proceedings of the 2006 International Conference on Parallel Processing, IEEE, Los Alamitos (2006)
8. Itaya, S., Hayashibara, N., Enokido, T., Takizawa, M.: Distributed Coordination for Scalable Multimedia Streaming Model. In: Proceedings of the 26th IEEE International Conference on Distributed Computing Systems Workshops. IEEE, Los Alamitos (2006)
9. Itaya, S., Hayashibara, N., Enokido, T., Takizawa, M.: Distributed Coordination Protocols to Realize Scalable Multimedia Streaming in Peer-to-Peer Overlay Networks. In: Proceedings of the International Conference on Parallel Processing, IEEE, Los Alamitos (2006)
10. Bertino, E., Catarci, T., Elmagarmid, A.K.: Quality of Service Specification in Video Databases. IEEE Computer Society, Los Alamitos (2003)
11. Tang, J.: An Agent-based Peer-to-Peer Grid Computing Architecture, A Thesis Submitted in Fulfillment of the Master, University of Wollongong (October 2005)
12. Li, S.: Quality of service control for distributed multimedia systems, A Thesis Submitted to the Faculty of Purdue University (1997)

13. Lu, S., Lyu, M.R.: Constructing Robust and Resilient Framework for Cooperative Video Streaming. In: ICME Confrance. IEEE, Los Alamitos (2006)
14. Amoretti, M., Conte, G., Reggiani, M., Zanichelli, F.: Designing Grid Services for Multimedia Streaming in an E-learning Environment. In: Proceedings of the 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises. IEEE, Los Alamitos (2004)
15. Shi, Z., Yu, T., Liu, L.: MG-QoS: QoS-Based Resource Discovery in Manufacturing Grid. Springer, Heidelberg (2004)
16. Guedes, L.A., Oliveira, P.C.: QoS Agency: An Agent-based Architecture for Supporting Quality of Service in Distributed Multimedia Systems. In: Proceedings of the IEEE Conference on Protocols for Multimedia Systems - Multimedia Networking, IEEE, Los Alamitos (1997)
17. Kerherve, B., et al.: On Distributed Multimedia Presentational Applications: Functional and Computational Architecture and QOS Negotiation. In: Proc. 4th Int. Workshop on Protocols for High-speed Networks, pp. 1–17. Chapman & Hall, Boca Raton (1994)
18. Zhang, M., et al.: Gridmedia: A Multi-Sender Based Peer-to-Peer Multicast System for Video Streaming. IEEE, Los Alamitos (2005)
19. Mine, T., Matsuno, D., Takaki, K., Amamiya, M.: Agent Community based Peer-to-Peer Information Retrieval. In: AAMAS 2004, Springer, New York, USA (2004)
20. Padmanabhan, V.N., Wang, H.J., Chou, P.A.: Distributing Streaming Media Content Using Cooperative Networking. In: NOSSDAV 2002, Miami, Florida, USA, May12-14 (2002)



# Real-Time Detection of Parked Vehicles from Multiple Image Streams

Kok-Leong Ong<sup>1</sup> and Vincent C.S. Lee<sup>2</sup>

<sup>1</sup> School of Information Technology, Deakin University  
Victoria 3125, Australia  
`kok-leong.ong@deakin.edu.au`

<sup>2</sup> Faculty of Information Technology, Monash University  
Victoria 3168, Australia  
`vincent.cs.lee@monash.edu`

**Abstract.** We present a system to detect parked vehicles in a typical commercial parking complex using multiple streams of images captured through IP connected devices. Compared to traditional object detection techniques and machine learning methods, our approach is significantly faster in detection speed in the presence of multiple image streams. It is also capable of comparable accuracy when put to test against existing methods. And this is achieved without the need to train the system that machine learning methods require. Our approach uses a combination of psychological insights obtained from human detection and an algorithm replicating the outcomes of a SVM learner but without the noise that compromises accuracy in the normal learning process. The result is faster detection with comparable accuracy. Our experiments on images captured from a local test site shows very promising results for an implementation that is not only effective and low cost but also opens doors to new parking applications when combined with other technologies.

## 1 Introduction

The motivation behind the work in this paper is the desire for a parking system that aims to reduce frustration for drivers in their attempt to hunt for a free parking lot. Especially under heavily utilised conditions, navigating a parking site and competing with other drivers for a free spot is often a time consuming and frustrating task. Current advanced parking systems at various sites in Australia implements a “sensor-to-lot” approach with signages near the site to assist drivers. Although such an implementation provided assistance to drivers, there are many drawbacks.

By using a sensor for each parking lot, a large parking site becomes costly to implement when the costs of fitting sensors and wiring them to the signage is considered. Consequently, the implementation is kept simple to contain the costs. As a result, the implementation failed to take advantage of the collective information provided by the sensors. In situations where the site is heavily utilised, drivers quickly face frustrations because (i) signage information become

inaccurate; (ii) sensor lights (that indicated free lots) become difficult to spot; and (iii) the effectiveness of light indicators are limited to a small range due to the “line of sight” approach. This a “local optimal” solution since drivers in a busy parking site can only depend on available information in the vicinity rather than the collective information provided by the sensors.

In search for a better car park system than the commonly used “sensor-to-lot” approach, we discovered that many research do not address the problem of informing drivers about free parking lots, and using that information effectively to reduce the frustration of drivers. A different solution is thus called for that started this investigation. As smart phones connected to the Internet via 3G networks become ubiquitous, we foresee that they may present the answer to ease, if not, end a driver’s car park hunting nightmare.

Our premise is that if drivers are informed in advanced about the situation of a parking site, it will enable decisions to be made to avoid the frustrations of not been able to secure a free parking lot. Extending this idea, it would become possible to use the technology in ways such as enabling guidance to parking lots on a large parking site, directing drivers to alternate parking sites under busy situations, and so on.

For such a system, the sensor in each parking lot needs to be wired to a server so that they can be mashed up with information on the Web to create the applications we envisioned. Doing so however will significantly increase infrastructure costs. The solution is to replace multiple sensors with a single IP-enabled camera. By reducing the number of input points, we lower costs but now require a method to detect the presence of a parked car. Our proposal is novel in terms of marrying image processing technologies and machine learning concepts to deliver a cost effective and accurate solution which we discuss in Section 3 along with experimental results in Section 4. We will also present existing works next, and our future work in Section 5.

## 2 Related Works

Our work comprises of two areas: (i) the design of smart car parks and (ii) the detection of free parking lots in images. On the design of car park systems, which is not the main discussion point of this paper but relevant, our survey revealed a focus on a number of key areas. Many address problems in aspects of parking such as smart payment systems [1,6,14], transit-based information [2,16], automated parking [1,11]. In these areas, the problems and objectives addressed are different from our motivation.

Two areas of car park design research are however of interest to us. The first is e-Parking systems such as those reported by [4,5] and parking guidance and information systems (PGIS) represented by the works of [9,8,12,21]. e-Parking systems focus on the use of the Internet to create a smart booking environment to inform drivers of available parking sites and free parking spaces. PGIS on the other hand, aim to provide guidance to help drivers locate a park. Best represented by information signages in the vicinity and vehicle detection sensors



**Fig. 1.** (a) A reference image where the parking lot is empty. The same set of filter is applied to the reference image as well as the incoming image stream represented by (b). For images streams where the vehicle colour is light (as in (b)), it is rather easy to obtain a high detection accuracy.

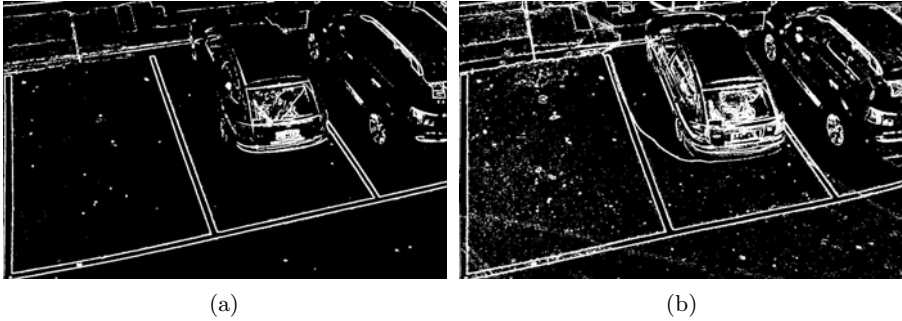
in each lot, they are the “sensor-to-lot” systems where we are keen to overcome its limitations discussed earlier and hence, the motivation of the work reported in this paper.

On the issue of detection, one approach is the use of machine learning techniques, where labeled images are used to train a classifier that will be deployed to detect the presence of a car. Here, different classifier technologies, methods of training and the structure of classifiers were explored. For example, [19] proposed a 8-class SVM classifier with probabilistic outputs while in [18], a simple (linear hyperplane) classifier was used to achieve above 90% accuracy by optimising the information of individual features in an image. Others used image processing techniques to achieve similar results. Interpretation of image sequences using visual surveillance techniques was proposed by [3] while [7] and [10] tracks movement of vehicles as the basis for determining free parking lots.

For [18], the major drawback is in the scalability of its solution. For an 8-class SVM classifier, the system is only capable of dealing with 3 parking lots in a single image. If a single camera captures 4 parking lots, a 16-class SVM classifier is needed. As the effort and computation requirements double with every additional parking lot, the solution’s practical significance is limited. In the area of image processing techniques such as [3,10,7], the detection mechanism requires incurs either high computational costs or large memory space. In comparison, our approach is far more scalable than the proposal in [18] and requires less computing resources than those proposed in [3,10,7].

### 3 Our Approach

In the context of image processing, which this system now depends on, the problem is a classic case of object detection [13,17]. The challenges of object detection are the high variability in appearances of objects in a given class (in our



**Fig. 2.** The only difference in the two images is the position of the sun. As a result, the intensity level affected the filters’ output as seen in (a) and (b) where edge detection and binarisation filters are applied using default parameter values. Clearly this impacted detection accuracy, especially in the case of the first parking lot. In our algorithm, we used a simple statistical method to adjust the filter’s parameters on-the-fly so that it can be compared to the reference image accurately.

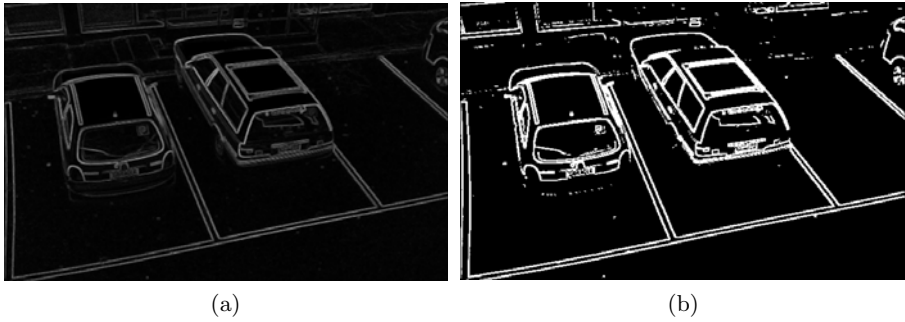
case, parked cars) and the added variability between instances of the same object due to alternate viewing angles and/or conditions (e.g., the same car viewed from the front, side, or back). As images are taken as specific time intervals from different cameras, we now have multiple image streams to be processed. The problem thus calls for real-time efficiency and accuracy.

To achieve this, we first address the question of accurate detection. As we learnt in Section 2, a popular approach is to use a classifier to distinguish between an image belonging to a target class (i.e., car present in lot) and one that doesn’t (i.e., car *not* present in lot). Usually a set of vectors, with each representing an image, is used in the training of classifiers to find discriminating features that separates two or more classes. In the case of the SVM [15], well-known for its binary classification accuracy under small data samples and high dimensionality, the set of discriminating features are identified by the hyperplane.

Simply put, the concept of a hyperplane is a cut that best separates the feature spaces into two distinct classes. In SVM, this cut is determined via the learning process using a series of vectors and its associated class label. The issue with this process is the dependency on the learning algorithm, which itself is dependent on the data, to find the best cut that defines the hyperplane. To increase accuracy, much of the work focuses on the training data either by stripping the vector down to key feature spaces and/or increasing learning instances. In any case, the idea is to reduce the noise in the training instances to allow a “cleaner cut” and hence, a better classification accuracy. We were however inspired by a different approach.

### 3.1 Recreating the Ideal “Hyperplane”

We asked if we can define the hyperplane directly. If we can do so, we will be able to eliminate the noise from the training instances giving rise to a significant



**Fig. 3.** Final image used to detect presence or absence of car: (a) before applying the binarisation filter; (b) after applying the binarisation filter, which improves accuracy. Clearly, the reason for the improved accuracy is the wider margin between the two intensities after the binarisation filter.

increase in accuracy. In pursuit of our ideal hyperplane, we begin by learning how the most accurate classification machine, i.e., the human subject, determines the presence of a free parking lot. Since we are no psychologist, we turned to existing literature for some guidance. Fortunately for us, Zhao and Nevatia [20] reported such an experiment with some useful findings. In the test they conducted, the factors most people mentioned about knowing the presence of a car are **(i)** their rather rectangular shapes; **(ii)** the visibility of front and rear windshields; **(iii)** evidence of a parking lot; and **(iv)** environmental conditions such as shadows or light.

During classification, these factors are the discriminating feature spaces to be used for detecting the presence of a car in an image. And in the specific case of the SVM, they will be the hyperplane we are seeking when feeding the learning instances to the SVM learner. While conceptually this is easy to explain, trying to implement this within the SVM isn't as straightforward. After all, the algorithm was designed to learn about the cut rather than to be told of the cut. While it is possible to process images of the noise to get close to the ideal hyperplane, it is not possible to automate this under multiple streams of images. This led us to consider an alternative.

In our opinion, these factors are clearly the key feature spaces to use in determining the presence of a car when given an image. In other words, the human subject would filtered other information focusing on the key features to arrive at the conclusion. In lingo of SVM, this would be the hyperplane we seek when feeding the learning instances to the SVM learner. While conceptually this is easy to explain, trying to implement this within the SVM isn't as straightforward. After all, the SVM was designed to *learn* about the discriminating features rather than been told what they are.

The immediate and apparent solution is to produce learning instances containing only these feature spaces. Instead of going with this option, we toyed with an alternative approach: *why not explicitly code the classifier instead of*

*feeding our psychological observations into the SVM learner?* Clearly, the benefit of doing so is performance, i.e., a custom classifier exploiting the psychological observations will result in real-time processing capabilities that the application needs. The idea of an accurate and fast classifier is very attractive to us. Hence, our decision to implement these findings using image processing techniques.

We first convert the colour images into 8-bit grey scale images allowing each pixel to be represented exactly in a byte for the ease of implementation. For now, we restrict ourselves to the analysis of a single parking lot taken in an image. Our solution will easily and directly scale to multiple parking lots. With the first factor been the shape, our intuition is to begin by applying an edge filter on the image. As shown in Figure 1(b) and Figure 2(a), the edge filter strips the noise in an image by dropping texture and tonal details but leaving behind structural properties to allow an object to be determined.

The next factor is crucial to the speed and accuracy of our proposal. The experiments at our test site revealed that the key determinant of an unavailable park lie in the visibility of the windshields. The windshields are glass surfaces that reflect light giving off a higher intensity within the parking lot relative to its environment. Even in dim multi-story parks, this remained the case even after the image was stripped off its details by the edge detection filter. Our algorithm uses this key observation, which will be discussed in Section 3.2.

The third human consideration is to look for evidence of a parking lot. In our case, this factor is built into the algorithm as we deal with a per parking lot basis during detection. For an image taken, we will predefine the boundary coordinates for each parking space in the image. In fact, the boundary coordinates defined an area smaller than the parking lot. In our experiments, we find that this gave rise to better efficiency *and* accuracy when the area is concentrate around the spot(s) where the windshields, i.e., factor (ii), are likely to appear.

The last human observation is by far the most challenging. Car colour and size, and varying lighting (or weather) conditions can cause false positives (or negatives) in the detection outcomes. For a dark car, there may be insufficient light from the windshield to conclude the presence of a car (i.e., false negatives). Likewise, a small car will give bigger variation in terms of where they can park within a lot space. And with windshields a key determinant in our algorithm, over variation in the position of the windshield will increase false negatives. Interestingly, dealing with lighting conditions was much easier than dealing with car colour and size. The issue with lighting conditions is mainly constrained to open parking spaces with natural lighting. As weather conditions (e.g., position of the sun) vary, the detection accuracy also fluctuates.

On weather conditions, we found rain to be an issue when our test camera was not properly sheltered. This caused significant problems when we applied the edge filter leading to false positives. This was easily overcome by mounting a shelter on the camera. Unlike fixed light sources in sheltered parking lots, the movement of sunlight throws varying light intensity (i.e., shadows) on the same parking lot resulting in both false positives and negatives. As Figure 2 showed, the movement of sunlight caused presence of noise when passing images through

the edge detection filter. Our approach is to use multiple reference images to compensate the varying levels of light due to the sun’s movement. Instead of a single reference image (such as Figure 1(a) taken at time  $t$ ), we used an array of reference images taken throughout the day to allow variation in the threshold thereby minimising the errors. With this intuition, we discuss the algorithm in the next section.

### 3.2 Algorithm

Let  $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$  be the set of cameras in a car park facility. For any camera  $c_i$ , we define a tuple  $\langle c_i, \mathcal{P}_i = \{p_1, p_2, \dots, p_k\} \rangle$  such that  $p \in \mathcal{P}_i$  is a parking lot monitored by  $c_i$ . We also define a tuple  $\langle c_i, \mathcal{R} \rangle$  such that  $\mathcal{R} = \{r_1, r_2, \dots, r_j\}$  is a set of reference images taken by  $c_i$  when the lots in  $\mathcal{P}$  are unoccupied and  $r \in \mathcal{R}$  is a reference image taken at some time period. For any  $p \in \mathcal{P}$ , the rectangular detection zone  $\mathcal{Z}(p) = \langle (x_1, y_1), (x_2, y_2) \rangle$  marks the area where the light intensity is measured in  $\mathcal{R}$  and also the image stream  $\mathcal{I}$  captured by  $c_i$ , which we define as a tuple  $\langle c_i, \mathcal{I} = \{r'_1, r'_2, \dots\} \rangle$ .

In defining  $\mathcal{Z}(p)$ , the coordinates are usually within the boundary defined by the parking lines *and* located approximately where the windshields are likely to appear for a given camera angle. As seen in Figure 3, after edge detection and binarisation, the edges of the windshields become a means to identify a change in the intensity reading in the ‘middle’ of the parking lot thus, suggesting the presence of a car. While it is possible to work on Figure 3(a), we find better accuracy after the binarisation filter as the margin of error is significantly increased – as shown in Figure 3(b).

The detection is made by comparing the intensity reading between  $r$  and  $r'$  for a given  $p \in \mathcal{P} \in \langle c_i, \mathcal{P} \rangle$  such that the intensity difference in the area defined by  $\mathcal{Z}(p)$  on  $r \in \mathcal{R} \in \langle \mathcal{P}, \mathcal{R} \rangle$  and  $r' \in \mathcal{I} \in \langle c_i, \mathcal{I} \rangle$  is above  $\varepsilon$ . In determining  $\varepsilon$ , the light intensity threshold that suggests the presence of a car, some calibration will be expected. This calibration is made with respect to the site condition and we believe is acceptable for a system of this nature. At our test site, we set this at a value of 15%. In other words, if the light intensity measured from the reference image  $r$  on the area determined by  $\mathcal{Z}(p)$  is 1, then the light intensity measured from  $r'$  on the same area must be  $> 1.15$  to conclude the presence of a car. The calculation to achieve this is given in Algorithm [1](#).

Since our images are grey scales, each pixel carries a value in the range of  $0 \dots 255$ , where 255 is a white that indicates the highest light intensity on the pixel. For any given  $\mathcal{Z}(p)$ , we are interested in the average intensity of light defined by  $\mathcal{Z}(p)$  on  $\mathcal{R}_{c_i}$  and  $r' \in \mathcal{I}_{c_i}$  respectively. We then compute the ratio to determine if the intensity in  $r'$  is  $\varepsilon$  higher than  $r$ . Equation [1](#) summarises this calculation.

$$\varphi(r, r') = \frac{1}{\mathcal{Z}(p)} \sum \frac{r'_{(x,y)}}{r_{(x,y)}} = \begin{cases} > \varepsilon & \text{Occupied} \\ \leq \varepsilon & \text{Free} \end{cases} \tag{1}$$

---

**Algorithm 1.** it DetectCars ( $\{\langle c_i, \mathcal{P}_i, \mathcal{R}_i, \mathcal{I}_i \rangle, \dots \}$ )

---

```

for all  $p \in \mathcal{P}_i$  do
   $r \leftarrow \sigma_{\text{Current time window}}(\mathcal{R}_i)$ 
   $r' \leftarrow \sigma_{\text{Most recent image}}(\mathcal{I}_i)$ 

   $\mathcal{F} = \{\text{Greyscale, Edge Detection, B/W Detection}\}$ 
   $r \leftarrow \text{ApplyFilter}(r, \mathcal{F})$ 
   $r' \leftarrow \text{ApplyFilter}(r', \mathcal{F})$ 

  if  $\varphi(r, r') > \varepsilon$  then
    print Occupied for  $c_i.p$ 
  else
    print Free for  $c_i.p$ 
  end if
end for

```

---

As mentioned and is the case with any threshold, this needs to be adjusted to suit individual cases. Once camera positions are fixed, reference images may be taken and the image streams can be used to empirically work out the best value of  $\varepsilon$  for a given camera. Once set, any variation in the environment is compensated using a different  $r \in \mathcal{R}$  instead. Thus  $\mathcal{R}$  is critical in cancelling out any noise that may impede accurate detection. We note that this calibration process is far more efficient than some machine learning approach, where training and verification can take longer.

We also apply an additional binarisation filter to eliminate pixel noise to achieve a cleaner wire frame of a car. We find that doing this will improve accuracy further when pixel values are cleaned up to either a value of 0 or 255. The challenge of using this filter is the need to provide a threshold  $\ell$ , where a pixel value  $> \ell$  will result in a white (and black otherwise). It is tempting to simply go for the mid-value of the intensity range, i.e., setting  $\ell = 127$ . However, doing so will not allow for varying light intensity in different photos and can, as our experiments show, result in a poorer detection accuracy. To determine the right  $\ell$ , we first find the average intensity in the image. Next, we adjust this mean value by adding 1-standard deviation to the threshold to derive  $\ell$ , which is the basis for the binarisation filter. We find that by adding 1-standard deviation to the mean intensity of the image, the results are more accurate as image noise are removed.

## 4 Experimental Results

In determining the effectiveness of our approach, we will benchmark our technique against that reported in [19]. Their proposal uses the SVM, where multiple classifiers were built to determine the availability of 3 parking lots. Whenever possible, we replicate the empirical conditions used in [19] so as to give an accurate comparison.



	Approach by [19] using 2400 samples	Proposed Technique (w/o training samples)
against SVM (3 spaces)	85%	93%
against SVM (3 spaces) + MRF	93.52%	93%
against SVM (1 space)	83%	97%

**Fig. 4.** A comparison of detection accuracy using highest level of training samples in [19] against the proposed technique, where such training is almost non-existent

In our setup, the same number of samples, i.e., 300, were taken on 3 parking lots as shown in Figure 1. Like our benchmark, the samples were taken over a day from the same position accounting for lighting conditions, changes in the colour intensity, and movement of vehicles in and out of the parking lots. We also experienced rain conditions that wasn't in the plan but nevertheless provided additional consideration in the design of such a parking system. Unlike the benchmark however, we do not require prior training. Instead, we define  $\mathcal{Z}(p_1)$ ,  $\mathcal{Z}(p_2)$  and  $\mathcal{Z}(p_3)$  indicating where the windshields are likely to be.

We also spent another day taking images for  $\mathcal{R}$  at 4 interval periods: early morning, late morning, early afternoon and late afternoon. We then recreate the 8-class SVM reported in the benchmark by replicating the training process. In doing so, the immediate difference is the amount of overheads required in the preparation of the benchmark method. For 3 parking lots, 2400 patches (300 for each class) of the image is needed. Acquiring these patches proved a very time consuming process that is unattractive when scaling up to large parking sites. Compared to our approach, there is no need to involve the mammoth task of training, which of course is an immediate benefit.

In [19], a range of classification accuracies were reported using different number of training samples. In this paper, we work directly with the highest number of samples so as to yield the most accurate version of their classifier. We then compare this accuracy level against our work using the same test images. On our tests (Figure 4), our approach achieved 93% in classification accuracy for 3 parking spaces and a very high 97% on a single parking space. This is on par with the reported 93.52% accuracy in [19], when a high level of training samples are used with the Markov Random Field (MRF) correction (for 3 spaces). When training samples are dropped, our technique becomes immediately attractive when the high cost of training is eliminated. In our case, the inaccuracies were a result of small cars giving rise to a bigger variation of the windshield positions within the parking lot. In such a situation, the light intensity gathered by our algorithm were too low to trigger a detection, i.e., false negatives.

In [19], the average conflict rate was also measured. This measure reflects the error as a result of camera angles capturing the presence of other cars beside the lot of interest, e.g., Figure 3(b) when camera angle is positioned on the left of the image. Again on this measure, the benchmark performed better only when there

	False Accept Rate	False Reject Rate
against SVM (3 spaces)	4.39%	8.73%
against SVM (3 spaces) + MRF	1.25%	3.56%
against SVM (1 space)	4.85%	8.12%
against proposed technique	3.86%	5.34%

**Fig. 5.** A comparison of false positives and false negatives using the techniques in [19] against the proposed technique

are sufficiently high training samples. In many cases, we can improve on this measure by reconsidering the camera positions. In positions where the overlap is minimal, we can improve on this measure without changing any part of the algorithm. We do recognise that this suggestion may increase the cost of the system but it will really depend on the decision maker to decide the balance to strike with accuracy on this matter and the costs.

The final measure is on the level of false positives. The challenge of false positives arise out of changing light conditions. Primarily due to the movement of sunlight in open spaces and the colour of the car in sheltered parking sites, our approach has a rate of 3.86% on average. While this is higher than the proposed SVM and MRF correction method in [19] (1.25%), it performs better than the benchmark without the MRF correction (4.39%). We intend to improve on this measure as part of our future work. Instead of just increasing the number of reference images in  $\mathcal{R}$ , we will consider similar correction techniques like the MRF used in the benchmark. On the performance of the false negatives though, our performance does not seem to differ greatly from the techniques evaluated (as reported in Figure 5).

We have also included a benchmark on execution performance of our proposal. Our implementation uses C# and the .NET's built in graphics library for bitmap manipulation. For the image filters, we used an Open Source library called AForge.NET ([www.aforge.net](http://www.aforge.net)) to allow us to quickly implement the algorithm to test the viability of our proposal. On this implementation, we found the execution performance to be acceptable when images captured are reduced to sizes of 640x480 or below. Any images larger in size, and hence with bigger number of pixels yield noticeable 'pauses' in intensity computation between images. On our implementation using images of different sizes, the time to process each image is shown in Figure X(a). We also noted that there are no significant accuracy issues in using the same image but of different sizes as long as  $\varepsilon$  is adjusted accordingly to match the number of pixels per image.

## 5 Conclusion

Despite advances in parking systems, we continue to face frustrations at heavily utilised parking sites. Current systems fail because drivers have no prior access to information until arrival. And upon arrival, much of the search for a free park is ad-hoc based on information from signages and light indicators within the

driver's line of sight. Our proposed system will ease driver frustrations through a system that integrates Internet-enabled smart phones. In Australia and many parts of the world, the ubiquitous adoption of such devices has made it feasible for drivers to access live parking information prior to arrival. When combined with other technologies, this opened up possibilities of a parking system that could inform drivers before arrival at site, direct drivers to parking lots, and thus regulating traffic in the surroundings. Our immediate future work is therefore to build applications on smart phones to demonstrate these ideas coming off a "camera-to-server" approach. Critical to achieving this is the development of a detection mechanism to fit the "camera-to-server" model, which is cost effective and technically viable. As argued earlier, existing systems and current experimental projects do not consider aspects of this problem. Our work thus fills this gap.

The research contribution is a method to enable a "camera-to-server" implementation by balancing the costs against the features needed to deliver the parking system. Unique characteristics of our approach include the applied insights of human detection (as reported in the psychological test conducted by Zhao and Nevatia [20]) in our algorithm, and the explicit coding of a detection behavior based on the learning characteristics of a SVM learner. By explicitly coding the classification behaviour of a SVM learner, we gain performance. At the same time, we can eliminate noise that would otherwise be embedded in the hyperplane through the normal training process. This gives us the improvement in detection accuracy. The final result is a detection mechanism that supports the "camera-to-server" approach with high efficiency and accuracy in an environment with multiple images streams.

## References

1. Chinrungrueng, J., Sunantachaikul, U., Triamlumlert, S.: Smart Parking: An Application of Optical Wireless Sensor Network. In: IEEE/IPSJ International Symposium on Internet Workshops and Applications, p. 66 (2007)
2. Farhan, B., Murray, A.T.: Siting park-and-ride facilities using a multi-objective spatial optimization model. *Computers and Operations Research* 35(2), 445–456 (2008)
3. Foresti, G.L., Micheloni, C., Snidaro, L.: Event Classification for Automatic Visual-based Surveillance of Parking Lots. In: International Conference on Pattern Recognition, vol. 3, pp. 314–317 (2004)
4. Hodel-Widmer, T., Cong, S.: PSOS: Parking Space Optimization Service. In: 4th Swiss Transport Research Conference, Verit/Ascona, pp. 1–22 (March 2004)
5. Inaba, K., Shibui, M., Naganawa, T., Ogiwara, M., Yoshikai, N.: Intelligent Parking Reservation Service on the Internet. In: Symposium on Applications and the Internet-Workshops, San Diego, CA, USA, pp. 159–164 (2001)
6. Jones, W.: Parking 2.0: Meters Go High-Tech. *IEEE Spectrum*, 20 (2006)
7. Lee, C.H., Wen, M.G., Han, C.C., Kou, D.C.: An Automatic Monitoring Approach for Unsupervised Parking Lots in Outdoors. In: International Carnahan Conference on Security Technology, pp. 271–274 (October 2005)

8. Li, Y., Ma, R., Wang, L.: Intelligent Parking Negotiation Based on Agent Technology. In: WASE International Conference on Information Engineering, vol. 2, pp. 265–268 (2009)
9. Liu, Q., Lu, H., Zou, B., Li, Q.: Design and Development of Parking Guidance Information System based on Web and GIS Technology. In: 6th International Conference on ITS Telecommunications, Chengdu, China, pp. 1263–1266 (2006)
10. Masaki, I.: Machine-Vision Systems for Intelligent Transportation Systems. *IEEE Intelligent Systems and their Applications*, 13(6), 24–31 (1998)
11. Mathijssen, A., Pretorius, A.: Verified Design of an Automated Parking Garage. In: Brim, L., Haverkort, B.R., Leucker, M., van de Pol, J. (eds.) *FMICS 2006 and PDMC 2006*. LNCS, vol. 4346, pp. 165–180. Springer, Heidelberg (2007)
12. Mo, Y., Su, Y.: Design of Parking Guidance and Information System in Shenzhen City. In: *ISECS International Colloquium on Computing, Communication, Control, and Management*, vol. 4, pp. 37–40 (August 2009)
13. Mohan, A., Papageorgiou, C., Poggio, T.: Example-Based Object Detection in Images by Components. *IEEE Trans. Pattern Analysis and Machine Intelligence* 23(4), 349–361 (2001)
14. Mouskos, K.C., Maria Boile, N.A.P.: Technical Solutions to Overcrowded Park and Ride Facilities. Technical Report: FHWA-NJ-2007-011, University Transport Research Centre, Region 2 (2007), <http://tris.trb.org/view.aspx?id=814921>
15. Osuna, E., Freund, R., Girosi, F.: Support Vector Machines: Training and Applications. Tech. Rep. Massachusetts Institute of Technology, Cambridge, MA, USA (1997)
16. Rodier, C.J., Shaheen, S.A., Kemmerer, C.: Smart Parking Management Field Test: A Bay Area Rapid Transit (BART) District Parking Demonstration. Research Report: UCD-ITS-RR-08-32, Institute of Transportation Studies, University of California, Davis (2008), [http://pubs.its.ucdavis.edu/publication\\_detail.php?id=1237detail.php?id=1237](http://pubs.its.ucdavis.edu/publication_detail.php?id=1237detail.php?id=1237)
17. Schneiderman, H., Kanade, T.: A Statistical Method for 3D Object Detection Applied to Faces and Cars. In: *International Conference on Computer Vision and Pattern Recognition*, Hilton Head, SC, USA, pp. 1746–1759 (2000)
18. Vidal-Naquet, M., Ullman, S.: Object Recognition with Informative Features and Linear Classification. In: *IEEE International Conference on Computer Vision*, Nice, France, pp. 281–288 (2003)
19. Wu, Q., Huang, C., Wang, S., Chiu, W., Chen, T.: Robust Parking Space Detection Considering Inter-Space Correlation. In: *IEEE International Conference on Multimedia and Expo.*, pp. 659–662 (July 2007)
20. Zhao, T., Nevatia, R.: Car Detection in Low Resolution Aerial Images. *Image and Vision Computing*, 710–717 (2001)
21. Zhong, H., Xu, J., Tu, Y., Hu, Y., Sun, J.: The Research of Parking Guidance and Information System based on Dedicated Short Range Communication. In: *Proceedings of the IEEE Intelligent Transportation Systems*, vol. 2, pp. 1183–1186 (October 2003)

# An Adaptive Framework for Personalized E-Learning

Ronnie Cheung and Hassan B. Kazemian

London Metropolitan University,  
United Kingdom  
ccheung@acm.org, h.kazemian@londonmet.ac.uk

**Abstract.** In this paper, an adaptive hypermedia framework is used in the design and development of an adaptive personalized e-learning system for Java programming. Learners with different learning goals, background and learning aptitudes are treated differently by including a model of knowledge and preferences in the system. A personalized e-learning framework is developed to provide a basic Java programming course that allows students to learn on his/her own pace and with adaptive features. The system evaluation shows that most students consider the personalized e-learning system useful for learning basic Java programming.

**Keywords:** e-learning; ontology; personalization; adaptive system.

## 1 Introduction

Traditional Web-based leaning systems are based on static contents that are accessed by different learners. Such kind of approach may not be effective for learners with different backgrounds and abilities. The ability to learn a topic is strongly related to an individual's background, learning aptitude and learning aptitude. For example, the courseware for a programming course may include a large number of programming examples and demonstrations. However, lengthy treatments of course contents may not be suitable for advanced learners who want to master the advanced features for the subject matter in a short time. Therefore, it is necessary to design an adaptive framework to implemented e-Learning systems that can cope with the differences in abilities of the learners in terms of background, preferences, learner aptitude, attitude and learner performance with previous activities in the system [2]. In this paper, we describe the development of a personalized e-learning system using an adaptive hypermedia approach. The Personalized E-learning System (PES) is developed as an educational hypermedia system using Semantic Web technologies. It is based on an adaptive framework that is capable of retrieving distributed learning repositories. A personalized adaptive Java course is adopted as the courseware so that students at different educational levels can learn basic Java programming. The initial course level is determined by the student's programming background. The course level of the subsequent chapters would be promoted to a more advanced level, or demote to a less advanced level depending on the chapter test results.

With the rapid development of distance learning and the Web technologies, Web-based learning has now become an important branch in the education technology. One of the benefits for e-learning is that the learning environment can be adapted to the individual's learning process and learning need. In fact, the adaptability and the personalization feature of the e-learning environment is one of the research areas that draw much attention in the e-learning field. The next generation of Semantic Web technologies provides a common framework that allows data to be shared and reused across applications, and it is considered to be a promising technology for implementing e-learning systems.

This paper describes the features of an adaptive e-learning system from the perspectives of content adaptation and generation based on different user models. Personalization and adaption are achieved by implementing an adaptive hypermedia architecture, which includes three models: the domain model, the learning models and the adaptation model. The implementation details of the adaption and personalization features are discussed with reference to the information stored in the domain model, the learner model and the adaptation model. The detailed system design aspects are described with details on how adaptive contents can be implemented and generated using a rules-based reasoning mechanism. Finally, a sample course model "guided study" is used to demonstrate the dynamic course sequencing and presentation processes.

## 2 Literature Review

### 2.1 Semantic Web Technology

The concept of the Semantic Web has emerged with the aim of making web resources more understandable by machines. It provides semantic-based access to the Internet, and extract information from texts in addition to being used in many applications to explicitly declare the knowledge embedded in them [4]. It provides a common framework that allows data to be shared and reused across applications, enterprise and community boundaries [9]. The main task of the Semantic Web is to "express meanings". In order to achieve this, several layers are needed including the XML layer which represents data, the Resource Description Framework (RDF) layer which represents the meaning of the data, and the Ontology layer, which represents the formal common agreement about the meaning of data, and Logic layer, It also enables intelligent reasoning with meaningful data. The effectiveness of the Semantic Web will increase drastically as more machine-readable Web content and automated services (including other agents) become available. This level of inter-agent communication will require the exchange of "proofs". Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF).

Ontologies are specifications of the conceptualization and corresponding vocabulary used to describe a domain [5]. They are well suited for describing heterogeneous, distributed and semi-structured information sources that can be found on the Web. By defining shared and common domain theories, ontologies help both people and machines to communicate concisely, supporting the exchange of

semantics and not only syntax. It is therefore important that any semantic for the Web is based on an explicitly specified ontology. By this, consumer and producer agents of the Semantic Web can reach a shared understanding by exchanging ontologies that provide the vocabulary needed for discussion. Ontologies typically consist of definitions of concepts relevant for the domain, their relations and axioms about these concepts and relationships. Several representation languages and systems are defined, but the more recent language, OWL (Web Ontology Language) is developed to unified different ontology languages. It is a representation language for describing web resources and supporting inference over those resources.

## **2.2 Personalization and Adaptive Educational Hypermedia Systems**

Personalization is an important issue which tailors and customizes learning experience to individual learners, based on the analysis of the learner's objectives, current status of skills and knowledge, and learning style preferences [3][8]. In a personalized e-learning system, the learner's interaction, information requests, problem-solving attempts will be recorded such that the system is able to recommend information to the learner or translate the learner's request into a query and send the query to the destination through education web service or other means.

In the hypermedia or hypertext paradigm, information is interconnected by links; different information items can be accessed by navigating through the link structure. Adaptive Hypermedia (AH) systems merge hypermedia with the user modeling technology, and can be applied in a variety of application area, of which one dominating area is education. In the e-learning area, personalization can be implemented by using Adaptive Hypermedia. Adaptive Educational Hypermedia (AEH) [1] deals with the issue of providing a personalized educational experience. An AEH system aims at providing adaptive presentation of multimedia or text according to the learner's needs.

Adaptive Educational Hypermedia Systems overcome the problem of presenting the same content to different users in the same way, regardless of their different interests, needs and backgrounds. AH systems provide two general categories of adaptation. Adaptive content presentation is presenting the content in different ways, according to the domain model (concepts, their relationships, prerequisite information, etc) and information from the user model. Adaptive navigation is which the system modifies the availability and/or appearance of every link that appears on a Web page, in order to show the user whether the link leads to interesting new information, to new information the user is not ready for, or to a page that provides no new knowledge. Effective adaptive hypermedia systems lead to impactful collaboration [10].

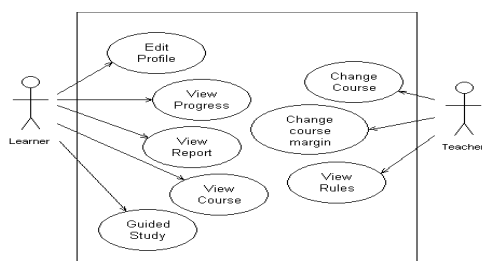
## **3 System Architecture**

### **3.1 Design Overview**

The personalized e-learning System (PES) is designed to use the Semantic Web technology to provide personalized e-learning experience for the learners based on a

hypermedia adaptive framework. The adaptive framework is based on the Adaptive Educational Hypermedia System (AEHS).

Figure 1 shows the use case diagram of the PES system. In this system, the learner is able to view and edit his profile, which contains the learner's programming background and experiences, and other preferences. These data will be used as our adaptation parameters. When the learner logs onto the system, a preset course will be loaded. The learner can view the table of content of the course, and access any unit within the course. In the basic Java programming course, a guided study feature is provided for the learners to learn the course sequentially. The sequence of the course and the course content display will be adapted according to the learner's learning aptitude and test results. To make this system more user-friendly, the learner can view his learning progress and his assessment report.



**Fig. 1.** Use case diagram for personalized e-learning system

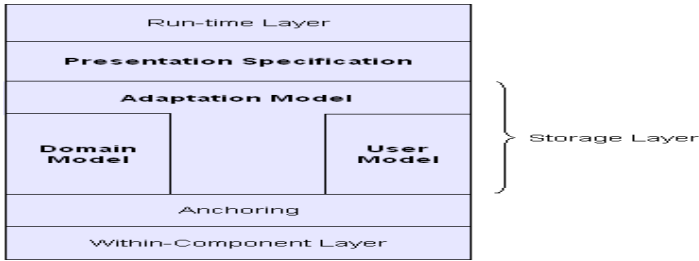
The teacher can log onto the system and change the course. In the personalized e-learning ontology-based system, the course change is apparently equivalent to changing the ontology file of the course. In the course sequencing, the test results scores are used to determine which course content and course level should be displayed. These scores are considered as the course margin that can be changed by the teacher so as to improve the effectiveness of this adaptive feature. In our proposed system, we use the Jess Expert System Shell to realize the adaptive features. The knowledge base, which is a set of Jess rules are stored in a text file to be loaded into the system during initialization. These rules can be viewed by the teacher.

### 3.2 System Architecture

The system architecture of the personalized e-learning system is shown in Figure 2. It provides a framework to express the functionality of adaptive hypermedia systems by dividing the storage layer into three parts that specify what should be adapted, according to what features should it be adapted, and how should it be adapted.

Adaptation to the user's individual characteristics is implemented according to the user model. A user model determines the user's goal, tasks, beliefs and characteristics that are important for adaptation. According to Brusilovsky [1], a hypermedia application can be adapted to the user's knowledge, goals, background and experience, preferences, interests, individual traits and environment. In addition, the following groups of individual user characteristics could be important for the adaptation of e-learning systems: personal data (including demographic characteristics such as age,





**Fig. 2.** System Architecture of PES

culture), psychological, cognitive and physiological parameters such as user's attention (simple and complex reaction time), memory (verbal working memory, long-term memory), cognitive abilities (spatial arrangement, etc), user's internality/externality, cognitive and learning styles, and personal decision abilities.

The storage layer consists of three models, mainly the domain model, learner model and the adaptation model. The domain model describes how the information content is structured. The user model describes the information about the user. The adaptation model contains the adaptation rules that define how the adaptation is performed. Each model is interrelated and can be accessed by the user.

### 3.2.1 The Domain Model

The domain model specifies the conceptual design of the adaptive hypermedia application, i.e. what will be adapted. The design of the domain model is to design the concept hierarchy of the learning objects. In our system, the learning object is the Java course provided through the distributed learning resources. There are three levels of the Java course, with different emphasis on the detail description of the subject depending on the learning preferences and performance of the learner. Each level has its own set of questions. The teacher can choose to display only the questions of the corresponding level, or display questions of all levels to determine the learner's aptitude of the learned course. The course level that is presented to the student is initially determined by the learner's background. Course level of subsequent chapter can be changed according to the learner's test results. If the learner has chosen to take pretest, the pretest result will be taken as the reference, otherwise, the chapter test result of the previous chapter will be used as a reference, assuming that if the user has fully understand the topic, he/she will get a high score in the chapter test. This result is used as a prediction for his future learning. If the learner desires to learn from more examples as his learning preference, more examples on the subject will be displayed.

### 3.2.2 The Learner Model

The learner model is designed based on the some general learning scenario of the students. In particular, the learner model described the learner's static information related to the following items:

- The name and student ID of the learner
- The education status, whether he is a high school student, bachelor degree or postgraduate level
- The learning goal, which course he wants to take,

- The motivation of the learner, whether he has high, middle or low level of motivation to learn the subject
- The learner's experience in the chosen subject, whether he has no experience, basic, intermediate or advanced level
- The learning style, whether he prefers principle-oriented or example-oriented.

The learner model also describes the learner's learning process and behavior related to the following item:

- What knowledge has the learner mastered? Has he mastered a certain topic in the chosen subject?
- What knowledge has the learner not mastered? What knowledge has he missed?
- At which part of the knowledge is the learner weak/strong?
- What examples were given to the learner?
- What answers did the learner give to a certain question?
- What score did the learner get in the test?
- How many times did the learner try before answering the question correctly?
- What is the rate of success and failure?
- How much time did the learner spend on the questions?
- How well does the learner master the subject?
- How much time did the learner use to master the subject?

The ontology-based learner model describes static and dynamic information that is related to the particular learner. The user preference module describes the static information of the learner that is set by the learner upon registration and the user knowledge module describes the dynamic information that is generated during the learning process. The static information will be created when the learner registered with the system. Registered user can log on to the system to edit the learner profile. The user knowledge is built as a hybrid of an overlay model and a historic model. The key principle of the overlay model is that for each domain model concept, an individual user knowledge model stores some data that can estimate the knowledge level on this concept. The historic model keeps some information about the learner's visit to individual page. In particular, the dynamic information in the learner model includes user visited pages, logged on duration, attempted test, test result, course he has completed, etc.

### 3.2.3 The Adaptation Model

The adaptation model specifies the specific adaptation semantics of the system. The adaptation model consists of adaptive presentation, adaptive navigation support and curriculum sequencing.

The adaptive presentation includes the optional pretest generation and course presentation. The pretest can be generated according to the learner's preference, background and the chapter end test (exit test) result of the previous unit. The adaptation rules in the system define the condition that generates the pretest question list. If the learner's programming background is "novice" and he has chosen to take pretest, the pretest question list will be the same as the last exit test question list. If the learner's programming background is "expert", the pretest questions will consist of

both the last exit test questions and the coming chapter end test questions, assuming that the learner may know the subject from other means. The adaptive presentation is also responsible for presenting relevant information to the learner to ensure that he can understand and learn the subject according to his preference and ability. In our system, one of the three different levels of the course will be displayed to the learner according to his learning progress. The default course view level will be set by the learner's background and programming experience. However, this level can be changed according to the learning progress. This level change can be determined by the chapter end test (exit test) result and if pretest is available, by the pretest result. The Jess rule defines the course view level to be displayed according to the learner's pretest result. If the pretest results of a particular document, which represents a unit or a subunit of a course, is higher than the course margin that has been set by the teacher, the system will promote the level of the document to be displayed. If the learner has chosen "no pretest" as his learning preference, the course level change will be determined by the result of the exit test. The system will present the next document only if the learner has passed the exit test.

The adaptive navigation support guides the learner towards the relevant, interesting information related to the learning topic. In our system, we provide additional examples and references hyperlinks to the learner on the particular topic. The learner can increase their knowledge on the particular topic through these additional links. The system will determine the additional links according to the viewed topic and the learner's preference. The adaptive navigation includes a set of recommendation rules to suggest the learners which topic is suitable. These recommendations are also used in the course sequencing. The system will use the prerequisite of a particular concept to determine if the related document should be recommended to the learner. If a document with concept has been visited, and there exists where the concept is a prerequisite for another concept, the document with this other concept will be recommended. When the learner has finished viewing the selected topic, the system will generate the exit test for the learner to make sure that he has understood the topic. The results will be recorded and updated in the learner's profile. The assessment item type of our system contains multiple choice questions only. The learner can set his exit test condition to either "no passing grade" or "pass at 50%". If the learner has failed the exit test, the system will suggest the learner to revisit the topic, assuming that he has not fully understand the topic. The idea of course sequencing is to generate an individualized course for each learner by dynamically selecting the most optimal teaching operation, which includes course content presentation, examples, questions or problems, at any moment of the learning process. In our curriculum module, we use the above recommendation to represent the chosen course sequence. The adaptive features of the personalized e-learning system are based on the learner model. The features include:

1. Adaptive Pretest Generation – the system generates a set of pretest questions according to the learner's preference. Research in the field of educational psychology demonstrates that students' prior knowledge in a key factor contributing to the learning process.

2. Adaptive Content Presentation – for a course with different presentation level (level 1: basic – detailed description of the topic with graphs, examples and other illustrations, level 2: intermediate – relatively fewer examples and illustrations, level 3: advanced – summary of the topic), the system can display the appropriate level to the learner.
3. Adaptive Course Sequencing – the system can guide the learner to the next appropriate page according to the student’s learning progress and test results.
4. Adaptive Link Annotation – the system can recommend the student to pages that according to the pretest result, chapter end test result or the pages that the learner has covered or “visited”.

## 4 Detailed System Design

The personalized e-learning system uses the Model-View Controller (MVC) architecture to separate application logic and page presentation as shown in Figure 3.

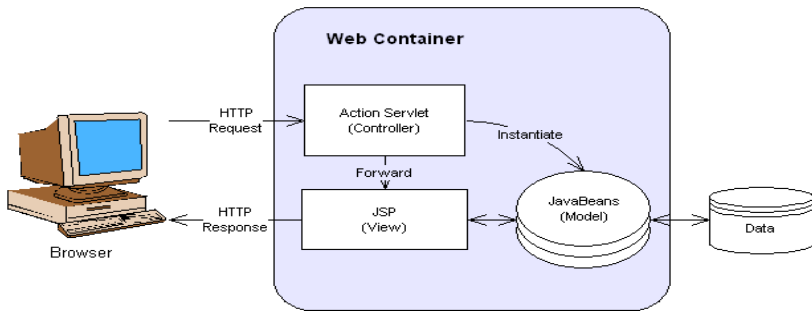


Fig. 3. The MVC Model

With this model, the servlets control the flow of the web application and delegate the business logic to the JavaBeans while the JSP pages generate the HTML for the web browsers.

In the PES system, the system data consists of the text file that stored the Jess rules and initial Jess facts and an OWL ontology file for each student describing the learner’s profile and his learning progress. Figure 4 shows the model view of the PES system. For each application session, the session object contains both the course bean and the student bean such that the Java Server Pages will handle the presentation logic and process the Java Beans to generate dynamic presentation code. The course bean contains the LearnerJessInfo object that stores the Rete object from Jess’s Java API [6]. The Rete object stores the initial Jess facts and Jess rules from the text file and Jess facts generated at runtime. The course bean also contains the LearnerOWLInfo object that stores the JenaOWLModel object from the Protégé-OWL API [7]. The JenaOWLModel object stores the learner and course content of the ontology.

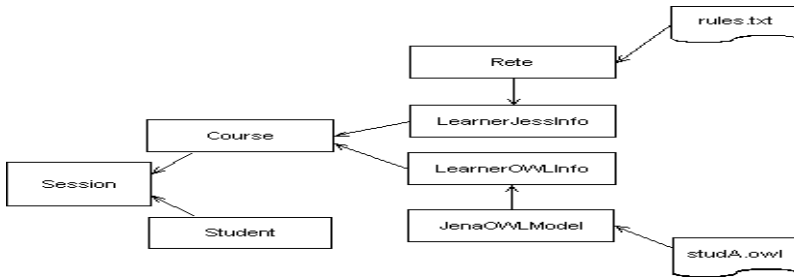


Fig. 4. The Model View of the PES system

### 5 Rule-Based Reasoning

In the PES personalized e-learning system, we use the Java Expert System Shell (Jess) [6] as our adaptation rule parser. The Jess architecture is shown in figure 5. It is designed as a library that can be embedded into many applications. It has an extensive Java API to interact with Jess. The core of the Jess library is the `jess.Rete` class. An instance of `jess.Rete` is an instance of Jess. Every `jess.Rete` object has its own independent working memory, list of rules and set of functions. The `jess.Rete` class exports methods for adding, finding, and removing facts, rules, functions, and other constructs. The Rete class is a façade for the Jess library.

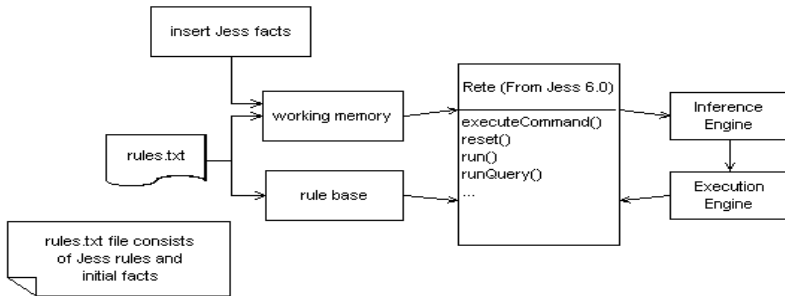


Fig. 5. The Java Expert System Shell

The following code extract creates the `jess.Rete` object when the personalized e-learning system starts. When the system starts, `jess.Rete` object will be created. By calling the `executeCommand()` to run the command “batch rules.txt”, the rules file “rules.txt” will be read into the system. The file “rules.txt” contains the initial Jess facts and Jess rules of the system. Calling the `reset()` command will reinitialize the working memory of the Jess’s rule engine. After reinitializing the working memory, all activated rules in the rule base will be fired by calling the function `run()`. Jess’s Java API reports errors by throwing instances of `jess.JessException`. Therefore, it is necessary to catch this exception when working with Jess in Java.

```

import jess.*;
try {
    rulesFile = "rules.txt";
    Rete engine = new Rete();
    engine.executeCommand("(batch \"" + rulesFile + "\")");
    engine.reset();
    engine.run();
} catch (JessException e) {
}

```

In the PES system, we use the forward-chaining rules in Jess. The Jess rules will be fired when the ‘if’ statement is matched. The new facts will be stored in the Rete object. Calling the runQuery() method can retrieve the facts, thus displayed on the web pages. During the learning process, additional Jess facts will be asserted into the working memory and the Jess rules will be executed accordingly. These Jess facts include learner’s visited pages and test results. The inference rules and inference processes are implemented according to the OPAL system implemented by Cheung et al. [3].

## 6 The Core Ontologies

Figure 6 shows the relationship between the ontologies used in the system. In the personalized e-learning system, the ‘doc.owl’ is a domain OWL ontology that contains the relation between the documents and concepts. The ‘qti.owl’ contains the relation between the test and questions. The ‘java.owl’ contains the ontology description of the Java course, which includes document relations and questions relations described in ‘doc.owl’ and ‘qti.owl’. In our system, we have created two sets of basic Java course, one uses the Sun Java Online Tutorial to learn Java, and the ontology instance is found in ‘java1.owl’. We further develop another basic Java course that has different course levels. The ontology instance is described in ‘java2.owl’.

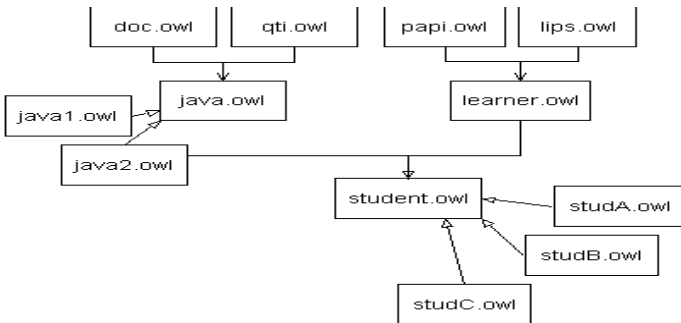


Fig. 6. The Relationship between the Ontologies in PES

The ‘papi.owl’ contains the semantic description of the learner based on the learner standard IEEE PAPI, whereas the ‘lips.owl’ contains the semantic description

according to IMS LIPS. The OWL ontology file ‘learner.owl’ describes the learner’s profile based using these two descriptions.

The ‘student.owl’ describes the learner who learns Java based on the ‘learner.owl’ and ‘java1.owl’ or ‘java2.owl’. In our system, we can interchange the two Java courses in the teacher’s mode. Each learner has his own ontology instance. If the learner’s ID is “studA”, the learner’s ontology instance will be ‘studA.owl’. When a learner logged in to the system, the system will check if there exists the learner’s ontology instance. If the ontology instance does not exists, the system will generate an instance file according to the learner’s ID.

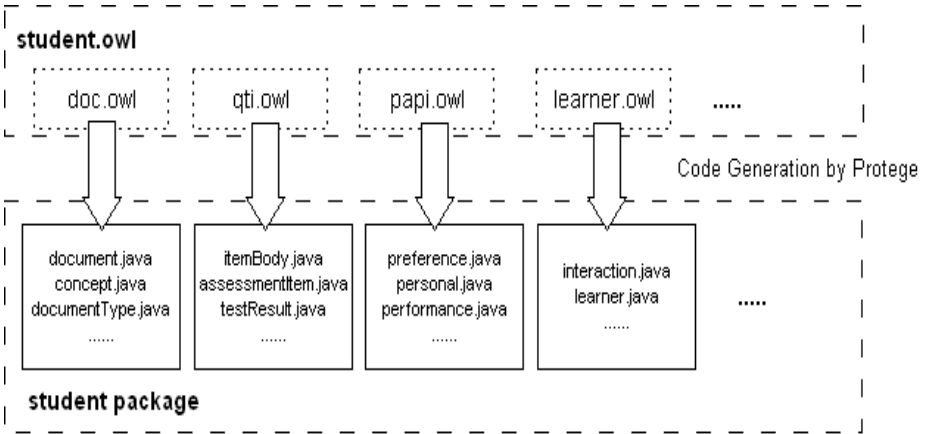


Fig. 7. The Core Ontologies

Figure 7 shows the relationship between the ontology ‘student.owl’ and the student Java package. All ontologies are created using Protégé [7], which is a free, open source ontology editor. In our proposed system, we use the Protégé-OWL editor, which is an extension of Protégé that supports the Web Ontology Language (OWL) to create OWL ontologies. The Protégé-OWL editor has an open-source Java API for the development of custom-tailored user interface components. It also comes with a code generator that is able to generate Java interfaces and implementation of OWL ontologies classes. In the student.owl described above, it is composed of the domain OWL ontologies and the learner OWL ontologies. The code generator converted the ontology classes into corresponding Java classes. The code generator also creates a Factory class so that it becomes easy to create, modify and query the ontological resources using standard library classes.

```
private Learner getAllLearner(String uriStr) {
    JenaOWLModel owlModel=ProtegeOWL.createJenaOWLModelFromURI(uriStr);
    MyFactory myFact = new MyFactory(owlModel);
    Learner learner = (Learner)myFact.getAllLearnerInstances();
    return learner;
}
```

As an example, the sample function getAllLearner() can get all the learner instances from the OWL ontology file, with filename “uriStr”.

## 7 Course Sequencing and Presentation

Figure 8 shows the flowchart for the main logic for course sequencing in the system. The “Guided Study” is an adaptive course sequencing feature for the learner to learn the course unit according to their past interactions. When the learner presses “Guided Study” in the course page, the system will find the latest course level and the current unit from the learner’s ontology instance. The initial course level is determined by the learner’s background and programming experience shown in table 1. The initial current unit will be the first unit of the course. If the learner has started the course, the latest course level and the current unit that has been stored in his last visit will be retrieved from the learner’s ontology instance. Each unit has a unit type. If the unit type is a pretest, the pretest will be displayed. The learner will be requested to answer the questions before he can proceed with the course content. This pretest is optional, and is set in the learner’s profile. The pretest result will be used to determine if the course level has to be promoted or demoted. When the learner has finished the pretest, the system will display the course unit content with the course level either determined by the learner’s profile, or the previous results. The system will also display the course content when the view type of the current unit is “lecture”, too. When the course unit content is displayed, the system will update the learner “visited” pages, assuming that the learner will read the content of the unit displayed. Upon completion of the course unit content, the learner can press “Next” to proceed to the next page. If the learner prefers “more examples” in his learner profile, an example page will be displayed to further explain the course unit content to the learner. Again, the system can display this page when the learner press “Guided Study” and the current unit view type happens to be an “example”. When the learner press “Next” again, the course unit end test will be displayed. The test questions can be either contains a single level of questions that is coherent with the course level, or a mixture of different levels. The test questions will be determined by the teacher. In the teacher’s page, the teacher can choose between “same”, which is the same level of question as that of the course level, or “mix”, which is a mixture of different levels of questions. In this way, the teacher can change the test questions to meet the needs of different learner group.

When the learner has completed the chapter end test questions, the system will check the results, and will determine the course unit and course level according to the learner’s result. There are four possibilities. If the learner has failed the test (rp), the system will demote the course level; if the learner has passed the course, but the score is lower than the demote score (du), the system will increase the course unit and demote the course level; if the result is above the demote score but under the promote score (ru), the system will increase the course unit only. If the learner test score is beyond the promote score (pu), the system will increase the course unit and promote the course level. The system has assumed that if the learner has a high score in the test, the learner understands the subject well and there is a high probability that the learner is able to learn faster by displaying course content that has less illustrations and examples. This means that the learner will be able to view the next unit in a higher level.



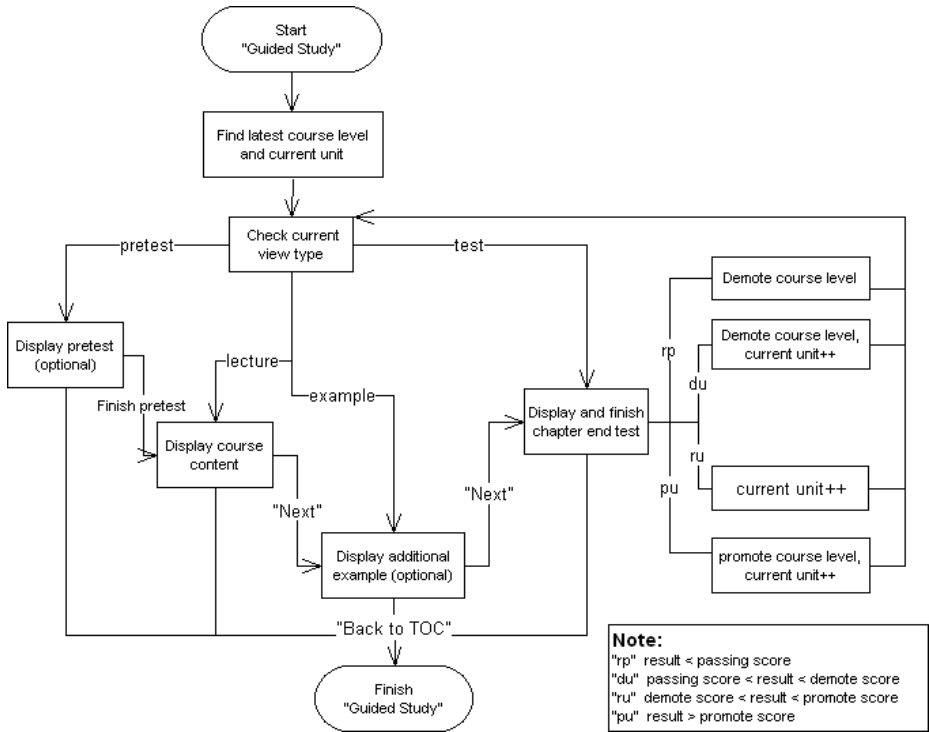


Fig. 8. Flow Chart for “Guided Study”

Table 1. The Initial Course Level

Programming background	Programming experience	Initial course level
Novice	Novice	Basic (1)
Novice	Intermediate	Basic (1)
Novice	Expert	Intermediate (2)
Intermediate	Novice	Intermediate (2)
Intermediate	Intermediate	Intermediate (2)
Intermediate	Expert	Intermediate (2)
Advanced	Novice	Intermediate (2)
Advanced	Intermediate	Advanced (3)
Advanced	Expert	Advanced (3)

The content presenter is responsible for presenting information to the learner or the teacher according to the user login. If the user is logged in as a learner, the learning environment will be displayed. If he is a teacher, the authoring environment will be displayed. Figure 9 shows an example of a course page in PES, with the details of the pre-tests and exit-test for the learner.

The screenshot shows a web browser window titled "Personalized E-learning System - Windows Internet Explorer". The address bar shows "http://158.132.10.171:25005/polyu/assessment.jsp". The page content includes a navigation menu with links for "Welcome Mary Lou", "User's Profile", "Learning Progress", "User's Report", "Course", and "Logout". The main heading is "Learning Progress of Mary Lou". Below this, the course title is "Basic Java Programming Tutorial", learning preference is "more examples", pre-test condition is "no pretest", and exit test condition is "pass at 50%". A table displays the learning progress for five units. The first unit, "Basic Java Programming I", has a pre-test score of 0, an exit test score of 52, and a result of "completed" on 2008-04-25 at 17:42:42. The other units have pre-test scores of 0 and exit test scores of 0. A "Return to Course" button is located below the table. The footer contains the copyright notice "© 2008 Hong Kong Polytechnic University. All rights reserved." and the browser status bar shows "Internet" and "100%" zoom.

Unit	Description	Pre-test	Test date	Exit test	Test date	Result
1	Basic Java Programming I	0		52	2008-04-25 17:42:42	completed
2	Basic Java Programming II	0		0		
3	Classes and Objects I	0		0		
4	Classes and Objects II	0		0		
5	Inheritance	0		0		

Fig. 9. A User Content Page in PES

## 8 Conclusion

Using an adaptive hypermedia framework, we describe the development of a personalized e-learning system (PES) for students to learn basic Java programming, with adaptive features according to students' learning preference and performance. In particular, we have developed a 3-level Java course to enable course content level promotion and demotion to help the student to learn the subject in different levels. The PES system also gives suggestions to the students by system recommendations.

Through the system, students can learn the basic Java course with adaptive features according to their chapter end test results. The course level promotion and demotion is able to present different course content level according to their test results. The results of the system evaluation have shown positive results on this approach. However, it has also shown that this system is not welcomed by high school students who do not have online learning experience. This is because our target audiences are those learners who have online learning experience, but want to improve their learning process and effectiveness through a learning platform that can be adapted to their needs.

In the PES system, we used the overlay user modeling approach to provide personalized learning. The personalization is based on both the static and dynamic information about the user model. The personalized courseware is initially based on the static information set by the learner. As the learner interacts with the system, the learner's interaction with the system and learning performance will constantly update the user model, thus provided course level changes according to this dynamic information. We used the rule-based shell approach so that the knowledge base of the system can be encoded as rules in a clear and logic-like way. This approach also allows the teacher to edit the adaptation rules and manage the adaptation behavior of the system according to different needs. The ontology-based approach is used to describe the domain model and the learner model. This approach provides an easy way to improve the course content, create more questions format and add in different courses using the same domain ontology. It also demonstrates that it is a flexible and easy way to manage distributed learning resource with a single learning model for the user.

**Acknowledgments.** The authors would like to thank the Hong Kong Polytechnic University for supporting this research.

## References

1. Brusilovsky, P.: Adaptive Hypermedia. *User Modeling and User Adapted Interaction* 11, 87–110 (2001)
2. Chen, J., Man, H., Yen, N.Y., Jin, Q., Shih, T.K.: Dynamic Navigation for Personalized Learning Activities Based on Gradual Adaption Recommendation Model. In: Luo, X., Spaniol, M., Wang, L., Li, Q., Nejdil, W., Zhang, W. (eds.) *ICWL 2010*. LNCS, vol. 6483, pp. 31–40. Springer, Heidelberg (2010)
3. Cheung, R., Wan, C., Cheng, C.: An Ontology-based Framework for Personalized Adaptive Learning. In: *Proceedings of the 9th International Conference on Web-based Learning*, Shanghai, China, pp. 52–61 (December 2010) ISSN 0302-9743
4. Dublin Core, Dublin Core website (2004), <http://dublincore.org/>
5. Gómez-Pérez, A., Corcho, O.: Ontology Languages for the Semantic Web. *IEEE Intelligent Systems* 17(1), 54–60 (2002) ISSN 1541-1672
6. Jess 6.1, Jess, the Rule Engine for the Java Platform, <http://herzberg.ca.sandia.gov/>
7. Protégé 3.3.2, The Protégé Ontology Editor and Knowledge Acquisition System, <http://protege.stanford.edu/>
8. Sampson, D., Karagiannidis, C.: Personalised learning: educational, technological and standardisation perspective. *Interactive Educational Multimedia* 4, 24–39 (2002) ISSN 1576-4990
9. The World Wide Web Consortium, <http://www.w3.org/>
10. Caballé, S., Lapedriza, À., Masip, D., Xhafa, F., Abraham, A.: *Journal of Digital Information Management* 8(5), 322–329 (2010)

# Gov 2.0 and Beyond: Using Social Media for Transparency, Participation and Collaboration

F. Dianne Lux Wigand

Institute of Government, University of Arkansas at Little Rock,  
2801 South University Avenue, Little Rock, AR 72204-1099, USA  
fdwigand@ualr.edu

**Abstract.** A recent paradigm shift has broadly impacted the evolution of electronic government: Gov 2.0. This shift represents a massive change from a static web presence for the delivery of information and services to using collaborative web technologies to engage citizens, foster co-production, and encourage transparency in government. Social media are creating new pathways between citizens and government to access information and services. The author argues that by applying Diffusion of Innovation, Social Influence, and Collective Intelligence theories, one can better understand how Gov 2.0 technologies and applications are adopted and may enable transformative changes in the delivery of online government information and services. U.S. federal government initiatives adopting social media are examined. These observations demonstrate how interactions between citizens and government are changing and creating entirely new online communities that defy traditional communication reach and organizational boundaries. Challenges and barriers to achieving open government initiatives are presented. Recommendations are offered.

**Keywords:** social media, eGovernment, Web 2.0, Gov 2.0, transparency, participation, collaboration.

## 1 Introduction

Gov 2.0, like Web 2.0, is an umbrella term used to refer to a new era of Web-enabled applications and tools such as blogs, micro blogs, podcasts, Really Simple Syndication (RSS), social networking sites, video sharing, web chat, and wikis used to encourage citizen participation, collaboration and transparency. Like Web 2.0, Gov 2.0 does not refer to a new software version, but rather to a new, i.e. second, phase of the evolving and extended Internet use in government. It is more than a mere set of technologies. Gov 2.0 uses web-based technologies; has a social dimension built around communities and social networks; is based on user-generated and control of content; emphasizes providing and remixing of data from multiple sources; uses increased simplicity in design, and features participatory, decentralized models and processes [23, p. 276].

The paradigm shift enabled by Gov 2.0 technologies is from end-users consuming information and data to producing and controlling information and remixing data for

new purposes to facilitate interaction and collaborative work. According to O'Reilly 2010, "Gov 2.0 is not a new kind of government; . . . [it] is the use of technology—especially the collaborative technologies at the heart of Web 2.0—to better solve collective problems at city, state, national and international levels" [17]. O'Reilly regards Gov 2.0 as an open platform that allows people inside and outside of government to innovate and has the potential to change the relationships among all stakeholders with government [17].

A primary driver for implementing Gov 2.0 in the U.S. federal government is outlined in the presidential Memorandum on Transparency and Open Government issued on January 21, 2009. The goal of the presidential directive was to create a level of openness in government by establishing a system of transparency, participation, and collaboration [16]. To achieve this broad initiative federal government agencies are employing social media to integrate technology, social interaction, and content creation. Consequently, the use of social media in government is creating entirely new online communities that defy traditional communication reach and organizational boundaries.

This author: 1) Examines theoretical approaches to explain the adoption of social media in government; 2) Explores the relationship among social media and participation, collaboration and transparency; 3) Reviews initiatives of social media use in U.S. federal government agencies; 4) Identifies challenges and obstacles for implementing social media in government.

In addressing these four points, this paper first presents a three-pronged theoretical overview describing broad, overarching theories: Diffusion of Innovations, Social Influence and Collective Intelligence. These theories provide an appropriate theoretical background to explain the characteristics as well as variables that affect the adoption and the use of social media. Next, the methodology employed for this study is presented. Then definitions of the social media to be examined and their relationship with participation, collaboration and transparency are described. A review of U.S. federal agencies initiatives presents a picture of prevalent uses of social media to achieve transparency, participation and collaboration. In addition a framework for how to assess these tools to determine the effectiveness and efficiency for the delivery of public online services is provided. It is argued that understanding these developments highlights how public administrators can harness the potential of social media for open government.

## 2 Theoretical Overview

Three theoretical approaches, Diffusion of Innovations, Social Influence, and Collective Intelligence theories offer insights into the characteristics and variables for why and how these new information and communication technologies (ICT) play an important role and are adopted in public organizations.

The first theory applicable to understanding the adoption of social media is **Diffusion of Innovations**. This theory explains how and why new ideas and technologies spread through certain channels among members of a social system. In order for an individual or organization to decide to adopt an innovation, there is a four stage decision-making process.

- *Knowledge* – awareness of an innovation,
- *Persuasion* – a favorable or unfavorable attitude toward the innovation,
- *Decision* – deciding to adopt or reject the innovation, and
- *Confirmation* – seeking reinforcement for the decision [20, p. 103].

Five characteristics of the technology explain the rate of adoption, i.e. the speed by which members of a social system adopt an innovation:

- *Relative advantage* – amount of improvement over existing technology,
- *Compatibility* – ability to incorporate the technology into an individual's life,
- *Complexity* – how difficult will it be to adopt,
- *Trialability* – how easy is it to experiment with the technology on a limited basis, and
- *Observability* – how visible is the technology to other users [20, p. 22].

Once an individual or an organization proceeds through the decision-making process, recognizes the value added, the ease and degree of complexity of adoption as well as easily experimenting with the technology, and observing how visible the technology is to others will determine how quickly the technology is adopted. Five adopter categories are described:

- *Innovators* – first individuals to adopt, including risk takers and opinion leaders,
- *Early adopters* – second group to adopt and individuals have a high degree of influence on later adopters (e.g., opinion leaders),
- *Early majority* – cautious and slower participants in the adoption process,
- *Late majority* – adoption occurs after the majority of the group has already adopted the innovation, and individuals are skeptical, and
- *Laggards* – last ones to adopt an innovation and individuals appear to have an aversion to change in general [20, p. 182-184].

These four stages of decision making, five characteristics of the technology, and five categories of adopters certainly create a foundation to understand and explain how and why social media are adopted by public organizations. For example, public organizations at all levels have knowledge, i.e. awareness of these ICT, and at some levels have not only a favorable opinion of social media, but have been persuaded to decide to use for either a specific task or for a specific audience. Moreover, the rate of adoption in public organizations is probably tempered by the improvement over existing technologies to build relationships with a targeted audience, its high compatibility with an agency and/or individual's environment. It is relatively easy to experiment with and these ICT are highly visible to others which can also spur its adoption. At this point in time it seems reasonable that all governmental agencies experimenting with social media may be regarded as early adopters.

**Social Influence Theory** postulates that individuals' thoughts or actions are affected by other people. Three categories of social influence are identified:

- *Compliance* – individuals appear to agree with others, but disagree privately,
- *Identification* – individuals are influenced by a respected person, and
- *Internalization* – individuals accept a belief or behavior [14, p. 51].

Informational influence can be defined as the desire to accept information from others as evidence of supporting their own beliefs [8]. In online environments social influence also is depicted by:

- *Liking* – people are persuaded by others they respect or like, and
- *Social proof* – people will engage in activities that they observe others doing [6, p. 76].

Interpersonal attraction, the familiarity with a respected individual, who is either in close proximity or only “a mere exposure” may increase an individuals’ attraction to an idea or information and lead them to share this information with others in their

### 3 Methodology

In exploring the relationships between social media and collaboration, participation, and transparency in government, this study used a multi-pronged, iterative design strategy that involved conducting a literature review, content analysis of documents, and the analysis of web sites. The literature review informed the author about previous studies conducted to measure the relationship between the use of social media and collaboration, participation and transparency as well as surveys about citizens’ usage of social media to contact government. The reviews revealed specific initiatives, methods of assessment, as well as challenges and barriers for implementation. The content analysis of government documents such as legislation, government directives, project reports, best practices guidelines provided the background for establishing policies and appropriate metrics for the deployment of social media in government. Finally an analysis of web sites, specifically of the 25 federal agencies identified in the Open Government Flagship Initiatives, provided a basis for assessing the progress of using social media to achieve the goals of collaboration, participation, and transparency. This multi-pronged approach provided a broad array of perspectives of the use of social media by government agencies.

### 4 Social Media and Open Government

U.S. federal agencies are using a wide variety of social media: Blogs, micro blogs, RSS feeds, wikis, video and photo sharing, podcasting, social networking sites, social bookmarking sites, mashups, widgets and virtual worlds. Through these social media, people or groups can create, organize, edit, comment on, combine, and share content. These social media tools use the “wisdom of crowds” concept to collaboratively connect online information [10].

There are two primary reasons for adopting social media in government. The first is the uptake of the use of social media by citizens, businesses and non-profit organizations. For the past five years public and private organizations have been adopting Web 2.0 ICT incorporating social media. Consequently, U.S. citizens have increased their daily online Internet activities. By December 2009 the number of adults in the U.S. using the Internet was 74 percent, up from 63 percent reported in

2003 [19]. The types of online activities U.S. citizens are engaged in vary widely: From using e-mail (89%), buying products (75%), searching for news or information about politics or upcoming campaigns (60 %), visiting government web sites (59%), using social networking sites such as MySpace, Facebook or LinkedIn.com (47%), reading blogs (32%), sharing user generated content (30 %), and using Twitter or other status-update service (19%) [19]. Similarly, U.S. citizens who access government information and services (31% of online adults) are moving beyond government web sites and using new online platforms such as blogs, micro blogs, social networking sites, email, online video, and text messaging [21]. Consequently, U.S. citizens are moving online in their private lives and using multiple channels including social media to access government information and services as well.

The second reason for government taking a more aggressive approach to using social media is the presidential directive for open government [16]. This directive stipulates that government should be:

- Transparent -- providing citizens with information about what their government is doing; in a rapid and timely fashion in accordance with laws; and in easy to search and usable formats.
- Participatory -- provide citizens with increased opportunities to participate in policymaking and to provide their government with the benefits of their collective expertise and information.
- Collaborative – engage citizens in the work of their government, and to encourage executive departments and agencies to use innovative tools, methods, and systems to cooperate among themselves, across all levels of government, with nonprofit organizations, businesses and individuals in the public sector. Additionally, all departments and agencies were directed to solicit public feedback to assess and improve their collaborative initiatives [16, p. 1].

Consequently, U.S. federal government agencies are adopting social to encounter citizens in their online world; to increase the range of communication and dissemination of information to diverse and targeted audiences; to facilitate interactive communication and community; to put a human face on government and to put a human face on government [10]. Government can leverage the unique characteristics of social media to achieve the goals of the open government directive [5, 18]. This study examines the most widely used social media by federal government agencies to ascertain their relationship with the three tenants of open government.

## 5 Types of Social Media

Before examining the relationship among these tools with the goals of open government, it is important to describe each as well as potential uses in government. The fundamental Gov 2.0 tools/technologies selected are: Blogs, microblogs, mashups, podcasts, RSS feeds, social networking sites, video sharing, and wikis. These eight tools were selected because each demonstrates one or more of the basic concepts of Gov 2.0: Enhancing user generated content, extending the reach of



communications to new audiences, building relationships via social networks, creating collaborative environments with internal and external stakeholders, increasing stakeholder engagement, and combining content from multiple sources for integrated purposes.

The following provides a description of each tool and its use in government [10, 5, p. 11]:

- **Blogs:** Chronological journal entries made on a web site about a particular topic. Blogs are primarily text and images, and are an historical archive of the topic. Used to disseminate information quickly, particularly through (RSS); increases outreach; places a human face on the organizations.
- **Micro blogs** are another form of blogging, but created in a short text message style. Twitter is a prime example of a popular micro-blog service (with a limit of 140 characters). Micro blog sites include: Twitter.com, Jaiku.com, Tumblr.com, and Plurk.com. Short conversations enable information sharing; provides a means to engage a community; extend outreach; broadcast emergencies; post key events, update content (via hashtags); track flow of information as well as reactions to issues and situations.
- **Mashups:** A web application combining data from multiple sources to be integrated into a single integrated service. The data from either source were created for different purposes. Examples using mapping data are Microsoft Virtual Earth, Google Earth, and Google Maps.
- **Podcasts:** A way for publishing MP3 audio files downloaded to computers and mobile devices. Podcasts are an efficient means for the user to receive up-to-date information on topics of interest.
- **Really Simple Syndication (RSS):** A web content format (XML) that enables the owner of a web site to alert users to new information. RSS feeds are usually used for updating blogs, news headlines, or podcasts to users. Subscribers to RSS feeds can gather information from their favorite web sites in one location without browsing and searching for it manually. This is a fast way to disseminate updated content from multiple sources to end-users. It increases awareness of government information; improves communication.
- **Social Networking sites:** Web sites designed to connect people in online communities. Participants are registered users who are allowed to interact with other users for social or professional purposes. Examples of social networking sites are MySpace, Facebook, and LinkedIn. All of these social networking sites enable users to create networks of contacts. Used to promote government services and information, for recruitment, and announce events.
- **Video Sharing:** Use videos, images and audio libraries to share information by enhancing communication with on-line audiences. YouTube is a prime example of this social media tool. Used for education and training, public outreach to new audiences, improves service delivery, and is a compelling channel to disseminate content. Government agencies are adding audio and video files to existing text as a way to enhance the message and increase awareness of the information

- **Wikis:** Collaborative publishing technology enabling multiple users to work on and publish documents online. Participants from different locations can contribute and modify existing documents (usually with editorial control). Documents are archived at a central location and users can access the documents by using hyperlinks. Used for knowledge sharing, cross boundary cooperation, engagement of contributors, increases transparency, increases project and time management, archives discussions and serves as a community repository.

## 6 Social Media and Open Government

The social media described above have been adopted widely by U.S. federal government agencies to achieve the goal of the open government directive. In addition, a progress report was issued in December 2009 outlining many of the initiatives undertaken by various agencies to achieve transparency, participation and collaboration [22]. For transparency, i.e. providing citizens with open data and information about what government is doing, this report lists the following efforts: For stimulus spending (Recovery.gov); general expenditures general (usaspending.gov); federal dashboard of information technology (IT) budgets (it.usaspending.gov); and entrepreneurship.gov and business.gov provide information on job creation and businesses. The primary purpose of Data.gov is to provide citizens usable data sets from various agencies that can be used in a variety of ways. In only one year data.gov has made available 272,677 datasets, 236 new applications have been created from Data.gov datasets, and eight cities and eight states have developed their own data sites [7]. Many agencies such as Department of Agriculture, Center for Disease Control, Department of Transportation provide high value data to improve the lives of citizens and each of these agencies use a variety of social media to share government data and information [22, p. 3]. Through a partnership with outside entities, the Federal Register, which records and archives the legislation and rules and procedures of government, is now available for download in a machine readable format. Fedthread.org is a web site that enables citizens to annotate the Federal Register and GovPulse (<http://govpulse.us/>) enables one to search by topic or location [22]. All of these web sites enable people to access government information and data and to use it for their own purposes.

To encourage participation, government agencies are launching blogs, Twitter, social networking sites, RSS, podcasts and videos through which citizens can leave comments and raise issues and provide feedback. A major initiative, IdeaScale, was launched by 23 agencies to solicit online public feedback on their open government plans. This feedback tool was open from February 6 to March 19, 2010 and recorded more than 1,400 ideas, 3,200 comments, and 32,000 votes from more than 6,400 users [1]. To see the results of these dialogues one can visit OpenGov Tracker at <http://www.opengovtracker.com/> and the Open Government Data and Reports available at [http://www.usa.gov/webcontent/reqs\\_bestpractices/open/data\\_reports.shtml](http://www.usa.gov/webcontent/reqs_bestpractices/open/data_reports.shtml) provides the raw data generated for each report. Other examples of engaging citizens and employees to participate are reflected by Open for Questions which allowed U.S. citizens the opportunity to ask questions about the economy and the Health IT Online Forum

enabled health care stakeholders to provide suggestions on improving health care through IT. Some agencies are launching competitions and contests to engage citizens and to disseminate information.

If participation efforts are designed to engage citizens and employees, then collaboration focuses on solving problems by engaging people to work together to offer solutions. Examples include the Department of Health and Human Services campaign to reach teens about the H1N1 flu by hosting video contests for the best H1N1 rap song. The National Lab Day effort created collaboration between government agencies and private and professional organizations to build communities of support for science, technology, engineering and math teachers [22]. Some examples of government agencies using wikis for internal collaborative work are: Intellipedia (U.S. intelligence community), Diplopedia (U.S. foreign affairs agencies within the State Department), OMB MAX Federal Community (Executive Branch personnel to collaborate on budget issues), and USGSA's USA Services Intergovernmental Collaborative Work Environment (incubator space for 20 intergovernmental communities). The OMB established a public Wiki (OMB USAspending.Gov) for the public to comment on the Federal Funding and Transparency Act. The U.S. Court of Appeals for the 7<sup>th</sup> Circuit, Practitioner's Handbook is a wiki that can be viewed by the public, but only edited by members of the bar. Wikis are seen as useful for government agencies to increase knowledge sharing, crossing boundaries within and across agencies as well as geographical boundaries, to engage more people in the process, transparency, and for more efficient project management. It is regarded as a very flexible channel for fostering collaborative work within government and with the community [11, 13]. As indicated by these multiple examples of using social media for transparency, participation, and collaboration, it becomes apparent that one size does not fit all and that each organization should select social media designed to be aligned with their mission and goals [15] and citizens' needs. Now that the uses of social media in Gov 2.0 have been examined the next step is to identify challenges and obstacles for implementing social media in government.

## **7 Are We There Yet? Challenges and Obstacles of Social Media**

No, but significant, first steps have been taken to begin to change the perspective of government employees about the relationships between government and stakeholders. In 2008 the Federal Web Managers Council published a whitepaper on the perceived and real barriers to implementing social media in government and to propose solutions [12]. One of the primary barriers at that time was that agencies perceived the use of social media as a technology issue and not as a communication tool that could extend the communication reach of organizations to new, targeted audiences. Other concerns centered around legal issues with third party vendors such as Twitter, Facebook, advertising, procurement, privacy, surveys, access, and administrative requirements for how agencies can communicate with the public during rulemaking [12]. Also in 2008 The Federal Web Managers Council published a whitepaper "Putting Citizens First: Transforming Online Government" that outlined six strategies to reform how the U.S. government delivers online services and information:

- Establish web communication as a core government business function,
- Help the public complete common government tasks efficiently,
- Clean up irrelevant and outdated content so people can find what they need online,
- Engage the public in a dialogue to improve customer service,
- Deliver the same answer from every service channel,
- Ensure underserved population can access critical information online [3].

The 2010 annual report by the Federal Web Managers Council recognized numerous accomplishments over the past two years, but more importantly that progress was made toward achieving its top priority of helping the public complete their top task easily. In addition, the Council noted that it supported participatory government by assisting 24 agencies to open online dialogues engage citizens to share ideas and to help shape an agency's open government plan. Although progress has been made, the Council noted several obstacles to achieving its goals:

- Web communications is still not managed as core business function,
- Top tasks are still difficult to complete,
- Web sites are still too cluttered,
- Public dialogues are not used to improve customer service,
- Consistent answers are not provided through all channels,
- Underserved populations are not a priority,
- Performance measurement is not consistent, and
- Need for resources to support open government [4, pp. 2-3].

While the Federal Web Managers Council is committed to overcoming these barriers, its top priority for the coming year is to improve government communication through all channels by delivering content that is clearly written, understandable, and engaging. This should help government agencies to send consistent messages through all channels and help the public to be able to complete their top tasks [4, p.3]. The Council is also committed to enable citizen participation, and to institutionalize the concepts of transparency and accountability,

One of the most striking aspects of the Council's plan is the focus on the needs of the end-user to be able to complete its top tasks. End-users, particularly online adults, are beginning to use digital channels to gather government information. A Pew Internet survey reported that 31 percent of Internet users did one of the following online activities in the last 12 months:

- 15% watched a government video,
- 15% signed up for received e-mail alerts,
- 13% read a government agency or official blog,
- 5% became a fan of a government agency or official,
- 4% of cell phone users received text messages from a government agency or official, and
- 2% followed a government agency or official on Twitter [21, p. 26].

Moreover, these government social media users reported that they used a broad array of online and offline platforms to accomplish their tasks. While 95% of these government social media users visited one or more web sites, 66% of these same respondents also contacted a government agency by phone, in person, or letter [21, p. 27]. Also these online government users, i.e. those who visited one or more government web sites, reported that they are more likely to use other social media tools to gather information. The most frequently used tool was video (31%), followed by email alerts (27%), and government blogs (25%) [21, p. 27]. These findings may indicate that more frequent government web site visitors are exposed to different platforms for communicating with government. When asked about their attitudes toward online government services, the respondents reported that providing general information (67%), contacting public officials via web sites (62%), and completing tasks online (62%) were most important. Posting information on social networking sites or alerts via Twitter were less important [21, p. 34]. The survey also measured respondents' attitudes of government using social media to engage citizens. Seventy-nine percent reported that having the ability via social media to follow and communicate online with government helps people be more informed about what government is doing (transparency). Seventy-four percent felt that social media made government agencies and officials more accessible. Internet users perceived using social media for engagement more favorably than non-users [21, p. 37].

The findings from the Pew Internet survey seem to be consistent with the goals of the Federal Web Content Managers Council. The Council emphasizes clear, consistent, understandable communication via all channels, and the survey finds that even frequent government online users also use multiple channels to contact government. Secondly, government online users want to gather information and complete tasks online, and web content managers realize that online government information is still too cluttered and that it is still difficult to complete a task. Finally, the emphasis on using social media for engagement and transparency seems to be perceived favorably by Internet users which support the continuation of these efforts.

## 8 Conclusions

Gov 2.0 applications and social media have signaled a paradigm shift in the evolution of e-government services. Collaborative technologies are being used to engage citizens, to foster co-production and encourage transparency in government. Social media may be forging new relationships between government and stakeholders. For government agencies the focus is shifting from the collector and repository of information and data to sharing and collaborating with end-users to solve problems at all levels of government. To understand this shift in perspective, the theoretical background detailed above helps to explain the adoption and use of social media by government and citizens. First, from the Diffusion of Innovations theory one recognizes that as knowledge and awareness of social media expand, then the willingness to use these digital tools increases. This is supported by the Pew Internet survey findings that the more exposure to online Internet government information and services by the public, the more favorable are the public's attitudes toward using other social media. Moreover, the more favorable attitudes Internet users have toward social media, the more likely they are to decide to use the medium and seek confirmation for

their decision. The rate of adoption both by government and citizens can be explained that social media tools provide an improvement over existing technologies and are relatively inexpensive to achieve communication goals and the goals of open government of transparency, participation and collaboration. In addition, social tools are easy to adopt and are compatible with existing technologies. Both government agencies and citizens can experiment with these technologies and the use of them is easily observable by others. Some government agencies may be considered early adopters, i.e. opinion leaders, while others are more cautious and fall into the category of early majority. In the online environment, both citizens and government agencies have been influenced by others they respect and who are engaging with social media successfully. Finally, a primary component of social media is the use for collaborative work which builds on collective intelligence to enable cooperation. Government agencies are using social media, both internally and externally, to solve problems.

After reviewing social media use in government, one finds that social media are indeed being used extensively for transparency, participation, and collaboration. Social media are being used to build relationships with stakeholders by engaging them in conversations and soliciting their ideas on issues and projects. By making public data readily available and usable, citizens can have a clearer picture of what their government is doing and this impacts their perception of transparency. Moreover, through collaborative projects, citizens and government employees can impact government policies. The roles of social media in government will develop as usage increases and expectations of end-users increases as their familiarity with the technologies evolve. Social media become platforms for interacting with citizens. Most importantly, it is not about the collaborative technologies, it is about the communication and relationships with the stakeholders that is at the heart of this paradigm shift.

## References

1. Anonymous. Open Government Public Engagement Tool. U.S. Government Service Administration (April 12, 2010),  
[http://www.gsa.gov/Portal/gsa/ep/contentView.do?contentType=GSA\\_BASIC&contentId=29495](http://www.gsa.gov/Portal/gsa/ep/contentView.do?contentType=GSA_BASIC&contentId=29495)
2. Atlee, T., Por, G.: Collective intelligence as a field of multi-disciplinary study and practice (2000),  
<http://www.community-intelligence.com/files/Atlee%20-%20Por%20%20CI%20as%20a%20Field%20of%20multidisciplinary%20study%20and%20practice%20.pdf>
3. Campbell, S., Flagg, R.: Putting Citizens First: Transforming Online Government (November 2008),  
[http://www.usa.gov/webcontent/documents/Federal\\_Web\\_Managers\\_WhitePaper.pdf](http://www.usa.gov/webcontent/documents/Federal_Web_Managers_WhitePaper.pdf)
4. Campbell, S., Flagg, R.: Putting Citizens First: Transforming Online Government 2010 Progress Report (April 2010),  
[http://www.usa.gov/webcontent/documents/2010\\_FWMC\\_AnnualProgressReport.pdf](http://www.usa.gov/webcontent/documents/2010_FWMC_AnnualProgressReport.pdf)
5. Chang, M., Kannan, P.: Leveraging Web 2.0 in government. IBM Center for the Business of Government (2008),  
<http://www.businessofgovernment.ort/pdfs/ChangReport2.pdf>

6. Cialdini, R.B.: The science of persuasion. *Scientific American* 284, 76–81 (2001)
7. Data. Gov. (June 14, 2010), <http://www.data.gov/>
8. Deutsch, M., Gerard, H.B.: A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology* 51, 629–636 (1955)
9. Flew, T.: *New Media an Introduction*. Oxford University Press, Melbourne (2008)
10. Godwin, B. *Open Government, Transparency, and Social Media: A slide presentation* (April 1, 2009), <http://www.usa.gov/webcontent/resources/previousnewmediatalks.html>
11. Godwin, B.: *Matrix of Web 2.0 technology and government. USA.gov and Web best practices*, GSA Office of Citizen Services (2008), [http://www.usa.gov/webcontent/documents/Web\\_Technology\\_Matrix.pdf](http://www.usa.gov/webcontent/documents/Web_Technology_Matrix.pdf)
12. Godwin, B., Campbell, S., Levy, J., Bounds, J.: *Social media and the federal government: Perceived and real barriers and potential solutions*. Federal Web Managers Council (December 2008), [http://www.usa.gov/webcontent/documents/SocialMediaFed%20Govt\\_BarriersPotentialSolutions.pdf](http://www.usa.gov/webcontent/documents/SocialMediaFed%20Govt_BarriersPotentialSolutions.pdf)
13. Godwin, B., Campbell, S., Levy, J., Bounds, J.: *Examples of agencies using online content and technology to achieve mission and goals*. Federal Web Managers Council (2008), <http://www.usa.gov/webcontent/documents/ExamplesofUsingTechnologyandContenttoAchieve%20Agency.pdf>
14. Kelman, H.: Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution* 1, 51–60 (1958)
15. Meijer, A., Thaens, M.: Alignment 2.0: Strategic use of new internet technologies in government. *Government Information Quarterly* 27, 113–121 (2010)
16. Obama, B.: *Memorandum for the Heads of Executive Departments and Agencies: Transparency and Open Government* (January 21, 2009), [http://www.whitehouse.gov/the\\_press\\_office/Transparency\\_and\\_Open\\_Government/](http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/)
17. O'Reilly, T.: *Government as a Platform*. In: Lathrop, D., Ruma, L.R.T. (eds.) *Open Government: Transparency, collaboration and Participation in Practice*, pp. 1–23. O'Reilly Media, Inc., Sebastopol (2010), <http://opengovernment.labs.oreilly.com/cho1html>
18. Osimo, D.: *Web 2.0 in government: Why and how?* European Commission Joint Research Center (2008), <http://ftp.jrc.es/EURdoc/JRC45269.pdf>
19. Pew Internet and American Life Project. *Generations online in 2009* (December 2009), <http://www.pewinternet.org>
20. Rogers, E.M., Shoemaker, F.: *Communication of Innovations: A Cross-Cultural Approach*, 2nd edn. Free Press, New York (1971)
21. Smith, A.: *Government Online: The Internet gives citizens new paths to government services and information* (April 27, 2010), <http://www.pewinternet.org>
22. White House. *Open government: A progress report to the American people* (2009), <http://www.whitehouse.gov>
23. Wigand, R.T.: *Web 2.0: Disruptive technology or is everything miscellaneous?* In: Huizing, A., de Vries, E.J. (eds.) *Information Management: Setting the Scene*, pp. 269–284. Elsevier Scientific Publishers, Amsterdam (2007)

# MATAWS: A Multimodal Approach for Automatic WS Semantic Annotation

Cihan Aksoy<sup>1,2</sup>, Vincent Labatut<sup>1</sup>, Chantal Cherifi<sup>1,3</sup>, and Jean-François Santucci<sup>3</sup>

<sup>1</sup> Galatasaray University, Computer Science Department, Ortaköy/İstanbul, Turkey

<sup>2</sup> TÜBİTAK, Gebze/Kocaeli, Turkey

<sup>3</sup> University of Corsica, Corte, France

caksoy@uekae.tubitak.gov.tr

**Abstract.** Many recent works aim at developing methods and tools for the processing of semantic Web services. In order to be properly tested, these tools must be applied to an appropriate benchmark, taking the form of a collection of semantic WS descriptions. However, all of the existing publicly available collections are limited by their size or their realism (use of randomly generated or resampled descriptions). Larger and realistic syntactic (WSDL) collections exist, but their semantic annotation requires a certain level of automation, due to the number of operations to be processed. In this article, we propose a fully automatic method to semantically annotate such large WS collections. Our approach is multimodal, in the sense it takes advantage of the latent semantics present not only in the parameter names, but also in the type names and structures. Concept-to-word association is performed by using Sigma, a mapping of WordNet to the SUMO ontology. After having described in details our annotation method, we apply it to the larger collection of real-world syntactic WS descriptions we could find, and assess its efficiency.

**Keywords:** Web Service, Semantic Web, Semantic Annotation, Ontology, WSDL, OWL-S.

## 1 Introduction

The semantic Web encompasses technologies which can make possible the generation of the kind of intelligent documents imagined ten years ago [1]. It proposes to associate semantic metadata taking the form of concepts with Web resources. The goal is to give a formal representation of the meaning of these resources, in order to allow their automatic processing. The process of defining such associations is known as semantic annotation (or annotation for short), and generally relies on libraries of concepts collectively described and structured under the form of ontologies. The result is Web documents with machine interpretable mark-up that provide the source material for software agents to operate. The annotation of Web resources is obviously fundamental to the building of the semantic Web.

According to Nagarajan and Uren *et al.*, in order to properly treat documents, annotating systems must follow a generic process [2] and meet seven different requirements [3]. The annotation process is composed of three primary steps that are



the identification of the entity to be annotated, its possible disambiguation and its association to a concept. The requirements are as follow. The first one is to use standard formats (R1). Indeed, they provide a bridging mechanism that allows the access to heterogeneous resources and collaborating users and organizations to share annotations. The second one is to provide a single point of entry interface (R2), so that the environment in which users annotate documents is integrated with the one in which they create, read, share and edit them. The third one is to support multiple ontologies and to cope with changes made to ontologies (R3). This last point ensures consistency between ontologies and annotations with respect to ontology changes. The fourth and the fifth requirements are related to the document to be annotated. An annotating system must support heterogeneous input formats (R4), and be able to manage the annotation consistency when the document evolves (R5). The sixth requirement is about the annotation storage (R6), for which two models are proposed: the annotations can be stored separately from the original document or as an integral part of the document. Seventh, and finally, as manual semantic annotation leads to a knowledge acquisition bottleneck, the last requirement deals with the automation of the annotating process (R7). Automated annotation provides the scalability needed to annotate existing documents on the Web, and reduces the burden of annotating new documents.

Besides static Web content such as textual or multimedia documents, semantic annotation also concerns dynamic content, and more particularly Web Services (WS). WS are non-static in nature; they allow carrying out some task with effects on the Web or the real-world, such as the purchase of a product. The semantic Web should enable users and agents to discover, use, compose, and monitor them automatically. As Web resources, classic WS descriptions such as WSDL files can be semantically enhanced using the annotation principle we previously described, i.e. by the association of various ontological concepts. However, due to the particular structure of WS descriptions, these associations must comply with very specific constraints, which are different from those encountered for other kinds of Web resources such as Web pages. [2]. Indeed, the semantics associated with WS need to be formulated in a way that makes them useful to the application of WS. Sheth presents four types of semantics for the complete life cycle of a Web process: data, functional, non-functional and execution [4]. *Data* semantics is related to the formal definition of data input and output messages. *Functional* semantics is related to the formal definition of WS capabilities. *Non-functional* semantics is related to the formal definition of constraints like QoS. *Execution* semantics is related to the formal definition of execution flows at the level of a process or within a WS. Semantically annotating a WS implies describing the exact semantics of the WS data and functionality elements, which are crucial for the use of the WS, as well as its non-functional and execution elements.

Efforts for WS annotation include WS semantic languages as well as tools to annotate legacy WSDL files. The most prominent semantic languages are OWL-S [5], WSMO [6], WSDL-S [7] and SAWSDL [8]. While OWL-S and WSMO define their own rich semantic models for WS, WSDL-S and SAWSDL work in a bottom-up fashion by preserving the information already present in WSDL. Those description languages are used in many research projects focusing on various semantic-related applications like automatic discovery and composition. In order to test these applications, one needs a benchmark, i.e. a large collection of annotated WS [9]. Such

collections exist, but are limited in terms of size, realism, and representativity. These limitations are due to the fact the annotation process is generally performed manually, and is therefore costly. The use of an appropriate annotation tool can help decrease this cost, especially if it is automated. However, because of the specific structure of this kind of document, automatically annotating a WS description is much different, from the natural language processing perspective, than annotating other Web documents such as plain text. It consequently requires to perform a particular form of text mining, leading to dedicated tools such as ASSAM [10] or MWSAF [11]. But those tools also have their own limitations, the main one being they are only partially automated and require human intervention, which is a problem when annotating a large collection of WS descriptions.

In this paper we present the first version of MATAWS (Multimodal Automatic Tool for the Annotation of WS), a new semantic WS annotator, whose purpose is to solve some of these limitations. MATAWS was designed with the objective of batch annotating a large collection of syntactic descriptions and generating a benchmark usable to test semantic-related approaches. It focuses on data semantics (i.e. the annotation of input and output parameters) contained in WSDL files, and currently generates OWL-S files (other output formats will shortly be included). Our main contributions are: 1) a full automation of the annotation process and 2) the use of a multimodal approach. We consider not only the parameter names, but also the names present in the XSD types used in the WSDL descriptions: type names, and names of the fields defined in complex types.

The rest of this paper is organized as follows. Section 2 presents both existing ways of retrieving a collection of semantic WS descriptions: recover a publicly available collection and annotate a syntactic collection using one of the existing annotation tools. In section 3, we introduce MATAWS and describe our multimodal approach. In section 4 we apply MATAWS to the annotation of a publicly available collection of syntactic WS descriptions. Finally, in section 5 we discuss the limitations of our tool and explain how we plan to solve them.

## 2 Solutions to Access an Annotated Collection

When looking for a collection of semantic WS descriptions, one can consider two possibilities: either using a predefined collection, or creating his own one. In this section, we first review the main existing collections and their properties. The creation of a collection can be performed either by using a random model to generate artificial descriptions, or by semantically annotating a collection of real-world syntactical descriptions. The usual goal when looking for a semantic collection is to test WS-related tools on realistic data. To our opinion, the WS collections properties are not known well enough to allow the definition of a realistic generative model, which is why we favor the second solution. For this reason, in the second part of this section, we also review the main tools allowing to annotate WS descriptions.

### 2.1 Collections of Semantic Descriptions

The main publicly available collections of semantic WS are those provided by the ASSAM WSDL Annotator project, SemWebCentral and OPOSSum. Their major features are gathered in Table 1.

The ASSAM WSDL Annotator project (Automated Semantic Service Annotation with Machine learning) [12] includes two collections of WS descriptions named *Full Dataset* and *Dataset2*. *Full Dataset* is a collection of categorized WSDL files, which contains 816 WSDL files describing real-world WS. *Dataset2* is a collection of OWL-S files, obtained by annotating a subset of the WSDL files using the ASSAM Annotator (cf. section 2.2). 164 descriptions were fully labeled, assigning ontology references to the WS itself, its operations and their inputs and outputs.

**Table 1.** Collections of semantic WS descriptions: main features

Name	Dataset2	OWLS-TC3	SAWSDL-TC	SWS-TC
Source	ASSAM project	SemWeb Central	SemWeb Central	SemWeb Central
Type	Real-world descriptions	Real-world descriptions, partially resampled	Real-world descriptions, partially resampled	N/A
Language	OWL-S	OWL-S	SAWSDL	OWL-S
Annotated Type	Data, Functional	Data	Data	Data
Size	164	1007	894	241
Particular features	Processed using Assam annotator	Single interface, one operation per service	Single interface, one operation per service	N/A

SemWebCentral [13] is a community whose purpose is to gather efforts from people working in the semantic Web area. Three semantic collections are available: *OWLS-TC* (OWL-S Test Collection), *SAWSDL-TC* (SAWSDL Test Collection) and *SWS-TC* (Semantic WS Test Collection). OWLS-TC3 is the third version of this test collection. It provides 1007 semantic descriptions written in OWL-S from seven different domains. Part of the descriptions were retrieved from public IBM UDDI registries, and semi-automatically transformed from WSDL to OWL-S. SAWSDL-TC originates in the OWLS-TC collection. It was subsequently resampled to increase its size, and converted to SAWSDL. The collection provides 894 semantic WS descriptions. The descriptions are distributed over the same seven thematic domains than OWLS-TC. SWS-TC is a collection of 241 OWL-S descriptions. There is not much information about this collection.

OPOSSum (Online PORTal for Semantic Services) [14] is a joint community initiative for developing a large collection of real-world WS with semantic descriptions. Its aim is to create a suitable test bed for semantically enabled WS technologies. OPOSSum gathered the three semantic collections of SemWebCentral, plus the *Jena Geography Dataset* collection, explicitly collected within OPOSSum. The collection contains 201 real-world WS descriptions retrieved from public. All the described WS belong to the domains of geography and geocoding. Unfortunately, for now, no semantic descriptions are available for the services of the Jena Geography Dataset, which is why this collection is absent from Table 1.

These collections have been widely used in semantic WS-related works [15, 16]. As shown in Table 1, they all focus on the annotation of the data elements, which corresponds to our objective. However, one can notice some limitations. SWS-TC

description is insufficient, it is not even clear if the WS descriptions are real-world. Dataset2 contains only real-world WS descriptions but it is very small, which can raise questions about its representativity. On the contrary, OWLS-TC3 and SAWSDL-TC contain a substantial number of descriptions. Nevertheless, these have been partially resampled in an undocumented way, which raises important questions concerning their realism.

## 2.2 Annotation Tools

From our point of view, WS annotation is considered as a one-time task, aiming at annotating legacy WS, which are described only syntactically. Newly created or modified WS should be (re)annotated manually by their authors, which is much preferable in terms of quality than any automatic processing. For this reason, and due to the specific nature of WS annotation, we are not concerned by all the 7 requirements stated by Uren *et al.* [3] for general annotation tools. It is of course necessary to use standard formats for input and output (R1). A polyvalent environment is not necessary, since we do not want to modify existing descriptions or create any new ones (R2). The support of multiple or changing ontologies is relevant (R3), but it is not the most important point, so we chose to ignore it in this first work. The input format is constrained to WSDL (R4), since it is the *de facto* standard for syntactical WS description. As stated before, we do not plan to maintain annotations if WS are modified (R5). The model of annotation storage (R6) is constrained by the output format: separate form for OWL-S and integrated for WSDL-S and SAWSDL. Finally, the level of automation is of great interest to us, given our context (R7).

Only a few publicly available tools exist to semantically annotate WS descriptions. Table 2 presents the main ones and summarizes their properties. They all take a set of WSDL files as input (R1 and R4), but differ on several properties such as their level of automation (R7) and the language used to output the semantic descriptions (R1). The tools are described in details in the rest of this subsection.

*Radiant* is an open source tool created at the Georgia University [17]. It takes the form of an Eclipse plug-in and can output both SAWSDL and WSDL-S files. It provides a GUI which presents the elements constituting the WS description and allows to select the concepts one wants to associate to parameters or operations, by browsing in the selected ontologies. This interface makes the annotation process easier, but the annotation is nevertheless fully manual.

*ASSAM* is an open source Java program developed at the University College Dublin [12], able to output OWL-S files. It provides assistance during the annotation process. First, the user starts manually annotating parameters and/or operations using an existing ontology. Meanwhile, *ASSAM* identifies the most appropriate concepts using machine learning methods. After enough information has been provided, the software is able to propose a few selected and supposedly relevant concepts when the user annotates a new WS.

*MWSAF* is another open source Java tool created at the Georgia University [11]. It outputs WSDL-S files, and like *ASSAM* it has a machine learning capability allowing it to assist the user during the annotation process. It is able to annotate not only parameters and operations, but also non-functional elements.

*WSMO Studio* is an Eclipse plug-in initially designed to edit semantic WS based on the *WSMO* model. An extension allows annotating WS parameters and operations, and outputting the result under the form of *SAWSDL* files [18]. However, the tool does not provide any assistance to the user and the process is fully manual.

**Table 2.** WS Semantic annotation tools and their properties

Name	Output Format	Annotated Type	Automation	Last Update
Radiant	SAWSDL, WSDL-S	Data, Functional	Fully manual	May 2007
ASSAM	OWL-S	Data, Functional	Assisted	May 2005
MWSAF	WSDL-S	Data, Functional, Non-Functional	Assisted	July 2004
WSMO Studio	SAWSDL	Data, Functional	Fully manual	Sept. 2007

Besides these annotation tools, several softwares allow to convert WSDL files to OWL-S files, but without performing any semantic annotation: they only apply a syntactic transformation and present the information contained in the original WSDL file under a form compatible with the OWL-S recommendation. *WSDL2OWLS* is an open source Java application created at the Carnegie Mellon University [19]. *OWL-S Editor* is a plug-in for Protégé (itself an ontology development environment) created at SRI [20]. Another software performing the same task is also called *OWL-S Editor*, but was developed at Malta University [21].

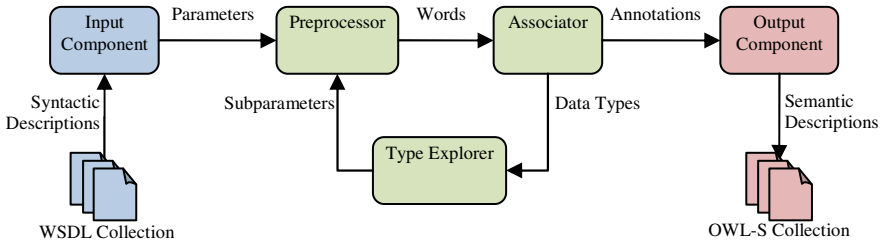
From this review, we can conclude the existing annotation tools present various limitations relatively to our goals. First, from a practical perspective, some of these tools are old and not supported anymore, which can cause installation and/or use problems. For instance, Radiant and ASAM are not compatible anymore with the current versions of some of the Eclipse plug-ins, libraries or API they rely on; meanwhile MWSAF installs and runs fine, but generates files without any of the annotations defined by the user. More importantly, these tools require important human intervention: Radiant and WSMO Studio are fully manual, whereas ASSAM and MWSAF only assist the user, after a compulsory learning phase. This justifies the development of our own tool, which we present in the next section.

### 3 Proposed Annotation Method

The absence of an existing solution fulfilling our needs compelled us to develop our own tool to semantically annotate WS descriptions. The main differences with the other annotation tools are the exploitation of several sources of information and the automation of the annotation process. In this section, we first describe the general architecture of our tool, which is made up of several independent components. We then focus separately on the components of interest, explaining their design and functioning.

### 3.1 General Architecture

MATAWS takes a collection of WSDL files as input and generates a collection of OWL-S files as output. Fig. 1 gives an insight of its modular structure, which includes five different components. Among these components, two are using external APIs (Associator and Output Component), whereas the three remaining ones were developed by us in Java. The Input and Output components are not of great interest with regards to the topic of this article, which is why we describe them shortly here. The other components are described in details in the following subsections.



**Fig. 1.** Architecture of MATAWS

The *Input Component* is in charge for extracting the set of all operation parameters defined in the considered collection of WSDL files. We designed a parser able (among other things) to retrieve the parameter names, type names and type structures (in the case of complex types) [22]. The *Output Component* is used after the annotation process to generate a collection of OWL-S files corresponding to annotated versions of the input WSDL files. For this purpose, we selected the Java *OWL-S API*, which provides a programmatic read/write access to OWL-S service descriptions [5]. Note we plan to add support for WSDL-S and SAWSDL by using other appropriate APIs.

The three remaining components correspond to the core of the annotation process. After the input component has parsed the WSDL files, it fetches parameters information to the *Preprocessor*. This one originally focuses on the parameter names, decomposing, normalizing and cleaning them so that they can be treated by the *Associator*. This component is based on the inference engine Sigma [23], whose role is to associate an ontological concept to a word. If Sigma is successful and manages to return a concept, this one is associated to the considered parameter. After all the parameters of a WS have been annotated, the *Output Component* is used to generate an OWL-S file with both the information contained in the original WSDL file and the selected concepts. However, for various reason explained later, it is not always possible for Sigma to find a suitable concept for every parameter. In this case, the *Type Explorer* accesses some properties related to the parameter data type, to obtain what we call subparameters. These are then fetched to the Preprocessor and the core processing starts again. In case of repeated annotation failure, this process can be repeated recursively until success or lack of subparameters.

### 3.2 Preprocessor

In order to work properly and propose a suitable concept, the Associator needs to process clear and normalized words. However, the names defined in real-world WS certainly do not meet this criterion. First, the meaning of an operation, parameter or type can hardly be described using a single word. For this reason, most names are made up of several concatenated words, separated either by alternating upper and lower cases or by using special characters such as dots, underscores, hyphens, etc. Second, sometimes the result is too long and abbreviations are used instead of the complete words. Finally, an analysis of any collection quickly shows different additional characters such as digits or seemingly useless separators can also appear.

Of course, there is no way to define an exhaustive list of the various forms a name can take in a WS description, but WS programmers actually follow only a few conventions, which allows performing very efficient preprocessing by applying a set of simple transformations to break a name into usable words. We distinguish three steps during name preprocessing: decomposition, normalization and filtering.

**Table 3.** Preprocessing examples

Transformation	Original Name	Extracted Words
Decomposition	WhiteMovesNext	White, Moves, Next
Decomposition	Number3Format	Number, Format
Decomposition	AUsername	Username
Decomposition	User_name	User, name
Normalization	no	number
Normalization	Password	password
Filtering	Parameter	-
Filtering	Body	-

The decomposition consists in taking advantage of the different types of concatenations we identified to break a name into several parts. It also involves some cleaning, in the sense all characters which are not letters are removed and diacritical marks are deleted. Table 3 shows some examples involving case alternation, and digit and underscore used as separators.

The normalization role is first to provide the Associator a clean version of the word, typographically speaking, by setting each word to lower case. Moreover, the normalization handles abbreviations, by replacing them with the corresponding full-length words. Table 3 gives an example of the name `no` being replaced by the word *number*. However, this last task is very context-dependent, because some strings are both full words and common abbreviations. For instance, `no` could simply mean the opposite of “yes”, used to negate the following concatenated word, e.g. `no_limit`. For this reason, human intervention can be necessary to set up this preprocessing, and adapt it to the considered collection. We chose to allow the user to define a list of common abbreviations.

Finally, we added a filtering step to deal with stop-words, i.e. words with no particular semantic information relatively to their context. For instance, the string `parameter` commonly appears in parameter names, without bringing any significant information, since the syntax of the WSDL file already allows to know if a certain

name points out at a parameter. For this reason, it can be considered as noise and ignored. Even more than before, the nature of the stop-words is closely linked to the domain of application, and requires human intervention to adapt the list of stop-words we defined.

Let us consider as an example the preprocessing of the name `ASessionId_02`. First it will be broken down to the words *A*, *Session* and *Id* while the numeric end of the name (`02`) will be ignored. The normalization step will transform them in *a*, *session* (lowercase) and *identity* (replacing an abbreviation). Finally, the filter will remove the article *a*, because it is a stop-word. Eventually, for this name `ASessionId_02`, the Preprocessor will output the two words *session* and *identity*.

### 3.3 Associator

As mentioned before, we use an existing tool called Sigma to associate a concept to a word. It is written in Java and allows to create, test, modify and infer ontologies [23]. It is provided with the Suggested Upper Merged Ontology (SUMO), which (unlike its name suggests) contains also mid-level and domain ontologies [24]. SUMO is free, covers a wide range of fields, and it has been mapped to the whole WordNet lexicon [25]. It was initially defined using the SUO-KIF language [26], and it is currently being converted in OWL [27].

**Table 4.** Concept association examples

Word	SUMO Concept associated by Sigma
buffalo	HoofedMammal
school	EducationalProcess
talk	Communication

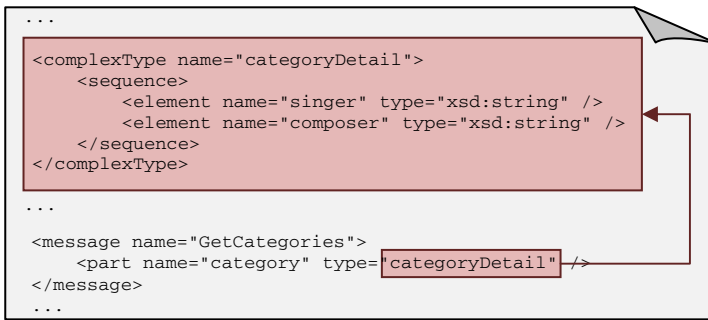
Although its main purpose is to work on ontologies, Sigma also offers a programmatic access to this mapping under the form of a method taking an English word as input and outputting a SUMO concept. Table 4 gives a few examples of such associations. The names we processed are most of the time not plain English words, which justifies our preprocessing.

### 3.4 Type Explorer

Although our focus is primarily on parameter names, we described the two previous components in general terms, because they can be applied to any kind of names. Indeed, different difficulties can arise, making it impossible to associate a concept to a parameter name. First, the Preprocessor might fail to break the name down to relevant words, hence fetching the Associator strings it cannot map to appropriate concepts. Second, the Preprocessor might filter all the words resulting from the name decomposition, meaning it will not be able to provide the Associator any word to process. This can be the case, for instance, when a name is composed of a single stop-word or several concatenated ones (e.g.: `SomeParameter_08`). Third, even if at least one correct English word can be fetched to the Associator, it is possible this one simply does not find any associated concept.



All three cases, or any combination of these three cases, result in the fact no concept could be associated to the considered parameter. To overcome this problem, we propose a multimodal approach taking advantage of latent semantics contained in the data type information available through WSDL files. First, in real-world WS, a large proportion of types have a user-defined name, whose meaning can be considered as complementary to the parameter name. Additionally, many of these custom types are complex in the XSD sense, i.e. they can be compared to the structured data types used in programming languages. A parameter whose type is complex is made up of several subparameters, which can recursively be composed themselves of other subparameters, if they have a complex type too. Therefore, by taking advantage of the data types, one can access the semantic information implicitly contained in the type names and subparameter names and types.



**Fig. 2.** Excerpt from a real-world WSDL file: parameter with a complex XSD type

Fig. 2 gives an example of a complex type extracted from a real-world WSDL file. A parameter named *category* has a complex type called *categoryDetail*, defined as a sequence of two strings: a *singer* and a *composer*. If we suppose the word *category* is a stop-word, the Associator will not be able to provide any concept for this parameter. However, considering the words *singer* and *composer* gives access to additional information usable by the Associator.

The principle of our Type Explorer component is as follows. It is activated when the processing of the parameter name could not be used to successfully identify any concept. We start with the type name: if it is custom, we process it exactly like the parameter name, going through the preprocessing and association steps. In case of failure to associate any concept, we go further and consider the type structure. If it is complex, we access the first level of subparameters. For now, we only consider XSD sequences, because these are the most widespread, however the same approach can be extended to the other kinds of XSD types. We first focus on the subparameter names, and if the association is inconclusive, on their type names. In case of failure, the process recursively goes on by analyzing the structure of the subparameter types to access the second level of subparameters. The recursion stops when there is no more level to process (permanent failure) or as soon as concept can be associated (success).

## 4 Application to Real-World Descriptions

To assess its performance, we applied MATAWS to a collection of syntactic WS descriptions. We wanted to use a large collection of real-world descriptions, in order to avoid specific cases and to get consistent results. Given these criteria, the best collection we could find is the *Full Dataset* collection from the ASSAM project [12], previously mentioned in our review of WS descriptions collections (section 2.1). It contains 7877 operation distributed over 816 real-world WS descriptions. In this section, we present the results we obtained on this collection. First we adopt a quantitative point of view and distinguish parameters only in terms of annotated or non-annotated. Second, we analyze the results qualitatively and discuss the relevance of the concept associated to the parameters.

### 4.1 Quantitative Aspect

We first focus on the proportion of parameters from the Full Dataset collection which could be automatically annotated by MATAWS. In this section, we consider a parameter to be successfully annotated if our tool was able to associate it to at least one concept. Table 5 displays several values, corresponding to the progressive use of the different components described in section 3. Each row represents the performance obtained when using simultaneously the specified functionality and those mentioned in the previous rows.

The first line corresponds to the direct application of the Associator, with no significant preprocessing. The only transformation consists in setting parameter names to lowercase, which is compulsory to apply Sigma. Under these conditions, MATAWS can propose a concept for 39.63% of the parameters. This means close to 40% of the parameters names are single words, which can be retrieved directly in WordNet. The rest needs more preprocessing to be successfully annotated.

The second row corresponds to the introduction of the decomposition step. The small improvement in the success rate (around +2%) allows us to think compound names do not contain directly recognizable words. By adding the normalization step, the improvement is extremely large (almost +48%). Further analysis shows this is only marginally caused by the replacement of abbreviations by full words. Among the remaining 10%, one can find specific parameter forms we plan to introduce in our preprocessing, and word variations such as plural forms, also easily integrable in our approach.

**Table 5.** Success rates obtained by using the different functionalities of MATAWS

Added Modification	Proportion of Annotated Parameters
No preprocessing	39,63%
Decomposition	41,94%
Normalization	90,01%
Filtering	69,06%
Type Explorer	72,04%

A strong decrease (-21%) can be observed when introducing the filtering step. This means that, among the associated words, many correspond to stop-words, or concatenations of stop-words. In this case, the Annotator might be able to retrieve a concept, but this one is useless in this context (e.g. parameter). The introduction of the Type Explorer allows improving slightly our success rate (+3%), but its effect is not as strong as we expected. This can be justified by the fact most parameters with a custom type where annotated using only their names. Moreover, the type structure is difficult to exploit in this collection, because some types defined as complex surprisingly do not actually have any content (i.e. no subparameters at all).

## 4.2 Qualitative Aspect

The quantitative analysis reflects the fact a large proportion of parameters could be associated to a concept. The question is now to know if these associations, which were automatically retrieved, are relevant relatively to the context. For this matter, we isolated all the words detected in the whole set of parameters, thanks to our Preprocessor and Type Explorer. Table 6 shows the first most frequent words with their associated concept.

Overall, most of the annotated words are associated to relevant concepts, leading to an approximate success rate of 83%. Words like *computer*, *month*, *numeric*, *password*, *customer* are perfectly recognized, but this is not the case of several widespread words such as *name*, *user*, *address* or *value*.

Irrelevant concepts are due to the fact some words have several meanings and can therefore be associated to several concepts. Such ambiguity can be raised directly when the considered word has most probably a unique meaning in the context of WS. For instance, when the word *user* is submitted to Sigma, it outputs three concepts, including the one expected in this case, i.e. “someone employing something”. However, the top result corresponds to “someone who does drugs”, which explains the associated concept (*DiseaseOrSyndrome*). Similarly, the appropriate concept for *name* is among the concepts returned by Sigma, but the top result correspond to its meaning in the expression “in the name of the law”, hence the concept (*HoldsRight*). The quality of the annotation could be improved for such common words by simply selecting *a priori* the appropriate concepts, like we defined lists of stop-words and abbreviations.

The selection of an accurate concept can also be context dependent, which makes it impossible or difficult to perform it *a priori*. For instance, the word *value* corresponds to many concepts equally likely to appear in a WS description: quantity, monetary value, time duration, etc. Regarding this problem, the quality of the automatic annotation can be improved by deriving concepts from several words, when they are available. For instance, if the parameter name is *value01* and its type is *myCurrencyType*, then we have enough information to infer the most relevant concept. This can be done, for example, by taking advantage of the WordNet textual definitions.

**Table 6.** List of the most frequent words, with their associated concept. Bold rows represent semantically irrelevant concepts

Word	Occurrences	Associated Concept
identity	1255	TraitAttribute
key	548	Key
<b>name</b>	<b>470</b>	<b>HoldsWith</b>
<b>user</b>	<b>424</b>	<b>DiseaseOrSyndrome</b>
<b>code</b>	<b>295</b>	<b>Procedure</b>
<b>number</b>	<b>294</b>	<b>Object</b>
<b>address</b>	<b>258</b>	<b>SubjectiveAssessmentAttribute</b>
<b>date</b>	<b>203</b>	<b>DateFruit</b>
city	168	City
amount	135	ConstantQuantity
administrator	128	Position
message	115	Text
<b>value</b>	<b>106</b>	<b>ColorAttribute</b>
password	98	LinguisticExpression
pass	70	ContestAttribute
customer	52	Customer
company	51	Corporation
phone	41	Device
electronic	35	ElectricDevice
computer	33	Computer
mailing	33	Transfer
month	32	Month
numeric	32	Number

## 5 Conclusion

In this article, we presented our tool MATAWS, which implements a new method to semantically annotate WS descriptions. It focuses on WS parameters, i.e. on the Data semantics [4], and implements most of the requirements defined by Uren *et al.* [3] and relevant to our context: it processes WSDL files and produces OWL-S files (R1 & R4), and is fully automated (R7). This automation level is enforced through the use of both an ontological mapping of the WordNet lexicon, and a multimodal approach consisting in using not only parameter names, but also data type names and structures to identify appropriate ontological concepts. When compared to existing annotation tools such as ASSAM [12] and MWSAF [11], it is important to notice that MATAWS is much less flexible, because it does not include any machine learning abilities. This is due to the fact our goal is different: we want to batch annotate a large collection of WS descriptions without any human intervention, whereas the cited works aim at helping human users to annotate individual WS descriptions. Moreover, we tested MATAWS on a large collection of syntactic real-world WS descriptions, and despite its simplicity, it obtained very promising results, with 72% of the parameters annotated.

The version presented in this article constitutes a first step in the development of our tool. Although some parameters could not be associated with relevant concepts, it

is clear that we reduced the manual labor required for the annotation of WS. However, for now this reduction is not important enough to spare human intervention, which is needed at least to control the result of the annotation process. To get around this limitation, we plan to improve our tool on several points. First, in order to lower the proportion of parameters we failed to annotate, we can use other sources of latent semantics present in the WSDL descriptions: natural language descriptions and names of messages and operations. Second, the association step can be improved in two ways. We can complete the Associator by including more tools able to map a lexicon to an ontology, such as DBPedia [28]. This would complete and enhance the results already obtained through Sigma. Also, by taking advantage of our multimodal approach, we can retrieve all the words related to a given parameter through its data type, in order to compare them with concept definitions expressed in natural language (as found in a dictionary).

**Acknowledgments.** The authors would like to thank Koray Mançuhan, who participated in the development of MATAWS.

## References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* (2001)
2. Nagarajan, M.: Semantic Annotations in Web Services. *Semantic Web Services, Processes and Applications* 3, 35–61 (2006)
3. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art. *Journal of Web Semantics* 4, 14–28 (2006)
4. Sheth, A.P.: Semantic Web Process Lifecycle: Role of Semantics in Annotation, Discovery, Composition and Orchestration. In: *Workshop on E-Services and the Semantic Web* (2003)
5. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: Owl-S: Semantic Markup for Web Services, <http://www.w3.org/Submission/OWL-S/>
6. de Bruijn, J., Bussler, C., Domingue, J., Fensel, D., Hepp, M., Keller, U., Kifer, M., König-Ries, B., Kopecky, J., Lara, R., Lausen, H., Oren, E., Polleres, A., Roman, D., Scicluna, J., Stollberg, M.: Web Service Modeling Ontology, <http://www.w3.org/Submission/WSMO/>
7. Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M.T., Sheth, A., Verma, K.: Web Service Semantics - Wsdl-S, <http://www.w3.org/Submission/WSDL-S/>
8. Farrell, J., Lausen, H.: Semantic Annotations for Wsdl and Xml Schema, <http://www.w3.org/TR/sawsdl/>
9. Küster, U., König-Ries, B., Krug, A.: Opossum - an Online Portal to Collect and Share Sws Descriptions. In: *International Conference on Semantic Computing*, pp. 480–481 (2008)
10. Hess, A., Johnston, E., Kushmerick, N.: Assam: A Tool for Semi-Automatically Annotating Semantic Web Services. In: *International Semantic Web Conference* (2004)
11. Patil, A., Oundhakar, S., Sheth, A., Verma, K.: Meteor-S Web Service Annotation Framework. In: *International Conference on the World Wide Web* (2004)

12. Hess, A.: Assam (Automated Semantic Service Annotation with Machine Learning) Wsdl Annotator,  
<http://www.andreas-hess.info/projects/annotator/index.html>
13. InfoEther, BBN Technologies: Semwebcentral.Org,  
<http://www.projects.semwebcentral.org/>
14. Küster, U., König-Ries, B., Krug, A.: Opossum Online Portal for Semantic Services,  
<http://hnsp.inf-bb.uni-jena.de/opossum/index.php?action=dataguide>
15. Skoutas, D.N., Sacharidis, D., Kantere, V., Sellis, T.K.: Efficient Semantic Web Service Discovery in Centralized and P2P Environments. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 583–598. Springer, Heidelberg (2008)
16. Ma, J., Zhang, Y., He, J.: Web Services Discovery Based on Latent Semantic Approach. In: ICWS (2008)
17. Gomadam, K., Verma, K., Brewer, D., Sheth, A., Miller, J.: Radiant: A Tool for Semantic Annotation of Web Services. In: International Semantic Web Conference (2005)
18. Dimitrov, M., Simov, A., Momtchev, V., Konstantinov, M.: WSMO studio – A semantic web services modelling environment for WSMO. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 749–758. Springer, Heidelberg (2007)
19. Srinivasan, N.: Wsdl2owl-S,  
<http://www.semwebcentral.org/projects/wsdl2owl-s/>
20. Elenius, D., Denker, G.: Owl-S Editor,  
<http://owlseditor.semwebcentral.org/index.shtml>
21. Scicluna, J., Abela, C., Montebello, M.: Visual Modelling of Owl-S Services. In: IADIS International Conference WWW/Internet, Madrid, ES (2004)
22. Cherifi, C., Rivierre, Y., Santucci, J.-F.: Ws-Next, a Web Services Network Extractor Toolkit. In: 5th International Conference on Information Technology (2011)
23. Pease, A.: Sigma Knowledge Engineering Environment,  
<http://sigmakee.sourceforge.net/>
24. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: International Conference on Formal Ontology in Information Systems, Ogunquit, US-ME (2001)
25. Pease, A., Niles, I.: Linking Lexicons and Ontologies: Mapping Wordnet to the Suggested Upper Merged Ontology. In: IEEE International Conference on Information and Knowledge Engineering, pp. 412–416 (2003)
26. IEEE: Suo-Kif (Standard Upper Ontology Knowledge Interchange Format),  
<http://suo.ieee.org/SUO/KIF/suo-kif.html>
27. McGuinness, D.L., Harmelen, F.: Owl Web Ontology Language,  
<http://www.w3.org/TR/owl-features/>
28. Universität Leipzig, Freie Universität Berlin, OpenLink: Dbpedia.Org,  
<http://wiki.dbpedia.org>

# Meshing Semantic Web and Web2.0 Technologies to Construct Profiles: Case Study of Academia Europea Members

Petra Korica-Pehserl<sup>1</sup> and Atif Latif<sup>2</sup>

<sup>1</sup> Institute for Information Systems and Computer Media

<sup>2</sup> Institute for Knowledge Management

<sup>1,2</sup> Graz University of Technology,

Infeldgasse, 8010Graz, Austria

pkorica@sbox.tugraz.at,

atif.latif@student.tugraz.at

**Abstract.** Web is gaining a huge amount of data with every passed second. It is often claimed that information is doubling tenfold as fast as knowledge. This outburst of data posed the real challenges of searching and information management. A naive web user today struggles to find their piece of information at one place e.g. if user looks for information on some topic he or she is likely to find many links to a site containing some information on the topic, but it is the user who then has to wade through many snippets in consolidating them into one coherent piece of information to get the answer. The idea of obtaining consolidated information is still an unsolved problem. Traditional search approaches are failing to generate a consolidated and aggregated view on data due to meaningless state of the most data. Meanwhile Semantic Web with structured, interlinked, machine-readable databases is gaining on relevance and opening new gateways for data betterment. The goal of this paper is to describe an approach which encompasses Web 2.0 and Semantic Web technologies to locate and aggregate person's relevant information into one user profile. In this study we experimented with our approaches on scientists of Academia Europaea for their consolidated information. We propose an application where this found information can be consolidated and presented in coherent / perceivable way. We also proposed a tool which can assist editors in selecting and gather pieces of related information from different sites by following aggregated links without infringing copyright. We hope that the combination of Web 2.0 and Linked Data technologies can resulted in better management of Information and our proposed application can help users to have better view about the person information which in other cases dispersed on the Web.

**Keywords:** Web 2.0, Semantic Web, Linked Data, User Profiles.

## 1 Introduction

The Era of Web 2.0 brought revolutions in the way we manage and share information on today's Web, giving more access, flexibility and control about the content. Many

Web 2.0 services like blogs, wikis, communities, comments and tags emerged on the scene changing stateless Web to participatory one. The ease of use offered by these services brought the flood of information along with them, in consequence lot of unmanaged information piled up in pre data stock. The bulk of information available on today Web is clearly very fragmented, diverse and hard for common user to search in, for their desired information. To cater to the problem of finding relevant information on the Web search engines offer their services to the users. Search engines provide the facility to search on the basis of some keywords which they map to the description of the page and rank them on the hit value. Some of the popular search engines are Google, Yahoo and Bing. Due to an innovative page rank algorithm and very large indexed corpus Google currently leads the search engine market. However, since its launch mid-year 2009 Bing has gained a relevant market share [1]. Search engines return different types of information (Web pages, pictures etc.) on the basis of entered keywords. Still, every day, it is getting harder to find the information we are actually looking for. This is perhaps one of the biggest problems of search engines. They have a problem coping with the amount of information and services available, particularly since so many similar pieces of information occur in different places. Thus there is a need of a robust aggregation and consolidated mechanism returning fewer but more reliable results [2].

More specifically if we talk about person's data search on the Web. Normally, a search engine returns just a bulk of non-clustered links of mixed quality, like the Web page of person, CV or multimedia content, and users have to dig down further by following the links to find intended information. So the first logical step would be the aggregation of the results. These aggregation services should be able to gather all the possible information about a person crafted from Web in order to offer users a unique and coherent view. As we will see in the related work section, such aggregation engines are getting increasingly popular. However the aggregation services still return links according to statistical computation how well a search term matches a document on the Web. But would it not be great if a machine could understand which website, person, recent tweet, Flickr photo, or Facebook message we are currently looking for? Then we could get much better and more relevant search results in shorter time. However, currently, this is not yet quite possible because machines lack the semantic understanding and common sense to build bridges between information. In order to accomplish this, machines need access to knowledge databases and a common query language for extracting the information from the databases. Under the term of Semantic Web some techniques and linked databases emerged which could help us accomplish this [3][4].

In this paper we concentrated on two streams of World Wide Web i.e. Web 2.0 and Semantic Web, as they offer certain advantages in their respective usage. On the one hand Web 2.0 provides more variety of authoritative data in more unstructured and machine un-friendly way by API services which need certain heuristics and machine learning algorithms for knowledge exploration, on the other hand Semantic Web techniques (Linked Data) provide limited datasets but in more structured form, which can easily be located and disambiguated. Taking these facts into account, the goal of the paper is to develop an approach which encompasses information retrieved from both Web 2.0 and Semantic Web technologies together to locate and aggregate person's relevant information into one user profile. We want to show that combining



search engines API's along with intelligent use of semantic technologies and datasets related information can be located, disambiguated and delivered to the user. Further we propose an application where the information found can be aggregated and presented in a coherent way as well as proving that Semantic Web technologies and Conventional Web applications assist each other in better information management.

## 2 Related Work

Important information about a company or a person is now-a-days not only stored on a traditional Web page but also in the Web 2.0 services like blogs, forums, Facebook, etc. Because the bulk of information is found online, having consolidated results is of great importance for both companies and people. While traditional search approaches like searching for the name return a bulk of unsorted information, search services specialized for searching people or company data like Zoominfo [6], 123people [7], Pipl [8], Intelius [9] are getting increasingly popular today. These services adopted conventional crawling, indexing and parsing of Web pages to retrieve and cluster the information. For example Zoominfo currently crawls and indexes 45million people and nearly 5 million companies from the open Web to locate details of individual people and companies and then constructs profiles using artificial intelligence techniques (see [10]). Some other search machines like Intelius or Zabasearch [11] use publicly available government records and commercial sources [12] in addition.

For example when a user searches for the person, in this case a scientist "Hermann Maurer" his personal information, served universities, invited speakers talks, awards, attended events, his co-authors and other relevant pieces of information need to be presented to the users. This kind of services will certainly help out users to find information in a more easy way and in a clustered form.

Furthermore it is important to conduct the search for the "right" person in case of two or more persons having the same name (note: this is usually not the case with companies due to market laws). The disambiguation of the person's name is always a challenging task due to invariabilities and similarities we found in names of the persons belonging to similar or different disciplines. In the future we want to solve this by using different heuristics inspired by popular approaches in disambiguation used by popular search services like presenting list of all possible people with similar name, using metadata for clustering of similar interest people and analysis of social networks. By analysis it is clear that the usage of metadata or additional information can help to be more domain-specific and can cut down the search results.

Meanwhile Semantic Web with structured, interlinked, machine-readable databases is gaining on relevance. Our approach differs from standard people search engines mentioned above (for example Zabasearch, Zoominfo, ...) by using a search engine API (currently Bing Search API [13]), which results we streamline using heuristics and cluster similar to people search engines, in combination with Semantic Web databases like Linked Open Data to enrich already found resources. This means we try to combine Semantic Web (e.g. Linked Open Data) and Conventional Web (Web services, Datasets especially Web2.0) to find information of a person, integrate it and present it to the user. To authors' best knowledge there are no similar concepts and frameworks currently. We envision that the use of this additional data which is very

well linked to other semantic resources can help us provide a more detailed, dynamic and interlinked user profile.

Semantic Web (Linked open Data) offers various machine readable structured datasets which could be read by so called software agents to “understand” the content and the relationships between the entities in and outside of the datasets. One of the core projects which have recently played a vital part in enlightening the idea of Semantic Web is known as Linked Open Data [5]. This project was initiated by Semantic Web Education and Outreach (SWEO) Interest Group in October 2007 and provides semantically rich metadata datasets in more structured way by use of well-established semantic technologies (RDF [23], RDFa [24], N3 [25]) and ontologies. The Linked Data Cloud continues to grow rapidly and currently there are about 13.1 billion RDF triples which came cross from different practical, social, business and research domains. These well-structured, openly available datasets along with extra added utility of query languages (SPARQL) can be explored for knowledge discovery and creating cross-references between relevant resources at more ease.

A dataset worth mentioning is DBpedia, a semantic flip of Wikipedia. DBpedia is considered one of the most promising knowledge bases, having a complete ontology along with Yago classification [14]. It currently describes more than 2.9 million things, including at least 282,000 persons, 339,000 places etc. [15]. The knowledge base consists of 479 million pieces of information (RDF triples). The openly available RDF dumps make DBpedia an interesting subject of study. Its heavy interlinking within the Linked Data cloud makes it a perfect resource to search URIs. For current study, we concentrated on the part of DBpedia that encompasses data about persons due to its large indexing of personal data. The availability of datasets containing information about persons (DBpedia [26], FOAF [27], SIOC [28], DBLP [29]) provides a test ground to make applications that can locate, discover, aggregate and link personal information on the fly in various contexts.

Various studies have been conducted to highlight the potential benefits in use of these Linked Data person datasets, for example a recent study by Stankovic [16] has shown the certain benefits and drawback of using Linked data Open datasets for the profiling and expert systems. Set of crafted heuristics introduced in this study showed that Linked Data semantically enriched datasets has on a very good chance to be exploited as a playground to find certain interlinked information about the person on focus i.e. FOAF profile, publication list, authored books, video talks , blogs and Wikipedia articles. This chunks of interlinked information can be further explored by the Query languages (SPARQL) to get more hidden facts by issuing a queries on this information (considered as a single database) resulting in a knowledge discovery.

We have already proposed and implemented a system highlighting the certain benefits which developer and end user can have by use of semantic technologies and linked data datasets.[17] This system is named as CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data) [18]. CAF-SIAL is a proof of concept application to discover and present informational aspect of resource (Person) from Linked Data. CAF-SIAL is based on a methodology for harvesting person's relevant information from the gigantic Linked Open Data cloud. The methodology is based on combination of information: identification, extraction, integration and presentation. Relevant information is identified by using a set of heuristics. The identified information resource is extracted by employing an

intelligent URI discovery technique. The extracted information is further integrated with the help of a Concept Aggregation Framework. Then the information is presented to end users in logical informational aspects. This system is recently tested on authors of an Open Digital Journal named as “Journal of Universal Computer Science”. Further information about this system can be referred at [19].

In this paper we concentrate on one group of persons – on scientists. Motivated from these related systems in Web 2.0 and Semantic Web domains we planned to experiment in our test application with the test set Academia Europaea on similar grounds. Our objective of this study is to combine both unstructured Web 2.0 information and structured Semantic Web information by using Concept Aggregation Framework in one single system to find and present the information about member of Academia Europea.

### 3 Test Set

In this paper we want to explain our approach using a simple set of scientists of Academia Europaea. The Academia Europaea is a European (located in London), non-governmental non-profit association acting as an Academy. The Academy was founded in 1988, and has over 2000 members from thirty five European countries and eight non-European countries. The Academy is divided into 19 Academic Sections, each representing an independent scientific discipline. The Sections are ruled by Section chairs and their committees. The membership includes leading experts from the physical sciences and technology, biological sciences and medicine, mathematics, the letters and humanities, social and cognitive sciences, economics and the law. The members are scientists and scholars who collectively aim to promote learning, education and research. Membership is by invitation only, following a peer review selection process. The Academy has over 40 Nobel laureates in different disciplines [20]. With the set of persons chosen, we ran into two difficulties:

First, since the selection process for Academia Europea is demanding, the average age of members is high, hence less computer savvy than would happen with a younger group. Thus, finding information on scientists from Academia Europaea on the Web is a challenging task. After all, search services of course can only find data which is stored somewhere on the Internet. This can be a problem when dealing with person search for elderly people as their profiles and work records (for example publications from Academy members) are normally not directly available or on their own sites (many do not have one), but can often (if at all) be found indirectly via Web pages from their universities or through reports on their achievements.

Second, the AE has three main groups: humanities, natural science and medical/biological sciences. Of those a high percentage of members of the humanities are less likely to use computers and the Web than members from the other groups.

Third, Academy members come from over 40 different countries. Information is therefore found in many different languages. Multilanguage websites can pose a problem for semi-automatic detection of information about the scientist. Ideally the heuristics for extracting the right information should be done in many different languages, and after detecting the language of the website some translation software or some other heuristics should be used.

Thus, looking at the situation in hand sight it is clear that our techniques would produce still better results for e.g. a group of young scientists in natural sciences using a common language. However, this does make our modest success even more significant.

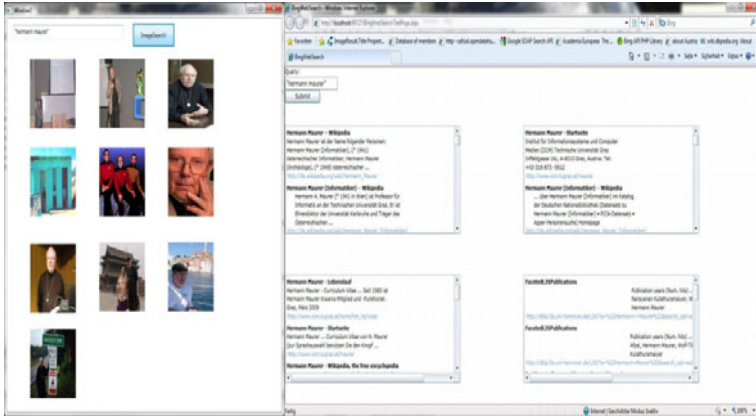
## 4 Approach

As described above we tested our application on the members of Academia Europaea test set. There were 2,249 test scientists from different fields of research for example molecular biology, economics, physics, astronomy, history, computer science and other scientific fields. As search query we automatically inserted full name of the scientist from the Academia Europaea member database into our application.

### 4.1 Conventional Web Search Approach

In the first experiment we used the search application for searching for images of Academy members on the Web. As mentioned above we wrote this application in C# using Bing version 2 API from Microsoft [13]. Bing Search API is used for querying the name of the searched person and it can return different types of results like Web pages, images, videos etc. The API returns various response fields for each result, see [21], amongst others it returns the *ImageResult.Title Property*. This property represents the so called “ALT” tag which represents the alternative text for the picture and usually gets displayed when user drives with the mouse over the picture or when the picture cannot be displayed on a web page. Given this, the ALT tag normally contains a quite good representation of what is shown by the picture. In this experiment we parse the ALT tag for the name of the scientist in the query and we only return images where ALT tag contains at least one part of his or hers name. After the query is finished, we display ten best images to the user. Each thumbnail leads the user to the web page containing the image so that he or she can then have the last decision whether the displayed result is correct.

The same approach can of course be used for various usages, for example also for text or video retrieval. A different part of our applications does this. Given the same query (= name of the searched person) for the image search our application starts various Web search queries using different metadata. It simultaneously performs a web search for Wikipedia pages, curriculum vitae, publication list, awards (for example there are over 40 Nobel prize winners in the Academy!), talks, special celebrations (for example celebration of 60th or 70th birthday), interviews, etc. A search for publications on the Web is also performed. In this case we parse the returned results' URLs and return only results where the URL contains at least one part of the name given in the query in order to get high quality results. The results are then clustered in four parts – Wikipedia, Homepage, Curriculum Vitae and Publications – which are displayed in separate frames. The described parts of the application are depicted in Figure 1 below.



**Fig. 1.** Results for image and Web (Wikipedia, homepage, CV and publications) search for Prof. Hermann Maurer, a member of Academia Europaea

Table 1 below shows the example results for our first experiment with image search for Academia Europaea test set.

**Table 1.** This table shows some of the example results for image search

Academia Europea test set Image Search	
Number of subjects	2.249
Returned images without heuristic	19.068
Returned images with heuristic	13.349

As we can see from the results of Table 1 this simple heuristic already helps in reducing the amount of false results. Note that just parsing URLs does not yet solve the disambiguation of the names. For example James Black is a name of a member of Academy but as this is a common name the probability is high that the Internet is full of URLs and images containing “James Black”. Even Wikipedia returns 23 results for “James Black” [22] and among them is our searched person James W. Black a British doctor and the winner of Nobel prize. One possible solution for this problem is adding some additional metadata to the query text like “Nobel” in this case, however this means that user has to have some knowledge about the domain of the searched person. In addition by restricting this application to search about various information concerning Academia Europaea scientists, or scientists in general, one could also profit of having a consistent domain-related metadata basis (e.g. scientific metadata) and therefore also better results.

### 4.2 Linked Data Search Approach

In our Linked Data (Semantic Web) approach, we used DBpedia and DBLP as our focused source to find URI of the Academia Europea members. DBLP dataset is used to specifically have the publication list to add more variety in the anticipated profile of the Person. In Linked Data a specific and unique URI is used to represent an entity

(Person, Place etc.), so locating a URI is considered as an important task. This URI is then further de-referenced to find additional information. We construct an algorithm which uses semantic technologies (ARC2, SPARQL) to locate the URI from DBpedia and DBLP. After applying this algorithm on the 2,249 members, we are able to locate 1,171 DBpedia URI and 505 DBLP URI. Further to nullify the ambiguous and wrong URI we passed it through set of heuristics. In the end we able to find 334 Valid DBpedia URI and 505 DBLP URI of the members as listed in Table 2.

**Table 2.** Result returned in Linked data search

Academia Europea test set Linked Data Search	
Number of subjects	2,249
DBpedia URI of the AEMembers	1171
DBLP URI of the AE Members	505
Authenticated DBpedia URIs	334

For example if we continued with our example of Hermann Maurer that we searched in conventional web previously, has successfully located his DBpedia URI and DBLP URI. After dereferencing these URI, in DBpedia we have found properties describing his biography, professional details as well as we got handful of his recent and old publications indexed by DBLP.

In concluding our two approaches, for future we envision organizing this information and presenting the results as shown in Figure 2 below: here we have aggregated the images, CV, Wikipedia links found by Bing search and Linked data retrieved information (biography and DBLP indexed publications) in one profile. We hope that this proposed system will give user a coherent and detail view of the information at one place. Still this initial aggregation returns a large set of links and documents and we believe that some sort of editorial process is needed due to sort out and consolidate this information.

The editorial process is also needed due to copyright issues it is important to think about how the consolidated information will be saved and presented to the user. It is not just possible to gather all the information from different sites and copy it to a consolidation server. This means we need to create a possibility to consolidate and display relevant data without infringing copyright. Instead of the trivial approach which would be to verify whether found pictures, videos and other multimedia content may be used, here we propose a tool where an editor can easily prepare steps necessary to gather and consolidate the information wanted and display it on-the-fly when user requests the information.

Our suggested approach is to gather a first version of consolidated information for a person with all relevant and unique links offered by the combination of Web 2.0 and Semantic Web. In the second step we ask a team of editors to go through the links, check them and annotate them. For example this editorial team first looks at the links and thinks about how to display relevant links and data on the server without infringing copyright. Then an editor uses a special plugin which records all necessary steps how the data of the interest is to be handled and displayed. For example Nobel Foundation has a copyright on all the pictures on their Web site. So if we take a picture of a Nobel prize winner as an example then the automatic part of our approach returns a link to a picture of this person on, among other image search result pages, the Nobel Foundation site. In order to be able to use this picture, editors start a plugin.

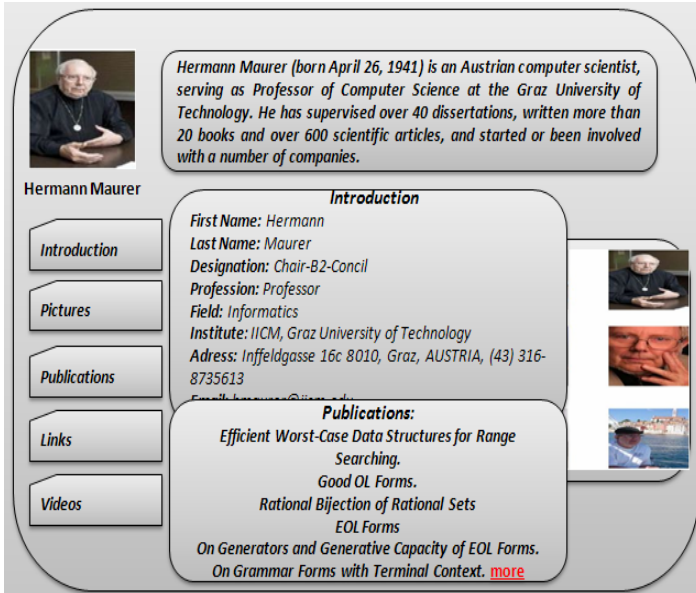


Fig. 2. Future Application Interface

This kind of plugin takes a group of parameters describing what to do in this case, for example do a screen dump, scroll to a part of page and cut out the picture using given coordinates, the date of link retrieve etc. These parameters would be added in as simple as possible manner – for example like recording a screen using Camtasia [30] and marking the relevant content on a screen dump. This plugin could be implemented as a server based application module that does not require the installation of any software on the user’s PC or workstation. It can therefore be used by anyone independent of hardware and operation system, as long as a Web browser is available. First author described a similar approach in [31] by developing a presentation tool called SIP which is able to avoid the violation of copyrights, allow access to data that may not be available to all and assure that the sources are always quoted properly. Our new plugin could also be installed on a publicly accessible server, in our case on the new Web Site of Academia Europaea which is currently programmed at our institute using JSP-Wiki Server [32].

We believe in using semi-automated approach for information retrieval in our application where firstly given query results from picture, video, audio and Web searches are combined into person’s profile and secondly a team of editors checks this profile for relevance of the results and copyright issues.

## 5 Conclusions and Future Work

We discussed an approach for combining results from Web 2.0 and Semantic Web into a structured person profile. In the future this approach could also be used for aggregation of results for objects like for example a famous city, university and other objects of interest.

A big advantage of using Web 2.0 in combination with Semantic Web is that currently more data (especially user generated content) can be found in the Web 2.0 giving extra edge in searching as compared to Semantic Web. Regarding Web 2.0 part of the application we want to develop more robust heuristics which can deal with multi-language environment of the Web - as for example our test data the members of Academia Europaea come from many different countries and therefore have lots of Web resources on different languages.

Currently an important advantage of our application is the semi-automatic search approach. As we already know, the areas where humans and machines excel are different. Therefore our software combines the best features from both perspectives and let the computer do the “hard” work of automatically searching the Web and presenting the results to the human, for example to an editor. Editors can easily understand the context of a given result and decide whether this it is relevant or not. Editors will also deal with the copyright issue by starting the plug-in described above and helping the machine to display the results correctly and without copyright violations. We believe that the idea of a plug-in for on the fly passing information to the user (i.e. using the server as client) to avoid copyright issues will gain even more importance in the future.

## References

1. [http://www.zdnet.de/news/digitale\\_wirtschaft\\_internet\\_ebusiness\\_comscore\\_bing\\_baut\\_marktanteil\\_in\\_den\\_usa\\_weiter\\_aus\\_sto-ry-39002364-41528722-1.htm](http://www.zdnet.de/news/digitale_wirtschaft_internet_ebusiness_comscore_bing_baut_marktanteil_in_den_usa_weiter_aus_sto-ry-39002364-41528722-1.htm) (last visited May 16, 2010)
2. Alexander Korth, A.: The Web of Data: Creating Machine-Accessible Information, Read Write Web (April 18, 2009), [http://www.readwriteweb.com/archives/web\\_of\\_data\\_machine\\_accessible\\_information.php](http://www.readwriteweb.com/archives/web_of_data_machine_accessible_information.php) (last visited May 30, 2010)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American Online Edition (May 17, 2001), [http://www.ryerson.ca/~dgrimsha/courses/cps720\\_02/resources/Scientific%20American%20The%20Semantic%20Web.htm](http://www.ryerson.ca/~dgrimsha/courses/cps720_02/resources/Scientific%20American%20The%20Semantic%20Web.htm) (last visited May 16, 2001)
4. Dbpedia.org (last visited May 16, 2010)
5. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking Open Data on the Web. In: Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria (May 2007)
6. Zoominfo, <http://www.zoominfo.com/> (last visited May 30, 2010)
7. 123people, <http://www.123people.at/> (last visited May 30, 2010)
8. Pipl, <http://pipl.com/> (last visited May 30, 2010)
9. Intelius, <http://www.intellus.com> (May 30, 2010)
10. <http://www.zoominfo.com/About/company/technology.aspx> (May 30, 2010)
11. <http://www.zabasearch.com> (last visited May 30, 2010)



12. Ramasastry, A.: Can We Stop Zabasearch – and Similar Personal Information Search Engines?: When Data Democratization Verges on Privacy Invasion, <http://writ.news.findlaw.com/ramasastry/20050512.html> (last visited May 30, 2010)
13. Bing, API, Version 2, <http://msdn.microsoft.com/en-us/library/dd251056.aspx> (last visited May 30, 2010)
14. YAGO: A Core of Knowledge, <http://www.mpiinf.mpg.de/yago-naga/yago/> (last visited May 30, 2010)
15. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., Bpeddia, D.: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference, Busan, Korea. Springer, Heidelberg (2007)
16. Stankovic, M., Wagner, C., Jovanovic, J., Laublet, P.: Looking for Experts? What can Linked Data do for You? In: Proceedings of the Linked Data on the Web Workshop (LDOW 2010), Raleigh, North Carolina, USA, April 27 (2010)
17. Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K.: CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data. In: Proceedings of International Conference on Networked Digital Technologies, Ostrava, Czech Republic, July 28-31, pp. 100–105 (2009)
18. Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K.: Harvesting Pertinent Resources from Linked Data. Accepted in Journal of Digital Information Management
19. Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H.: Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). In: Proceedings of the Linked Data on the Web Workshop (LDOW 2010), Raleigh, North Carolina, USA (April 27, 2010)
20. <http://acadeuro.org/index.php?id=6> (last visited May 30, 2010)
21. <http://msdn.microsoft.com/en-us/library/dd250942.aspx> (last visited April 18, 2010)
22. [http://en.wikipedia.org/wiki/James\\_Black](http://en.wikipedia.org/wiki/James_Black) (last visited April 24, 2010)
23. <http://www.w3.org/TR/rdf-primer/> (last visited June 15, 2010)
24. <http://www.w3.org/TR/xhtml-rdfa-primer/> (last visited June 15, 2010)
25. <http://www.w3.org/DesignIssues/Notation3.html> (last visited June 15, 2010)
26. <http://dbpedia.org> (last visited June 15, 2010)
27. <http://www.foaf-project.org/> (last visited June 15, 2010)
28. <http://sioc-project.org/> (last visited June 15, 2010)
29. <http://dblp.l3s.de/d2r/> (last visited June 15, 2010)
30. Camtasia – a screen recording tool, <http://www.techsmith.com/camtasia.asp> (last visited June 17, 2010)
31. Trattner, C., Helic, D., Korica-Pehserl, P., Maurer, H.: Click, Click– and an Educational Presentation is Available on the Web. Accepted in ED-MEDIA 2010
32. JSPWiki - A feature-rich and extensible WikiWiki engine built around the standard J2EE components, <http://www.jspwiki.org> (last visited June 19, 2010)

# Towards Ontology-Based Collaboration Framework Based on Messaging System

Gridaphat Sriharee

Department of Computer and Information Science  
King Mongkut's University of Technology North Bangkok  
1518 Piboolsongkram, Bangsue, Bangkok 10800  
gridaphat@kmutnb.ac.th

**Abstract.** This paper proposes a collaboration framework to support the participants to realise the workflow of business process transaction regarding the message exchange. The collaboration framework is implemented into functional management layers: business process, ontology and technical layer. In ontology management layer, a collaboration ontology is created and used for realising the collaboration messages. The sequence diagram and state diagram are designed according the requirement of business process management to model the collaboration flow. In technical management layer, the framework is implemented on the JMS messaging platform. A scenario of logistics collaboration is demonstrated and it shows that effective collaboration can be achieved. In addition, a discussion of job tracking is presented to suggest how the system benefits from the ontology-based collaboration description.

**Keywords:** Collaboration, Messaging System, Ontology.

## 1 Introduction

Organisations use collaboration as a method to exchange information inside and across their boundaries. A collaboration framework is designed to help the participants in a (business) community to achieve clearly defined outcomes. The collaboration framework enables integration and automation of business processes. Good collaboration can reduce transaction costs. Therefore, organisations that collaborate efficiently have greater revenues than their competitors. Organisations can create various business models that will be offered to customers. The business model combines a set of services while the collaboration framework provides a collaboration process to support such a model.

A simplified mechanism for collaboration can be implemented through a messaging system. The messaging system supports collaboration within and across organisations. Basic messaging models are point-to-point and publish-subscribe messaging models. The message exchanging is implemented by an asynchronous communication protocol. The messaging system provides a means for collaboration by allowing the actors (collaboration applications) to post events and react to the events posted by other actors. The messages posted are sent to message queues in the

messaging system. The message queues are implemented by database and/or file. The recorded messages in the message queues represent states of the collaboration and that information can be used for job tracking.

The collaboration framework is a framework that supports business process management. The business process has its business workflow (e.g. described by BPEL) which may require collaboration across manual or automated tasks during business process transactions. Such business process is typically of long duration and some of the exchanging messages are routed to the collaboration framework that works with the intention to achieve effective communication across various business applications and delivery channels. The collaboration framework gives information to detect the progress of business process execution and guarantees delivery of the messages which lead to a collaborative business process.

This paper proposes a collaboration framework from technical perspective that is realisable by three functional logic of management layers: business process layer, ontology layer and technical layer, as follows:

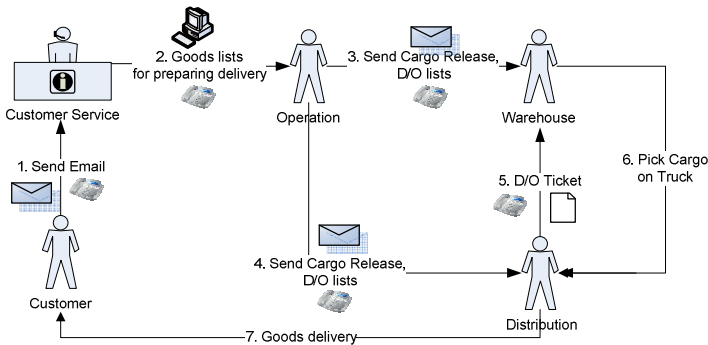
- *Business process layer* focuses on the function to realise the business operations through collaboration message concerning collaboration in the transacted business process. Modeling the collaboration message is realised by the UML sequence diagram and the UML state diagram.
- *Ontology layer* focuses on the formal semantics of the collaboration message. A collaboration ontology is proposed as a shared knowledge of the collaboration message exchanging in the collaboration. Also, the ontology-based collaboration description is created and represented by OWL [1]. The collaboration ontology is realised in both the business process and technical layers.
- *Technical layer* focuses on the architecture of the collaboration framework that is developed on JMS messaging platform [2]. In this layer, the collaboration tracking component is introduced. It uses the ontology-based collaboration description for event tracking relating to the collaboration.

A collaborative logistics is demonstrated to present how the collaboration is realised according to the mentioned management layers. The rest of this paper is laid out as follows. Section 2 introduces the need of collaboration with an example of logistics delivery process. Section 3 presents a collaboration framework with the functional management in the business process layer, ontology layer and technical layer. Section 4 discusses the use of ontology-based collaboration description for job tracking purposes. The requirements of message tracking and some examples of tracking are given. Section 5 discusses some related works and section 6 presents a conclusion and discusses future work.

## 2 A Scenario of Logistics Delivery Process

This section presents an example of collaboration required in the logistics. Collaborative logistics is an important process to reduce the cost of cooperation management within and between organisations. It enables the delivery process to be conducted with high performance that meets the requirements of the customers. Thus, the efficient collaboration is required in the logistics organisations.

Fig. 1 depicts a simple scenario of a logistics business process for goods delivery analysed by [3]. A logistics company provides a value chain business with four main sections: *customer service*, *operation*, *warehouse* and *distribution* sections. These sections use telephone and email communication for collaboration. Each section may have its own particular logistics application used to support its operation and each such application may link to one or more applications belonging to the other sections. Note that the automated collaboration framework executed by software application is not yet addressed in this scenario. Collaboration in the logistics delivery process is initiated by a customer who sends an email to customer services requesting goods delivery. The customer service section coordinates with the operation section to prepare goods delivery through supporting applications. Later, the operation section sends e-mails to the warehouse and distribution section for preparing the cargo release of goods with D/O (Delivery Order) lists requested by the customer. The warehouse prepares goods packages following the D/O lists for shipment. The distributions section issues a D/O ticket that is used for picking goods from the warehouse and later transports the goods to the customer. To enable collaborative logistics, software integration is required. A messaging system is one of the technologies that can be considered for enterprise application integration [4], and this can be implemented for collaboration.



**Fig. 1.** An interaction among the participants in the delivery process of a logistics company

### 3 The Collaboration Framework

Fig. 2 depicts an overview of the proposed collaboration framework and is logically classified into business process, ontology and technical layers.

In the business process layer, the collaboration manager creates the collaboration specifications: a UML sequence diagram and a state diagram which represent the collaboration. These diagrams are analysed according to the business process specification created by the business process manager. The sequence diagram presents the interactions among the actors in the collaboration. The state diagram presents the states of the collaboration messages of a particular business function. The BPM model is a standard for capturing business processes at the level of domain analysis [5]. In this research, the BPM model is addressed as a foundation of the collaboration

specifications. Here, the collaboration specifications represent business operations from a technical perspective rather from the high-level business process prospective.

In the ontology layer, the collaboration manager interacts with the collaboration information manager, a component for generating ontology-based collaboration description that is compatible with collaboration specifications. The collaboration information manager is presented as a tool that is capable of transforming UML state diagram schema (e.g. XMI [6]) to ontology-based schema i.e. OWL schema. This will be developed by providing the mapping parser. The ontology-based collaboration description is created by using the definitions of collaboration messages defined in the collaboration ontology (created by the collaboration manager).

The messaging system supports asynchronous communication protocol with reliable delivery. In this research, collaboration is realised by the ontology-based messaging system in the technical layer. It is a messaging system developed on a JMS platform [2] and is enhanced by using ontology-based collaboration description for job tracking purposes. The participants (customer service, operation, warehouse and distribution sections) can query particular events of interest using such ontology-based collaboration description.

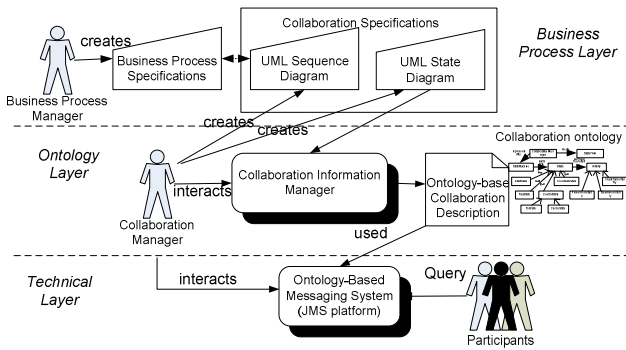


Fig. 2. The collaboration framework

### 3.1 Business Process Layer

This layer focuses on how to describe the collaboration and collaboration message between the actors participating in the collaboration. The interaction between the actors is thus considered according to the sequence of sending/receiving messages in the collaboration.

Fig. 3 depicts a UML sequence diagram that represents the collaboration in a logistics delivery process according to the collaboration mentioned in Section 2. The actors are collaboration applications: customer service, operation, warehouse, distribution and system applications. The system application is introduced to the model for message synchronisation and administration purposes. Fig. 4 depicts a UML state diagram of the collaboration. The collaboration message is represented by some states according to analysed sequence diagram (Fig. 3). A state has its transition. A state changes to another state when there is a call for sending a message.

A transition has a sender and a receiver actor (indicated by parenthesis); the state is owned by the sender, so for example *NewTask* is the state of the operation actor. A particular collaboration message is hence defined according to a series of states occurring in the various actors and that indicates about a whole activity chain of a particular process. Note that each actor also has its own states to complete its individual task. For example the operation actor is in the waiting state before it changes into the *NewTask* state.

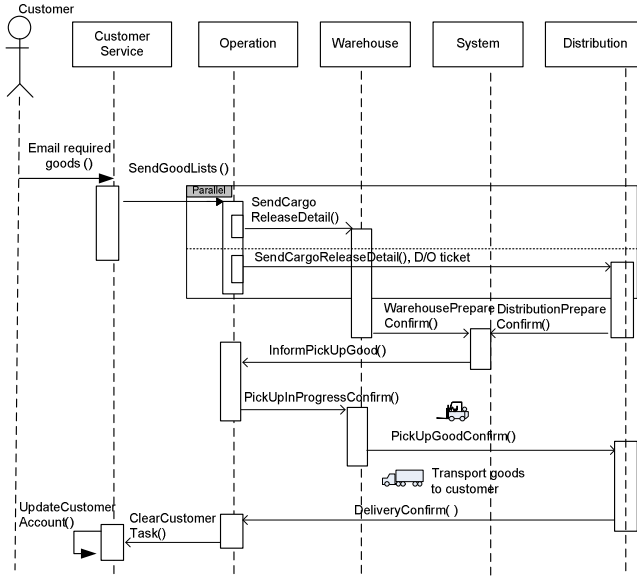


Fig. 3. Sequence diagram of the collaboration

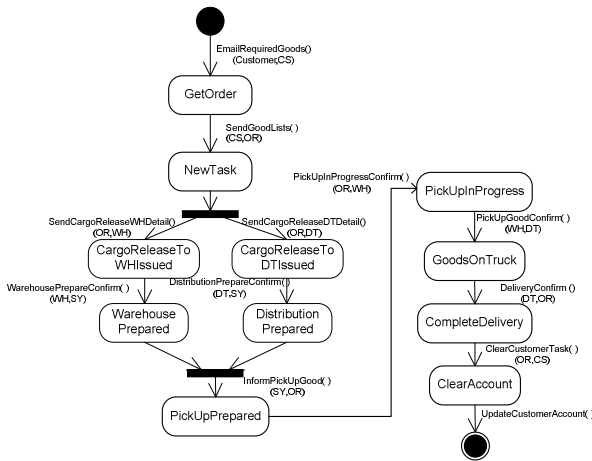


Fig. 4. State diagram of the collaboration

### 3.2 Ontology Layer

Ontology is a shared model to represent the conceptualisation of things in a particular domain [7]. Ontology is also used as an agreement to achieve a common understanding of information use between the participants in a community. Regarding the business process, ontology is created to represent a high level description of the information used in the system; such a description is easily understandable by humans and is available for the machine to query.

The collaboration ontology is created for describing the collaboration messages (represented by states) and relevant properties according to the collaboration. The collaboration is represented in terms of state machine model with reference to the state diagram analysed in the business process layer. The state associates to the messaging information that is described by some attributes such as message id, send date, receive date, job number and additional detail. The collaboration comprises states and collaboration activities with upper concepts depicted in Fig. 5. The collaboration activity concept represents the transition associating to the state. The collaboration activity can be an operation activity relating to the messaging function and also a manual activity such as phone calling or emailing to the users.

In this layer, the ontology-based collaboration description described by OWL is created. Such a description is used for a particular collaboration; the example shown in Fig. 6 is the collaboration information for the collaboration of the logistics delivery process according to the state diagram analysed in the previous section.

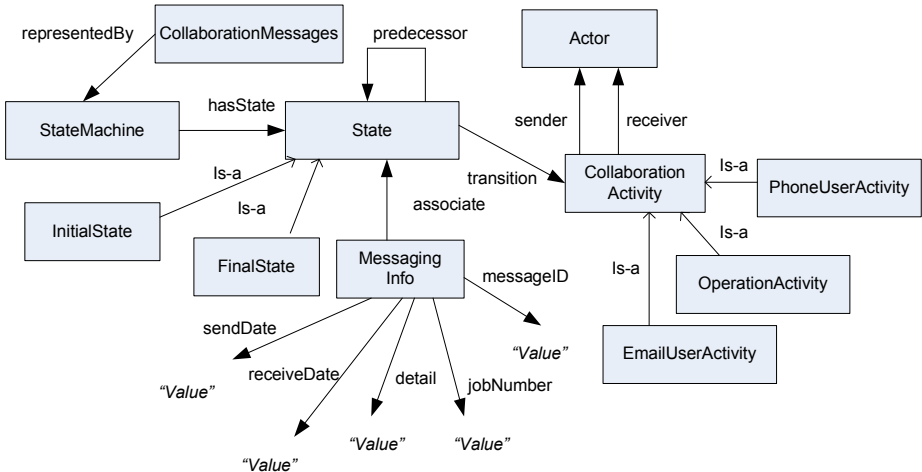


Fig. 5. A Collaboration Ontology

```

<rdf:RDF
  xmlns="http://www.example.com/collaborationonto.owl#"
  xml:base="http://www.example.com/collaborationonto.owl"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:owl="http://www.w3.org/2002/07/owl#" >
  <CollaborationMessages rdf:ID="LogisticsDeliveryProcess">
    <representedBy
      rdf:resource="#LogisticsDeliveryProcessStateMachine"/>
  </CollaborationMessages>
  <StateMachine
    rdf:ID="LogisticsDeliveryProcessStateMachine">
    <hasState rdf:resource="#CargoReleaseToWHIssued"/>
    <hasState rdf:resource="#DistributionPrepared"/>
    <hasState rdf:resource="#GoodsOnTruck"/>
    <hasState rdf:resource="#CargoReleaseToDTIssued"/>
    <hasState rdf:resource="#PickUpPrepared"/>
    <hasState rdf:resource="#PickUpInProgress"/>
    <hasState rdf:resource="#WarehousePrepared"/>
    <hasState rdf:resource="#InitState"/>
    <hasState rdf:resource="#GetOrder"/>
    <hasState rdf:resource="#NewTask"/>
    <hasState rdf:resource="#ClearAccount"/>
    <hasState rdf:resource="#CompleteDelivery"/>
  </StateMachine>
  <State rdf:ID="NewTask">
    <transition rdf:resource="#SendCargoReleaseWHDetail"/>
    <transition rdf:resource="#SendCargoReleaseDTDetail"/>
    <predecessor rdf:resource="#GetOrder"/>
  </State>
  <OperationActivity rdf:ID="SendCargoReleaseWHDetail">
    <sender rdf:resource="#Operation"/>
    <receiver rdf:resource="#Warehouse"/>
  </OperationActivity>
  <OperationActivity rdf:ID="SendCargoReleaseDTDetail">
    <sender rdf:resource="#Operation"/>
    <receiver rdf:resource="#Distribution"/>
  </OperationActivity>
</rdf>

```

**Fig. 6.** Part of an ontology-based collaboration description for collaborative logistics

### 3.3 Technical Layer

In this research, the ontology-based messaging system is developed on a JMS messaging platform. Fig. 7 depicts the components of an ontology-based messaging system. Client applications are developed according to the actors involved in the collaboration.

The message queues are managed for particular actors and each is represented by a table in the database. The messages sent and received in the framework are represented by states. Their associated messaging information are stored in the table of the message queue database. The JMS object (java class) is created for each message queue. A collaboration tracking component provides job tracking function to the framework. Collaboration manager creates message tracking requirements



(discussed in Section 4) which are some events of interest. The message tracking requirements are used as inputs of the collaboration tracking component for monitoring tasks or assigning automatic tasks in the collaboration. Fig. 8 depicts an example of customer service client to issue a new task and to track the delivery process according to the message exchange.

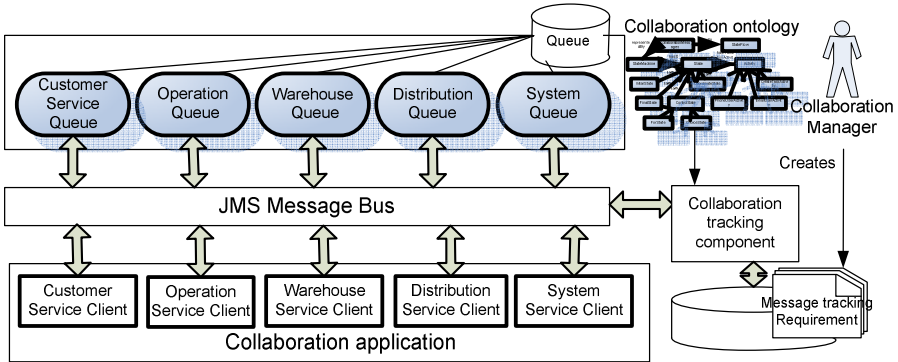


Fig. 7. The technical layer



Fig. 8. Example of client application

## 4 The Use of Ontology-Based Collaboration Description

In the technical layer, the collaboration tracking component is introduced. It requires the message tracking requirement as input information. The message tracking requirement represents the events of interest specified by the collaboration manager.

Such events may relate to a particular actor or the entire collaboration process and may relate to some external actors not directly involved in the computing framework (e.g. customers). The purposes of tracking vary from simple to complex questions for example:

- (i) *If the goods are not delivered to the customer within 2 days after the goods are picked up from the warehouse, then the system must notify the distribution manager.*
- (ii) *If the goods are ready to deliver then customer service should call to the customer to confirm the delivery date.*
- (iii) *If the goods have been delivered to the customer already, then the system reports the lists of tasks to history log file.*
- (iv) *Is the goods delivery finished?*
- (v) *Where is the current state of the task in the workflow?*
- (vi) *How is the sequence of states of the collaboration of particular job?*
- (vii) *What activity must be performed when the goods are ready for delivery?*

From the examples mentioned above, it is clear that tracking purposes can be considered as belonging to the need to manipulate task and the need to question the knowledge of the transacted process during execution time and in general events.

The implementation of the message tracking is considered from two different angles: *how to represent the message tracking requirement?* and *how to use such requirement in execution framework?* The latter point of view requires the output from the former to be the input of the execution system. Regarding the former point of view, there are a number of representation languages that can be used to represent the message tracking requirement. Some rule-based languages can be used for representation. For example, XML rule-based language (e.g. RuleML [8]), semantic rule-based language (e.g. SWRL [9], WRL [10]) and RDF language (e.g. TRIPLE [11]). With such representation, humans can understand the description and the described description enables the machine to query. For example, the rule that satisfies requirement (ii) can be defined as the expression

*State(GoodsOnTruck) →  
CollaborationActivity(CallUserToConfirmDeliveryDate).*

This rule states that if the antecedent is true then the consequent must be true.

Regarding the execution framework, the execution can be manipulated in application logic and by querying technique and rule-reasoning. For an execution that requires rule reasoning, the rule engine is needed as well as the transformation from the rule representation to another rule format that the rule engine can execute (e.g. SWRL to Jess). An identifying a suitable implementation depends on some technique and the experiences of the developer. For example, the requirement (iv) can be implemented by querying technique that is checking whether the collaboration message is in the state `CompleteDelivery` which means that the goods are delivered to customer already. Below is an example of the OWL/RDF instance description that created from messaging information recorded in the message queue database.

```

<MessagingInfo rdf:ID="MsgJob01">
  <messageID rdf:datatype="xsd:int">1      </messageID>
  <detail rdf:datatype="xsd:string">
    Good delivery is confirmed.      </detail>
  <jobNumber rdf:datatype="xsd:int">111 </jobNumber>
  <receiveDate rdf:datatype="xsd:dateTime">
    2010-09-05T19:37:27      </receiveDate>
  <associate rdf:resource="#CompleteDelivery" />
</MessagingInfo>

```

The SPARQL [12] query that satisfies the requirement (iv) is specified as follows:

```

SELECT ?job ?date
WHERE {
  ?messageInfo <#associate> <#CompleteDelivery>.
  ?messageInfo <#jobNumber> ?job.
  ?messageInfo <#receiveDate> ?date.
}

```

With separating the management of the ontology layer from the technical layer, job tracking is more flexible. For example, enrichment description relating to tracking requirement can be implemented into the collaboration ontology while the information stored in the message queue database may contain simple information only. Also, various queries can be defined.

## 5 Related Works

There are some related works as follows. [13] presented a collaboration ontology used to promote integration among collaborative software tools such as e-mail, chat and forum. [14] discussed an ontology that is developed to represent different aspects of workflows for collaborative ontology development. [15] proposed a conceptual model for business process collaboration using the process ontology. The process ontology represents formal semantics of the process elements (e.g., entities, objects, activities) and gives formal understanding of the process model. According to message tracking, [16] proposed the message tracking for publish-subscribe messaging middleware based on the JMS API and [17] proposed the message attributes and the message tracking in web services framework. The message is defined according to the SOAP messaging protocol. [18] presented a JMS agent gateway that is responsible for message exchange from JMS providers to logistics multi-agent brokers. They used ontology as an information model to define a task for message exchange. [19] proposed a collaborative network that uses ontology-based information to automate the specification of BPMN collaborative processes.

This research adopts ontology as a shared model for use in the collaboration but uses it for a different purpose from [13] and [14]. However, some defined concepts are similar in which tracking process relates to state and transition and activity as basis information. This work is close to [15] and [19] in which ontology is used to represent formal semantics of the collaboration. However, the proposed collaboration ontology is created from technical perspective rather from business process perspective discussed in [15] and [19]. This research focused on the job tracking; in

contrast, work [14] tracks ontology development and work [16] and [17] track the message and its routing. In this research, JMS messaging platform is a centralised system for providing the knowledge of the collaboration but work [17] tracks the message in decentralised manner.

## 6 Conclusion

This research presented a means of developing a collaboration support system using an ontology approach. A collaborative logistics is studied and developed in a previous work [3]. In this research, the messaging system is enhanced by incorporating ontology and ontology-based description for realising the information in the collaboration. The proposed collaboration ontology contains some simple concepts which represent simple states and transitions, and are used for querying purposes. The collaboration is easily realisable regarding the proposed logical functions: business process, ontology and technical layer. In this work, message exchange of logistics collaboration is implemented based on point-to-point messaging model using asynchronous communication protocol and this can be extended to support publish-subscribe messaging model. With the mapping between semantic information and functional logic designed by sequence and state diagram, the participants thus are able to realise the workflow during business transaction.

## References

1. OWL Web Ontology Language, <http://www.w3.org/TR/owl-features/>
2. <http://java.sun.com/products/jms/>
3. Pongphagha, P.: A Development of Collaborative Logistics of Kerry Logistics Co. Ltd. Thailand. Master Project, Thai (2010)
4. Gawlick, D.: Message Queuing for Business Integration. *eAi Journal* (October 2002)
5. <http://www.bpmi.org/>
6. XML Metadata Interchange, <http://www.omg.org/technology/documents/formal/xmi.htm>
7. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2), 199–220 (1993)
8. Rule Markup Initiative, <http://ruleml.org/>
9. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, <http://www.w3.org/Submission/SWRL/>
10. Web Rule Language (WRL), <http://www.w3.org/Submission/WRL/>
11. Triple, <http://triple.semanticweb.org/>
12. SPARQL Query Language for RDF, <http://www.w3.org/TR/rdf-sparql-query/>
13. Oliveira, F.F., Antunes, J.C.P., Guizzardi, R.S.S.: Towards a Collaboration Ontology. In: *Anais do 2nd Workshop on Ontologies and Metamodels in Software and Data Engineering*, João Pessoa/PB (2007)
14. Sebastian, A., Noy, N.F., Tudorache, T., Musen, M.A.: A Generic Ontology for Collaborative Ontology-Development Workflows. In: Gangemi, A., Euzenat, J. (eds.) *EKAW 2008*. LNCS (LNAI), vol. 5268, pp. 318–328. Springer, Heidelberg (2008)

15. Gong, R., Li, Q., Ning, K., Chen, Y., O'Sullivan, D.: Business process collaboration using semantic interoperability: Review and framework. In: Mizoguchi, R., Shi, Z.-Z., Giunchiglia, F. (eds.) ASWC 2006. LNCS, vol. 4185, pp. 191–204. Springer, Heidelberg (2006)
16. Jun, S., Astley, M.: Low-Overhead Message Tracking for Distributed Messaging. In: Proceedings of the ACM/IFIP/USENIX 2006, International Conference on Middleware, Australia, pp. 363–381 (2006)
17. Sahai, A., Machiraju, V., Ouyang, J., Wurster, K.: Message Tracking in SOAP-Based Web Services. HPL-2001-199 (2001)
18. Curry, E., Chambers, D., Lyons, G.: Enterprise Service Facilitation within Agent Environment. In: Proceedings of the IASTED Conference on Software Engineering and Application, pp. 601–606 (November 9-11, 2004)
19. Rajsiri, V., Lorré, J., Bénaben, F., Pingaud, H.: Collaborative Process Definition Using An Ontology-Based Approach. In: IFIP International Federation for Information Processing, vol. 283, pp. 205–212 (2008)

# A QoS and Consumer Personality Considered Services Discovery

Xiuqin Ma, Norrozila Sulaiman, and Hongwu Qin

Faculty of Computer Systems and Software Engineering  
Universiti Malaysia Pahang

Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia

xueener@yahoo.com.cn, norrozila@ump.edu.my, qhwump@gmail.com

**Abstract.** How to dynamically find web services which best meet the requirement of consumers from the massive services is an ongoing research direction. In this paper, we propose a Quality of Service (QoS) and Consumer Personality Considered services Discovery (QSCPCD) mechanism, which complements the existing service discovery. QSCPCD is a consumer-oriented service discovery mechanism, which builds at the side of service consumer and is only used by consumer himself. Initially, we obtain a variety of related web services by calculating similarity, which can improve recall ratio. From previous results, we then find services based on consumer's requirement for QoS, which can improve precision ratio. As a result, it can speed up the discovery of services, ensure high precision of service discovery, and take into QoS and consumer personality consideration. Experiment results demonstrate the contribution of the proposed model.

**Keywords:** service discovery; QoS ; UDDI; consumer personality.

## 1 Introduction

With the development of distributed computing technology, it is desirable to facilitate communication between applications and resource share in geographically distributed systems. As a result, the emergences of web services [1] and global computational grids [2] bring changes to the traditional paradigm of distributed computing.

A basic service in a service framework is service discovery: given a description of services desired, a service discovery mechanism returns a set of services that match the description. Service discovery has been an active research topic in recent years. UDDI [3] is a fully centralized service discovery mechanism. To use UDDI, consumers must be familiar with the rule of classification. And it can easily cause error due to heavy access requests. Based on UDDI, two kinds of related work were done: supplement and extension to UDDI. As for supplement, there are UDDIe [4], activeUDDI[5], UDDI- $M^T$  [6], etc.; as for extension, there are WS-Inspection [7], "my service"[8], etc. However, little support is provided for searching for a service based on QoS and consumer personality. In this paper, we propose a service discovery model that is defined as QoS and Consumer Personality Concerned services Discovery (QSCPCD).

The rest of the paper is organized as follows: Section 2 presents a basic QoS and Consumer Personality Considered services Discovery framework model. Section 3 shows the performance of this service discovery framework, using our simulation results. Finally, Section 4 concludes this paper.

## 2 QoS and Consumer Personality Considered Discovery Model

### 2.1 Roles of This Model

(1) Service Provider: It can provide services, which can be invoked by consumers. Service providers can publish service information in Service Registry such as UDDI.

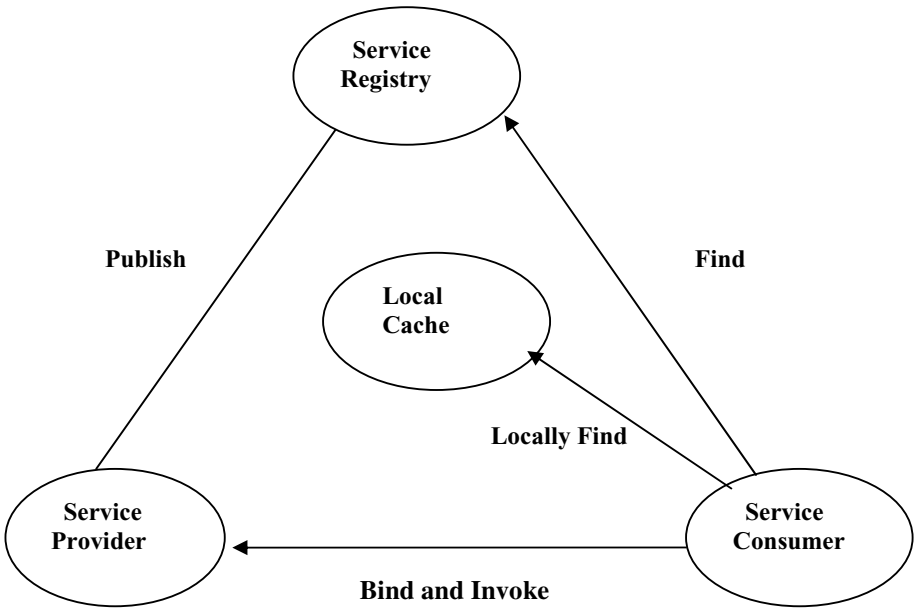


Fig. 1. System Model

(2) Service Consumer: It invokes all kinds of various services which are provided by service providers. In this model consumers send the service request to service registry and local cache when they need services. Local cache searches services based on service request. As a result, we get some service information, in which local cache carries out the second search by QoS. If we can find the service meeting the consumer’s needs in local cache, we can invoke the service; if there is no applicable service in local cache, we have to turn to Service Registry.

(3)Service Registry: where providers register services and consumers discover services.

(4)Local cache: There is the most frequently-used services information during a given time. The consumer can achieve applicable service information by related query

interface. In order to keep the most frequently-used services information in local cache, we carry out maintenance of information coherence and service overlay.

### 2.2 Operation of This Model

- (1) Publish: Providers publish service information in Service Registry by this operation.
- (2) Find: Consumers search the available service in Service Registry.
- (3) Locally find: Consumers can search service satisfying the needs in local cache by this operation which consist of two steps.

The first step of this operation is to discover some service information based on similarity between consumer’s service description and services description in local cache. Numerical calculation [9] is applied with the propose of getting similarity between WC (WC denotes service of consumer request) and WL (WL denotes service in local cache). We construct the following similarity function:

$$S(W_C, W_L) = k_1 S_{name}(W_C, W_L) + k_2 S_{text}(W_C, W_L) \tag{1}$$

$$\sum_i k_i = 1, \quad 0 \leq k_i \leq 1, \quad i = 1, 2.$$

Where  $S_{name}(W_C, W_L)$  and  $S_{text}(W_C, W_L)$  are similarity functions based on name description and text description respectively, which can be achieved by string match algorithm.  $S_{name}(W_C, W_L)$  and  $S_{text}(W_C, W_L)$  are not of equal importance to consumer, who likes to impose weights on his choice parameters. So there is a weight  $k_i$ . Consumer can set a threshold  $\lambda$ . If  $S(W_C, W_L) < \lambda$ ,  $W_L$  will be omitted, else  $W_L$  will be kept in service pool which briefly stores high similarity services from local cache.

The second step of this operation is to discover some service information based on Quality of Service (QoS) from service pool. Our definitions of QoS attributes are as follows: (1) Cost (2) Response-time (3) Availability (4) Accessibility (5) Integrity (6) Reliability (7) Reservability (8) Security. There are the larger attribute value and the worse QoS for Cost and Response-time, which are defined by cost attribute. While there are the larger attribute value and the better QoS for Availability, Accessibility, Integrity, Reliability, Reservability and Security, which are described by benefit attribute. And differences are large among all kinds of QoS attribute values. In order to overcome these problems, all of attribute values are standardized between 0 and 1 and ranks in ascending sequence.

Formula of solving cost attribute:

$$q_{ij} = \begin{cases} 1 & x_j^{\max} = x_j^{\min} \\ \frac{x_j - x_j^{\min}}{x_j^{\max} - x_j^{\min}} & x_j^{\max} \neq x_j^{\min} \end{cases} \tag{2}$$



Formula of solving benefit attribute:

$$q_{ij} = \begin{cases} 1 & x_j^{\max} = x_j^{\min} \\ \frac{x_j^{\max} - x_{ij}}{x_j^{\max} - x_j^{\min}} & x_j^{\max} \neq x_j^{\min} \end{cases} \quad (3)$$

Where  $x_{ij}$  denotes attribute value before standardization,  $q_{ij}$  denotes attribute value after standardization.

We can gain the final QoS requirement value:

$$q_{QoS} = m_1 q_1^{ave} + m_2 q_2^{ave} + \dots + m_n q_n^{ave} \quad (4)$$

$$\sum_k m_k = 1, \quad 0 \leq m_k \leq 1, \quad k = 1, 2, \dots, n$$

Where  $m_k$  means importance degree of QoS attribute which is given by the consumer.

$q_k^{ave}$  is average of the QoS attribute.

(4) Bind and Invoke: Consumers bind and invoke service, if service requests are satisfied.

(5) Maintenance of local service information which consists of three suboperations.

① Localization of service information: The consumer can make use of familiar languages and terms to record service information and then retrieve service information by names, function and keys denominated by himself in local cache. Moreover the consumer can design local service discovery interface in the light of own interests and custom.

② Maintenance of information coherence: Service information may changes frequently during the whole lifetime of a service in the distributed system. Changed state such as service unregistry, service error and so on may lead to failure of service access. After service state changes, that the consumer still make use of the old service information stored in the local cache causes error or no response. So it is necessary to maintain the coherence of service information. Within this framework, we mainly adopt Immediate Query, Improved Immediate Query, Periodic Poll, Self-adaptive period poll, TTL-based information updating, and Self-adaptive TTL-based information updating so on.

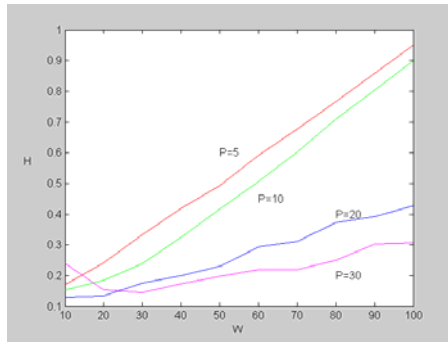
③ service overlay: With the accumulation of service information, the amount of service information of local cache will reach saturation due to memory capacity limitation of local cache. Beyond this, if new coming service will enter local cache, we have to choose a service to be substituted. Overlay strategy within this framework has many common characteristics with replacement strategy in the simulated memory system, but from the perspective of this framework, we firstly overload invalid service. And then we mainly adopt Least Recently Used (LRU) algorithm. However we have to take into account users' desired QoS, we bring forward Highest-Price Substituted algorithm, Lowest-Speed Substituted algorithm (LSS) and Lowest-Evaluation Substituted algorithm (LES) so on as complementary approaches.

### 3 Experiment Results

(1) Hitting Rate locally ( $H$ )

$H$  which directly determines speed of discovery is the key parameter in this framework. It is influenced by many factors, such as the largest number of services stored locally  $C$ , the total times of service access  $T_a$  during some time, intensive level of service request  $X$  (This consumer intensively accesses  $P$  services  $W$  times in  $C$  services for some time.  $W$  divided by  $P$  is equal to  $X$ ), Hitting rate locally  $H_1$  when service reservation is employed, Hitting rate locally when service reservation is not employed  $H_2$  ( $H_1$  and  $H_2$  equals  $H$ ).

In these experiments, we assume that  $T_a$  is equal to 100 and the number of services in UDDI is 50.  $C$  is an important factor to influence  $H$ . It is obvious that the larger  $C$ , the higher the hitting rate.  $X$  is another important factor that influences  $H$ . Generally, larger  $X$  is associated with higher hitting rate.



**Fig. 2.** We assume that  $C$  equals 10. There are four curves which respectively means when  $P=5$ ,  $P=10$ ,  $P=20$ ,  $P=30$ , the different hitting rate with the increase of service request  $W$ .

(2) Speed of Service Discovery

Now we will discuss some important factors, such as average time to discover a service locally  $T_l$ , average time to discover a service in UDDI  $T_r$ , time of discovering a service without using this mechanism  $T_1$ , time of discovering a service using this mechanism  $T_2$ . What we can see in  $T_2$  is that higher hitting rate locally means higher speed of service discovery; meanwhile, the difference of  $T_1$  and  $T_2$  is enormous, too. When hitting rate is rather high, the time saved is considerable.

$$T_1 = T_a \times T_r \quad \text{and} \quad T_2 = T_a \times H \times T_l + T_a \times (1 - H) \times T_r \tag{5}$$

$$\Delta T = T_1 - T_2 = T_a \times H \times (T_r - T_l) = T_a \times T_r \times H \times (1 - T_l/T_r) \tag{6}$$

In a general way,  $T_l$  is much less than  $T_r$ . Consequently,

$$\Delta T \approx T_a \times T_r \times H \tag{7}$$

(3) Recall Ratio

We compare “my service” [5] and this framework in recall ratio illustrated in Fig.3. It is seen that recall ratio of “my service” is only 52%, our recall ratio is up to 90%.

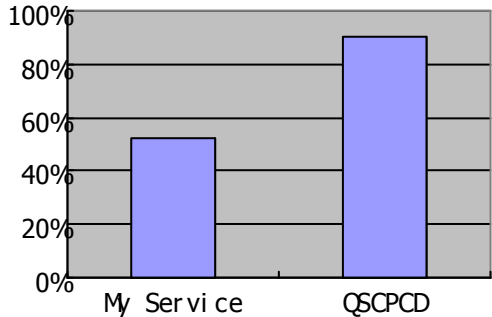


Fig. 3. Recall Ratio

(4) Precision ratio

We compare “my service” [5] and this framework in precision ratio illustrated in Fig.4. It is seen that precision ratio of “my service” is only 65%, our precision ratio is up to 93%.

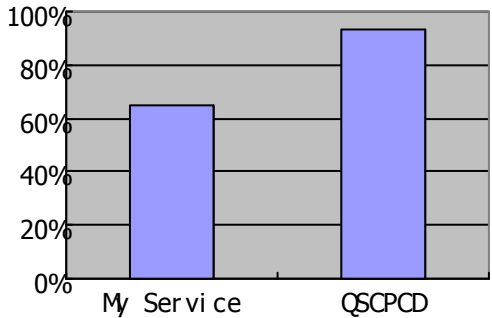


Fig. 4. Precision ratio

## 4 Conclusion

Current service discovery systems focus on the distribution and matching of service information and queries, with less emphasis on QoS and consumer personality. We argue that service discovery should consider QoS and consumer personality. As a result, we propose QSCPCD. A large number of services can be found locally, so this mechanism can speed up discovery of service. Furthermore it reduces access to remote servers. The consumer can make use of familiar languages and terms to record service

information and then retrieve service information by names, function and keys denominated by himself locally. Consequently it achieves satisfactory accessibility. In locally finding, initially we obtain a variety of related web services by calculating similarity, which can improve recall ratio. From previous results, we then find services based on consumer's requirement for QoS, which can improve precision ratio. As a result, it can speed up the discovery of services, ensure high precision ratio and recall ratio of service discovery, and take into QoS and consumer personality consideration.

## References

1. Kreger, H.: Web Services Conceptual Architecture (WSCA 1.0) (May 2001)
2. Tuecke, S., Czaikowski, K., Foster, I.: Grid service specification
3. <http://www.UDDI.org>
4. ShaikhAli, A., Rana, O.F., Al-Ali, R., Walker, D.W.: UDDIe: An Extended Registry for Web Services. In: Workshop on Service Oriented Computing: Models, Architectures and Applications at SAINT Conference. IEEE Computer Society Press, Los Alamitos (2003)
5. Jeckle, M., Zengler, B.: Active UDDI-an Extension to UDDI for Dynamic and Fault-Tolerant Service Invocation, <http://www.netobjectdays.org/pdf/02/papers/ws-rsd/1200.pdf>
6. Miles, S., Papay, J., Dialani, V., Luck, M., Decker, K., Payne, T., Moreau, L.: Personalised Grid Service Discover. In: Proceedings of 19th Annual UK Performance Engineering Workshop, pp. 131–140. University of Warwick, Coventry
7. Web Services Inspection Language (WSIL) 1.0
8. Feng, B., Liu, X., Li, W.: A consumer-oriented mechanism of service discovery. *Computer Research and Development* 40(12), 1787–1790 (2003)
9. Jianqiang, H.: Research on key technology in web services (doctoral thesis) National University of Defense Technology (2005)

# User Centric Homogeneity-Based Clustering Approach for Intelligence Computation

Yun Wei Zhao<sup>1</sup>, Chi-Hung Chi<sup>1</sup>, and Chen Ding<sup>2</sup>

<sup>1</sup> School of Software, Tsinghua University, Beijing China 100084

<sup>2</sup> Department of Computer Science, Ryerson University, Toronto, Canada  
chichihung@mail.tsinghua.edu.cn

**Abstract.** Clustering is a classic technique widely used in computation intelligence to study similarity measure among entities of interest. The output measurement of clustering, however, is often computation centric (e.g. number of peaks, K) instead of user centric (e.g. quality of the clusters). This creates a big gap between the algorithms and the users, in particular when they are applied to areas such as software services. To address this issue, we propose to use the expected homogeneity degree among entities within a given cluster as the input quality requirements specified by the users to drive the data clustering process. We evaluate the effectiveness of our proposal by modifying two most widely used clustering methods, K-means and hierarchical, according to the homogeneity degrees of the clustered output results.

**Keywords:** Artificial Intelligence, Data Clustering, Computational Intelligence, Homogeneity Measure.

## 1 Introduction

In many analytic domains such as information retrieval and service and business intelligence, clustering is one important technique to study the similarity measure among entities of interest [11][16]. There have been lots of previous research efforts on clustering algorithms, most of which are tailored to different focuses such as accuracy and performance.

In the selection process for an appropriate clustering algorithm, there are usually two different approaches. The first approach is to base on the specifications defined by a given application domain and algorithm. A good example is the K-means clustering used in image processing, where K is predefined by the algorithm. This is what most of the clustering algorithms are designed and optimized toward to. The second approach is user centric; it is based on the user expectation on the quality of the clustering result, independent of neither the algorithm nor the input data.

With the popularity of the software service provisioning, the second approach is getting increasingly attention because, by definition, service is user centric. In addition, there is another requirement that complicates the situation. Under a given

software service operation environment, the clustering requirements for intelligence analytics and decision making on the same data set are diversified, depending on: (i) whether it is from the user view or the provider view, and (ii) different needs from different providers or different users. These end into two design considerations for user-centric clustering algorithms: user perceived quality and real-time performance.

To provide a better user centric quality requirements specification capacity on the clustering algorithms and to support wide variation of clustering needs from different users/providers, we propose a new set of clustering algorithms based on user pre-defined homogeneity expectation. This homogeneity-based clustering algorithm only requires user to input his expected homogeneity degree on the clustering result instead of the distance or radius among the data entities under study. This is very important because from the user viewpoint, he is likely not able to have enough knowledge about the intrinsic properties of the given data set. The homogeneity index is a much better human's intuitive measurement concept than distance is, thus giving more meanings to non-professional / business users. Moreover, homogeneity is an efficient and effective way to describe data distribution, independent of whether the data is dispersive or intensive. This gives potentials for progressive clustering according to different requirements from different users and providers, which will result in good time performance.

The rest of the paper is organized as follows. Section 2 surveys on existing works related to this paper. Section 3 gives definition on the homogeneity index used in this paper. Section 4 describes our proposed homogeneity based clustering algorithms. Some theoretical analysis is also given in this section. Finally, the paper concludes in Section 5.

## 2 Related Work

Clustering techniques play a very important role in the description and visualization of the distribution of a given set of data entities. These data entities might come from the real-time monitored behaviors of software, systems, and human. They are widely used in e-business, image processing [1][2], and data pattern analysis [8].

As for software services, a lot of intelligence computation models have been proposed based on clustering. Usually, they are used in one of these two ways. Service providers can identify user and market requirements according to the clustering results of service user behavior and instance behavior, and this in turns will define the requirements set for future products and services. Service requesters can also make use of the clustering results for the selection of the best-fit service provider and for the definition of individual service quality requirements set. These are critical foundations to support trust [11][12], recommendation[13], and other intelligence computation in service intelligence provision system [7].

With numerous efforts having been put in clustering research, there are basically five main categories of clustering approaches. They are partitioning [14][15], hierarchical [6], density-based, model-based, and grid-based methods; each of them has its own strengths and weaknesses.

Most of clustering algorithms deal with vector data, each item of which is a numeric value. There is also a branch of clustering algorithms which specially deal with the interval data and/or more general symbolic data. Symbolic data analysis [17] is a unique way of analyzing discrete multi-valued data variables. It can handle variables of type numerical (traditional single point data), interval, categorical, enumeration, and modal.

Clustering is always based on similarity measure. Some of the most common similarity measures include Minkowski distance (absolute distance, Euclidean distance, and Chebyshev distance), Lanberra distance, and Mahalanobis distance, among which Euclidean distance is the most widely used. However, up to now, there is still no satisfactory distance measure defined for two vectors if each element of these vectors is a range data (with the possibility of open end) rather than a single value data.

Clustering can also be thought of as an unsupervised process that is different from classification because there are no predefined classes or examples that specify what kind of relations among the data are desirable. Thus, it needs validity analysis [5][9][10] to evaluate clustering algorithm from different aspects, and this helps to select an appropriate clustering method. In [10], the evaluation of a specified clustering algorithm can be based on Davies-Bouldin index, Dunn’s index, Calinski-Harabasz index, and a recently developed index I, which imposes an ordering of the clusters in terms of its goodness/validity.

Finally, a lot of work has been concentrated on evaluating the clustering results based on homogeneity analysis [3][4][5]. Clustering based on homogeneity is also under research, mainly in image processing [1][2]. However, the current clustering methods based on homogeneity often require a priori knowledge of data, and the user has to input an appropriate variable, such as distance between two vectors or radius of the cluster result set [1].

### 3 Definition of Homogeneity

In this section, we would like to give a precise definition on the homogeneity measure used in this paper. Consider a vector data set  $X = \{x_1, x_2, \dots, x_N\}$  and  $(x_i = \langle x_{i1}, \dots, x_{iM} \rangle)$ , where  $N$  is the number of entities in  $X$  and  $M$  is the number of attributes of the vector data, the definition of the normalization [2] of  $X$  is given by:

$$x_{ij}' = \frac{x_{ij} - \min_{l=1, \dots, n} \{x_{lj}\}}{\max_{l=1, \dots, n} \{x_{lj}\} - \min_{l=1, \dots, n} \{x_{lj}\}}$$

Thus, the value of each attribute is normalized to [0, 1]. Our definition of homogeneity is based on the variance of the data set. After certain steps of a given clustering algorithm are calculated, we have  $p$  clusters, each of which is a subset of  $X$ . The definition of a cluster  $C_t = \{x_1', x_2', \dots, x_{N_t}'\} (t = 1, \dots, p)$ , where  $N_t$  is the number of entities in  $C_t$  and  $\sum N_t = N$ , is given by

$$v(C_t) = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} d^2(x_i', \mu)}$$

where  $d(x_i', \mu)$  is a distance metric between two vectors and  $\mu$  is the mean of  $C_t$ , and the attribute of  $\mu_j$  ( $j = \{1, 2, \dots, M\}$ ) is given by

$$\mu_j = \frac{1}{N_t} \sum_{i=1}^{N_t} x_{ij}'$$

Note that particularly when  $X'$  is one-dimensional and  $d(x_i', \mu)$  is the Euclidean distance,  $v(X)$  becomes the statistical variance of the data set  $X'$ .

If all the data in one cluster are exactly the same, the variance will reach the minimum value 0. If, for each  $x_i'$  in  $C_t$ , all  $d(x_i', \mu)$  is equal to 0.5, the variance will reach the maximum value  $M^{1/2}/2$ , and we hope the homogeneity will reach the minimum value 0%. For example, a cluster contains only two vectors:  $\min(X')$  and  $\max(X')$  (there are also some other cases). That is to say,  $v(C_t) \in [0, M^{1/2}/2]$ , where  $M$  is the number of attributes of the vector.

Variance is an effective way to describe the dispersion of a data set, but it is harder to interpret and less intuitive. So what we propose is to map it to homogeneity. Since  $h(C_t) \in [0\%, 100\%]$ , we need a monotonically decreasing function  $f: h(C_t) = f(v(C_t))$ , which satisfies

$$\begin{cases} \{h(C_t) = 100\%, \text{ when } v(C_t) = 0 \\ \{h(C_t) = 0\%, \text{ when } v(C_t) = \frac{\sqrt{M}}{2} \end{cases}$$

There are three proposals of function  $f$ :

- (i) Exponential function

$$h(C_t) = e^{-\frac{v(C_t)}{\gamma}}, \gamma = -\frac{\frac{\sqrt{M}}{2}}{\ln 0.00000001}$$

- (ii) Parabolic function

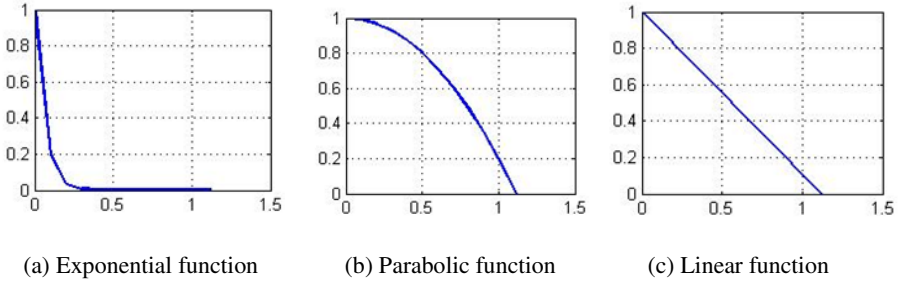
$$h(C_t) = av^2(C_t), a = -\frac{4}{M}$$

- (iii) Linear function

$$h(C_t) = kv(C_t) + 1, k = -\frac{2}{\sqrt{M}}$$



Their function curves with 5 attributes in the data vector are shown in Fig. 1. The tendency of decrease of the three functions varies a lot. Exponential function varies sharply at first and quickly goes down to nearly zero percent. As a result, the homogeneity will be mapped to a very low degree even when two data vectors are actually close to each other. Parabolic function varies smoothly at first, thus, it puts less restrictions on defining similarity between two data vectors.



**Fig. 1.** Curves for Three Homogeneity Function when  $M=5$

## 4 Homogeneity-Based Clustering

In this section, we are going to propose our user centric homogeneity based clustering algorithms, followed by its theoretical analysis. Two algorithms are proposed. They are: sequentially allocated and nearest allocated.

### 4.1 Sequentially Allocated Homogeneity-Based Clustering Algorithm

The idea of the sequentially allocated homogeneity-based clustering algorithms can be described as follows. The algorithm is initialized by choosing a vector as the first cluster randomly. The next step is to assign the remaining vectors to the already formed clusters in turn and calculate the corresponding homogeneity. If the new homogeneity is lower than the input homogeneity, the vector to be assigned will be deleted from the current cluster assigned to the next cluster. And the vector will form a new cluster if the homogeneity of each existing cluster added with the vector to be assigned is lower than the given homogeneity.

Based on the analysis of the characteristics of the mapping function in the definition of homogeneity in Section 3, the choice of the mapping function depends on the specific application circumstances. In case of the evenly-distributed raw data, the linear function will perform the best. However, there might be one extreme case in which people still want to make reasonable clustering even if most of the data have relatively rather high similarity. On the other extreme case, parabolic is a best choice.

Assume  $X$  is the normalized vector data set,  $N$  is the size of  $X$ ,  $M$  is the number of attributes,  $h$  is the expected homogeneity, and  $H(C[i])$  is the function of mapping variance to homogeneity. The pseudocode of the sequentially allocated homogeneity-based clustering algorithm is shown below:

Procedure SEQUENTIALLY-ALLOCATED-HOMOGENEITY-CLUSTER( $X, h$ ):

```

1  C[1]←CreateNewCluster()
2  addCluster(X[1],C[1])
3  clusterNum←1
4  flag←true
5  for i←2 to N
6      do for j←1 to clusterNum
7          do addCluster(X[i], C[j])
8              flag←true
9              h_temp←h(C[j])
10             If h_temp≤h
11                 remove(X[i],C[j])
12                 flag←false
13             else
14                 break
15         If flag=false
16             clusterNum←clusterNum+1
17             C[clusterNum]←CreateNewCluster()
18             addCluster(X[i],C[clusterNum])
19 return (C[] and clusterNum)

```

The time complexity of calculating the homogeneity of Cluster  $C[i]$  in line 9 is  $O(N_i)$ , where  $N_i$  is the size of cluster  $C[i]$ . And the execution time of the algorithm is  $O(N)$ , where  $N$  is the size of the input data set.

## 4.2 Nearest Allocated Homogeneity-Based Clustering Algorithm

The ideas of the nearest allocated homogeneity-based clustering algorithms can be described as follows. The algorithm is first initialized by choosing a vector as the first cluster randomly. Then the next step is to assign the remaining vectors in turn to the nearest cluster of the already formed ones and calculate the corresponding homogeneity. If the homogeneity is lower than the input homogeneity, the vector to be assigned will be deleted from the current cluster and it will form a new cluster. The pseudocode of the nearest allocated homogeneity-based clustering algorithm is given below:

Procedure NEAREST-ALLOCATED-HOMOGENEITY-CLUSTERING( $X, h$ ):

```

1  C[1]←CreateNewCluster()
2  addCluster(X[1],C[1])
3  clusterNum←1
4  flag←true
5  for i←2 to N
6      do d←maxNum
7          for j←1 to clusterNum
8              do Cen[j] ←getCentroids(C[j])
9                  dt←getEuclideanDistance(X[i],Cen[j])
10                 if dt<d
11                     then min←j
12         addCluster(X[i], C[min])
13         h_temp←h(C[j])
14         if h_temp<=h
15             then remove(X[i],C[j])
16             clusterNum←clusterNum+1
17             C[clusterNum]←CreateNewCluster()
18             addCluster(X[i],C[clusterNum])
19 return (C[] and clusterNum)

```

The time complexities of calculating the centroids of  $C[i]$  in line 8 and the homogeneity of Cluster  $C[i]$  in line 13 are both  $O(N_i)$ , where  $N_i$  is the size of cluster  $C[i]$ . And the execution time of the algorithm is  $O(N)$ , where  $N$  is the size of the input data set.

### 4.3 Theoretical Complexity Analysis

To evaluate the performance of our homogeneity-based algorithms, we would like to analyze at them from two different aspects: (i) time complexity, and (ii) characteristics of the clustering results.

**Table 1.** Complexity Analysis of Clustering Methods

	Time Complexity	Space Complexity
K_Means	$O(N)$	$O(N)$
Hierarchical	$O(N^2)$	$O(N^2)$
Sequentially Allocated Homogeneity	$O(N)$	$O(N)$
Nearest Allocated Homogeneity	$O(N)$	$O(N)$

Based on the analysis given in the last section, we can get the time complexity comparison shown in Table 1. In our comparison, we also include two most commonly used clustering approaches, K-means and hierarchical. Note that as for K\_Means, the time complexity cost is  $O(Nt)$ , where  $t$  is the number of iterations involved. And generally speaking, when the number of clusters is smaller, the

iteration time (or the time cost) is smaller. However, as for homogeneity-based clustering, the time cost is slightly larger when the input homogeneity is close to 100% and slightly smaller when the input the homogeneity is close to 0%.

Related to the characteristics of the clustering results, the sequentially allocated homogeneity-based clustering algorithm tends to generate clusters with relative population of the first cluster much greater than those of the other ones. However, this does not apply to the nearest allocated homogeneity-based clustering algorithm, of which the relative population distribution among different clusters are more similar

## 5 Conclusion

In this paper, we proposed a user centric homogeneity based clustering algorithm. We claim that this approach will be more suitable for environment where users care about the quality of the output result and they do not have knowledge on the distribution properties of the input data set. We also give a theoretical study on the strength and weakness of the algorithm by comparing with two other main classes of data clustering techniques (namely partitioning and hierarchical). Although our algorithm is sensitive to the input order of the raw data (which is similar to K-Means), the quality is a lot better. Besides, it can be applied to different circumstances simply by changing the mapping function in the definition of homogeneity. As for future work, homogeneity as an important criterion to theoretically define the number of clusters required in the pre-stage of hierarchical approaches and how to get a reasonable clustering results by considering both homogeneity and relative population will be our next focus.

**Acknowledgement.** This paper is supported by the China 863 HighTech Program under Project Number 2008AA01Z129 and National Natural Science Foundation of China, Project Number 61033006.

## References

1. Bajcsy, P., Ahuja, N.: Uniformity and Homogeneity Based Hierarchical Clustering. In: Proceedings of the 13th International Conference on Pattern Recognition, vol. 2, pp. 96–96 (1996)
2. Chen, L.F., Jiang, Q.S., Wang, S.R.: A Hierarchical Method for Determining the Number of Clusters. *Journal of Software* 19(1), 62–72 (2008)
3. He, J., Tan, A.H., Tan, C.L., Sung, S.Y.: On Quantitative Evaluation of Clustering Systems. *Clustering and Information Retrieval*, pp. 105–133. Kluwer Academic Publishers, Dordrecht (2003)
4. Mika, S.I.: On Evaluation of Clustering using Homogeneity Analysis. In: IEEE International Conference on Systems, Man, and Cybernetics (2000)
5. Heiser, W.J., Meulman, J.J.: Homogeneity Analysis: Exploring the Distribution of Variables and their Nonlinear Relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, pp. 179–209. Academic Press, New York (1994)

6. Bajcsy, P., Ahuja, N.: Location-and Density-based Hierarchical Clustering Using Similarity Analysis. In: Proceedings of the 13th International Conference on Pattern Recognition, vol. 2, pp. 96–96 (1998)
7. Vu, L.H., Hauswirth, M., Aberer, K.: QoS-based Service Selection and Ranking with Trust and Reputation Management. In: Proceedings of the International Conference on Cooperative Information Systems, pp. 446–483 (2005)
8. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster Analysis and Display of Genome-Wide Expression Patterns. Proceedings of the National Academy of Sciences of the United States of America 95(25), 14863–14868 (1998)
9. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster Validity Methods: Part I. ACM SIGMOD Record 31(2), 40–45 (2002)
10. Maulik, U., Bandyopadhyay, S.: Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(12), 1650–1654 (2002)
11. Vu, L.H., Hauswirth, M., Aberer, K.: QoS-Based Service Selection and Ranking with Trust and Reputation Management. In: Proceedings of OTM Confederated International Conferences on Cooperative Information Systems, pp. 466–483 (2005)
12. Golbeck, J., Parsia, B., Hendler, J.: Trust Networks on the Semantic Web. In: Proceedings of Cooperative Information Agents VII, vol. 2782, pp. 238–249 (2003)
13. Maulik, U., Bandyopadhyay, S.: An Adaptive Recommendation System Without Explicit Acquisition of User Relevance Feedback. Journal of Distributed and Parallel Databases 14(2), 173–192 (2003)
14. Li, S.M., Ding, C., Chi, C.H., Deng, J.: Adaptive Quality Recommendation Mechanism for Software Service Provisioning. In: Proceedings of IEEE International Conference on Web Service, pp. 169–176 (2008)
15. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
16. Liu, X.Z., Zhou, L., Huang, G., Mei, H.: Consumer-Centric Web Services Discovery and Subscription. In: Proceedings of IEEE International Conference on e-Business Engineering, pp. 543–550 (2007)
17. Diday, E., Noirhomme-Fraiture, M.: Symbolic Data Analysis and the SODAS Software. Wiley Interscience, Hoboken (2008)

# Mining Temporal Association Rules with Incremental Standing for Segment Progressive Filter

Mohsin Naqvi<sup>1</sup>, Kashif Hussain<sup>1</sup>, Sohail Asghar<sup>1</sup>, and Simon Fong<sup>2</sup>

<sup>1</sup> Center of Research in Data Engineering (CORDE),  
Mohammad Ali Jinnah University, Islamabad, Pakistan  
sohail.asg@gmail.com

<sup>2</sup> Department of Computer and Information Science,  
Faculty of Science and Technology,  
University of Macau, Macau SAR  
ccfong@umac.mo

**Abstract.** Association rule mining is a popular data mining technique which dredges up valuable relationships among different items in a dataset. A variant called temporal association rule mining finds relationship between items with respect to particular time periods. Databases are frequently updated; therefore temporal association rules that we discover should be corresponding to the updates in databases. Most of the existing data mining techniques however do not cover revising associate rules from the latest updates in the dataset. Some form of incremental mining technique is also needed to embrace the fresh elements that are updated continuously in the transaction database. In this paper we propose a technique that modifies the frequent patterns in pace with changes to the database over time. An Incremental Standing method for Segment Progressive Filter (ISPF) is proposed. ISPF algorithm is used for supporting the temporal association rule mining in transaction database with different exhibition periods. Our algorithm is optimized such that scanning of database is minimized. Scan reduction technique is applied here to generate all candidate k-item sets to form 2-candidate item sets directly. Working of the proposed algorithm is tested and illustrated with examples and a case study respectively.

**Keywords:** Temporal association rules.

## 1 Introduction

Association rule problem is first introduced by R. Agrawal [2] that is designed for finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Association relationship is useful in selective marketing, decision analysis and market basket analysis fields [1]. Several techniques have been developed for mining association rules [17] such as FP-Growth algorithm [18], mining of generalized and multi-level rules [19], constraint based rule mining and mining multi-dimensional rules [20].

Temporal association rule mining is first introduced by Wang, Yang and Muntz in years 1999-2001 together with the introduction of the TAR (Temporal Association Rule) algorithm [3]. Temporal association rule mining has been introduced in order to solve the problem on handling time-series by including time expression into association rules [4]. Temporal association helps to find the valuable relationship among the different item sets, in temporal database. Temporal association rules are largely different from traditional association rules by the fact that temporal association rules attempt to model temporal relationships in the data. There are different types of temporal association rules defined in the literature such as intertransaction rules, episode rules, trend dependencies, sequence association rules and calendric association rules.

In literature [4, 5, 6] most of existing techniques are developed based on temporal content analysis. New TAR algorithms that have been presented for general temporal association rule mining in database are PPCI algorithm [11], SPF [9], and ITRAM [4]. Temporal association rules have various kinds like Calendric Association rule [12], Cyclic Association rule [13], Association rule based on partition [14], progressive weighted miner [10], incremental temporal association rule [4] and periodic temporal association rule [15].

Temporal databases are known to be continually being updated or appended. Temporal association rule mining must synchronize with these updated transactions, without any loss of time granularity. Existing temporal association rule mining techniques cannot deal with the upcoming transactions of database as they might operate in batches. New rules may get omitted, and we need to address this issue. Let  $n$  be the number of partitions and  $m$  be the number of updates of the database. We need to generate the temporal association rules without loss of time granularity. In order to solve this problem, the INCREMENTAL STANDING FOR SEGMENT PROGRESSIVE FILTER (ISPF) algorithms are proposed. ISPF effectively divide the database item set with their common start and end times. It is a common phenomenon that items in the real transition database have their dissimilar exhibition periods.

The rest of paper is organized as follows. Section 2 provides the review of some related works. Section 3 describes the proposed algorithm. Performance result is shown in section 4. And section 5 gives the conclusion of the paper.

## 2 Related Works

Several algorithms have been proposed for mining the temporal association rules in temporal database. Among these algorithms, Tarek et al proposed the ITARM algorithm to discover the temporal frequent item set after the temporal transaction database has been updated [4]. The basic idea of ITARM algorithm depends on previously generated 2-candidate item set with their supports. ITRAM works as it checks first the extension of the pervious partition and attempts to find 2-candidate item set from the new partition; if it succeeds then it merges the current partition with the pervious partition, and from there it finds the 2-candidate item set. This approach is basically introduced to facilitate incremental mining techniques over an ever updating transaction database.

Another approach proposed by C. H. Lee et al, is progressive partition miner (PPM) [6]. In PPM the database is first partitioned by the size of time granularity. The PPM algorithm is applying with a filtering threshold mechanism on each partition of the database to prune out those cumulatively infrequent 2-itemsets. PPM also employs database scanning reduction technique. However, the limitation of this technique is its ability to deal with problems of incremental mining.

Cheng. Y. Chang et al proposed an algorithm called segmented progressive filter (SPF) [9] that is based on the Segmentation and progressive filtering. The basic idea of segmentation is to first divide the database into certain imposed time granularity. Then in illumination of exhibition period of each item, it further segments the database based on their common starting and ending times. For each part of the database it finds the 2-candidate item set with a cumulative filtering threshold. SPF applies also scanning reduction technique for generating candidate K-item set. After generating all candidates it generates the sub-candidate and counts for the value of support. Temporal databases are continuously updated or appended. But SPF does not perform any incremental mining technique on the refreshing database.

Moreover J. M. Ale et al expands the notion of association rule incorporation of the time to their frequent item sets [7]. Thus it tries to extend the existing non-temporal mining model by introducing the concept of temporal support. Discovery of association rule is done in a two-phase process; it first finds the frequent item set according to the lifespan of the item set and secondly it uses these frequent item sets to generate the rules. These rules are checked based on the confidence. In this proposed technique it however does not consider the updates of the database.

M. Chen et al developed a temporal association rule model to be used in video database for video event detection [5]. In this approach it captures the characteristics of temporal patterns with respect to the event of interest. M. Chen et al proposed their framework based on feature extraction, hierarchical temporal association mining and multimodal data mining. Often traditional association rule mining approaches use a manually assigned threshold. The advantage of Chen's model is the use of an adaptive mechanism for determining the essential threshold.

Byon et al proposed an Exponential Smoothing (ES) filter for temporal association rule mining [8]. ES filter takes two steps; one is to partition the database and then feed them into a Progressive Weighted Miner (PWM). PWM has a weight function that gives greater weights to recent data than old data; each is divided by equal period [10].

Ru Miao et al presented the idea of Apriori-extended mining periodic temporal association rules (MPTAR) [15]. Previously techniques of TAR did not consider the individual item exhibition period. MPTAR solved this problem, by including the exhibition period of individual item. Again MPTAR is a two-step periodic rule mining mechanism. The first step is mining the trend of continues attribute through cycle curve and the second step is calculating the period of the attribute. MPTAR did not define the cumulative threshold, and it is short of embracing upcoming transaction entries in the association rule mining.

Edi Winarko et al invented a new algorithm called, ARMADA (mining richer temporal association rules from interval-based data) [16]. While reading the database into memory, it counts the support of each state and generates frequent 1-patterns. By using a recursive find-then-index strategy, the algorithm discovers all temporal patterns from the in-memory database.



### 3 Temporal Association Rule Mining with ISPF

There are two major challenges in general temporal association rules methods which we will have to overcome. The first major challenge is to tackle the problem of updating the association rules while temporal database are continually being updated. The second challenge is the exhibition period of the item set in the database that should be allowed to differ from one to another. In the light of these challenges, we combine ISPF into TAR mining. ISPF consists of three major procedures; one is the database updating, the second is database segmentation and the third is candidate generation from the segments.

At each ending interval of the database update, the database is divided based on the imposed time granularity. The database is divided in the light of item set's common start and finish period. Then it checks the latest update of the database. By using this technique the number of segmentation is minimized and it is small when compared to the other previous methods. This feature provides the capability of filtering the candidate item set in either the forward or backward direction. After the segmentation it generates the 2-candidate item set in each sub database. When all sub databases are processed, all these 2-candidate item set are merged in union. After this, scan reduction technique is utilized over these candidate item sets and it generates the k-item set. As the last step of the algorithm, when all the k-item sets are generated, TIS and SIS are computed, and it counts the support for each rule.

We present the proposed algorithm, ISPF that is to be used for mining incremental temporal association rule, in the form of pseudo code as below. The advantage of this technique is its ability to deal with the problem of incremental mining techniques in mining temporal association rules.

#### Algorithm: ISPF

Input: transaction database DB, minimum support, time granularity, update of database db.

Output: frequent item sets.

Step#1: Divide the database based on the imposed time granularity.

Step#2: Check the update of the database. If the database is updated, append this update with pervious database transactions.

Step#3: Partition (in the light of exhibition period) the database based on either common star or end time

Step#4: Generate the 2candidate item set from each sub database

Step#4.1: Merge the new and pervious partition frequent 2-candidate item set.

Step#4.2: Count the relative support of each item set.

Step#4.3: Apply pruning.

Step#4.4: Proceed to next partition

Step#4.5: Go to step#4.1.

Step#5: Generate the k-item set through scan reduction technique.

Step#6: Count the support and apply pruning.

Step#7: Generate the sub candidate item.

Step#8: Count the support.

Step#9: Prune.

**Fig. 1.** Proposed Algorithm ISPF

## 4 Case Study Results

The efficacy of algorithm ISPF is tested by a given case study. Consider the transaction database shown in the Table 1. A set of time series in the database indicate the transaction records dated from January to March. They are the archival records which have already existed. A new portion of transactions that represent the incremental update of the database is recorded in the month of April. These transactions are shown in the last part of the table. Minimum support 30% and minimum confidence 75% are set for the experiment. The scanning direction of partitions 1 and 2 are from left to right, whereas the direction of partitions 3 and 4 are from right to left.

**Table 1.** Transaction database sample used in the experiment

	P1	Date	TID	Item set
		Jan 03	TID1	A F
		TID2	D C F	
		TID3	A C F	
		TID4	A D	
Database	P2	FEB03	TID5	C D
			TID6	B C D F
			TID7	A B C
			TID8	B
	P3	MAR03	TID9	E F
			TID10	B C F
			TID11	A B
			TID12	A E
UPDATE DATABASE	P4	APR03	TID13	AB
			TID14	A B E
			TID15	E
			TID16	B

Table 2 illustrates the start time and end time of the item set. By using this information, the database is partitioned based on either the common starting time or common ending time which could be optionally chosen by the user. The choice has little difference on the results when the sample size is large enough.

The results of the database partitioning are shown in Table 3. They include the partition 1-candidate item sets, their supports and the partition number of each candidate item set. The support values of the AD, AF, CF candidate item sets are equal to the defined threshold; AC and DF are pruned because their support values are lower than the defined threshold.

**Table 2.** The start and finish time of the item sets

Item	Start	End
A	Jan-02	Apr-02
B	Feb-02	Apr-02
C	Jan-02	Mar-02
D	Jan-02	Feb-02
E	Mar-02	Apr-02
F	Jan-02	Mar-02

**Table 3.** Partition 1

P1		
C	Start	Count
AC	1	1
AD	1	2
AF	1	2
CF	1	2
DF	1	1

Table 4 demonstrates the partition 1 frequent candidate item sets and partition 2 candidate item sets. Their support and partition values of the candidate item set are shown. The supports of the CF, BC, BD and CD candidate item sets are equal to or higher than the defined threshold; other item set are pruned because their support are less than the defined threshold.

**Table 4.** Partition P2+P3

P1+P2		
C	Start	Count
AD	1	2
AF	1	2
CF	1	3
AB	2	1
BC	2	2
BD	2	2
BF	2	1
CD	2	2

Table 5 illustrates the partition 4 frequent candidate item sets, as well as their support and partition values. AB is the only candidate item set that is qualified by the given minimum support; other item set are pruned because of their low support values.

**Table 5.** Partition 4 candidate item set 1

<b>P4</b>		
<b>C</b>	Start	Count
<b>AB</b>	4	2
<b>AE</b>	4	1
<b>BE</b>	4	1

Table 6 shows the partition frequent candidate item set of 4 & 3 candidate item sets, the support values and partition numbers of the candidate item set. AB and EF are the only candidate item sets that meet the defined threshold, other item sets are pruned away because their supports are less than the defined threshold.

After the scanning through all the sub databases, the resulting frequent candidate sets are AB BC BD CD CF and EF. Using scan reduction technique it generates k-item set. BCD and CDF are generated as a result.

**Table 6.** Partition 4 & 3 candidate item sets

<b>P3+P4</b>		
<b>C</b>	Start	Count
<b>AB</b>	4	3
<b>AF</b>	3	1
<b>BF</b>	3	1
<b>CE</b>	3	1
<b>CF</b>	3	1
<b>EF</b>	3	2

The problem of mining temporal association rules basically consists of two steps. Firstly it generates all frequent maximal temporal item sets called TIS, and the corresponding temporal sub-item sets namely SIS. SIS are generated based on these TIS. Both TIS and SIS item sets carry relative supports that would have to be greater than the predefined minimum value. The subsequent step is to derive all the frequent general temporal association rules that are frequent enough to meet the minimum required confidence value. Generating the frequent general temporal association rules is simple when the frequent TIS and SIS and their corresponding support values are known by scanning the whole database once.

In our experiment, a list of SIS and TIS candidate item sets are generated and their support values are shown in Table 7. We can observe that SIS are subset of the frequent item sets TIS. The qualified SIS candidates are A(2,4), B(2,3), B(2,2), B(2,4), C(1,2), C(1,3), C(2,2), D(2,2), C(2,3), D(1,2) and TIS candidate item sets are AB(2,4), BC(2,3), BD(2,2), CD(2,4), CF(1,3), EF(3,3), BCD(2,2).

**Table 7.** SIS and TIS

Candidate item set		count
<b>SIS</b>	A(2,4)	4
	B(2,3)	5
	B(2,2)	4
	B(2,4)	6
	C(1,2)	5
	C(1,3)	6
	C(2,2)	3
	C(2,3)	4
	D(1,2)	4
	D(2,2)	2
<b>TIS</b>	AB(2,4)	3
	BC(2,3)	2
	BD(2,2)	2
	CD(2,4)	2
	CF(1,3)	4
	EF(3,3)	1
	BCD(2,2)	2

The following table shows the information about the final frequent candidate item sets. The support values of both SIS and TIS, and their start and end partition information are shown. They are the ingredients for temporal association rule mining.

**Table 8.** Temporal item sets

Item set	S	E	S	E	TIS		
<b>AB</b>	A		B				
	1	4	2	4	AB(2,4)		
	B		C				
<b>BC</b>	2	4	1	3	BC(2,3)		
	B		D				
<b>BD</b>	2	4	1	2	BD(2,2)		
	C		D				
<b>CD</b>	1	3	1	2	CD(2,4)		
	C		F				
<b>CF</b>	1	3	1	3	CF(1,3)		
	E		F				
<b>EF</b>	3	4	1	3	EF(3,3)		
	B		C				
			D				
<b>BCD</b>	2	4	1	3	1	2	BCD(2,2)

Temporal association rule mining techniques are generating the rules based on the temporal information of transactions. The process is the same as the existing technique that was already published in previous research papers. In our case, the database is continuously being updated; the updates result in a number of useful new rules that can potentially be extracted, but they are neglected by the existing techniques of temporal association rules. To alleviate this issue, our proposed ISPF algorithm is used. The results from our experiment demonstrate the significance of the proposed work. For instance, in the case study, P4 contains the updated transactions of the database. The existing techniques would have generated the rules based on the database partitions P1, P2 and P3. These techniques however do not cater for the P4 (which holds the updated transactions of database). When the latest part of the database transactions is omitted, intuitively the resulting rules would miss out the elements of the latest information; therefore it is losing their timeliness and appeal in the knowledge discovery process. Our proposed algorithm generates rules covering all parts of the database, from P1 to P4. The tables above show that the proposed algorithm generated the most updated rules. These updated rules may contribute to effective and complete decision-making.

## 5 Conclusion and Future Work

Temporal association rules mining is a technique that incorporates the temporal characteristics in the association rule mining process over the frequent item sets. Temporal databases are known to be continually updated in reality. Existing temporal association rules mining techniques have not covered the most recently updated part of the data and hence the temporal association rules miss out the latest information elements. In this paper we explored the problem of incremental mining problem in general. In particular, we proposed the INCREMENTAL STANDING FOR SEGMENT PROGRESSIVE FILTER (ISPF) algorithm in order to align the updating of the database and the temporal association rules mining. ISPF first divides the database according to the common start and end times of the item sets, and it considers the updates of the temporal association rules. The case study results show the significance of the ISPF which performs better than the existing techniques and overcome the existing problem.

Our new method called ISPF theoretically should work with other variants of temporal association mining, as an add-on process rather than a revolutionary replacement. More complex examples would be tested in the future, as well as investigating the possibility of integrating ISPF into other mining algorithms.

As a future work, we opt to automate the proposed algorithm for real world applications domains, such as finance, marketing, medical, and security monitoring where real-time information streaming is typical and results of temporal association rules are critical. The other direction is to enhance the current user-interface of the temporal association mining program which facilitates the end-user to obtain temporal patterns and rules from relational temporal databases easily with a press of button. The temporal elements of the rules should be automatically evaluated and visualized for easy referencing.

## References

1. Tan, P.-N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Pearson Addison, London (2006)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rule Between sets of items in large database. *Proceeding of ACM SIGMOD*, 207–216 (1993)
3. Wang, W., Yang, Y., and Muntz, R.: Temporal Association Rules with Numerical Attributes. NCLA CSD Technical Report 990011 (1999)
4. Gharib, T.F., Nassar, H., Taha, M., Abraham, A.: An efficient algorithm for incremental mining of temporal association rules. *Data & Knowledge Engineering*, 800–815 (2010)
5. Chen, M., Chen, S.-C., Shyu, M.-L.: Hierarchical Temporal Association Mining for Video Event Detection in Video Databases
6. Lee, C.-H., Lin, C.-R., Chen, M.-S.: On Mining General Temporal Association Rule in Publication of database. In: *ICDM (2002)*
7. Ale, J.M., Rossi, G.H.: *An Approach to Discovering Temporal Rules*. ACM Press, New York (2000)
8. Byon, L.-N., Han, J.-H.: Fast for Temporal Association Rule in a Large Database. *Key Engineering Materials*, 287–279 (2005)
9. Chang, C.-Y., Chen, M.-S., Lee, C.-H.: Mining General Temporal Association Rule for item with different exhibition period. *IEEE, Los Alamitos (2002)*
10. Lee, C.-H., Ou, J.C., Chen, M.-S.: Progressive Weighted Miner: An efficient method for time constraints mining. In: Whang, K.-Y., Jeon, J., Shim, K., Srivastava, J. (eds.) *PAKDD 2003. LNCS (LNAI)*, vol. 2637, Springer, Heidelberg (2003)
11. Pandey, A., Pardasani, K.R.: PPCI algorithm for mining temporal association rules in large database. *International Journal of information and knowledge (April 2009)*
12. Lin, Y., Ning, P.: Discovering Calendric based temporal association rule. In: *Proceeding of the 8th International Symposium on Temporal and Reasoning (2001)*
13. Ozden, B., Ramaswamy, S., Silberschatz, A.: Cyclic Association rule. In: *Proceeding of International Conference on Data Engineering*, pp. 412–421
14. Chen, X., Petrounias, I.: A framework for temporal data mining. In: Quirchmayr, G., Bench-Capon, T.J.M., Schweighofer, E. (eds.) *DEXA 1998. LNCS*, vol. 1460, p. 796. Springer, Heidelberg (1998)
15. Miao, R., Shen, X.-J.: Construction of Periodic Temporal Association Rules in data mining. In: *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)*. IEEE, Los Alamitos (2010)
16. Winarko, E., Roddick, J.F.: ARMADA – An algorithm for discovering richer relative temporal association rules from interval based data. *Data & Knowledge Engineering*, 76–90 (2006)
17. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules in large database. In: *Proceeding of 20th International Conference on Very Large Databases*, pp. 478–499 (1994)
18. Han, J., Pei, J., Vin, V.: Mining frequent pattern without candidate generation. In: *Proceedings of 2000 ACM SIGMOD Int. Conference on Management of Data*, pp. 486–493 (2000)
19. Han, J., Fu, V.: Discovery of multiple level association rule from large database. In: *Proceedings of the 21th International Conference on Very Large Databases*, pp. 420–431 (1995)
20. Tung, A.H., Han, J., Lakshmanan, L.S., Ng, R.: Constraints based clustering in large databases. In: *Proceeding of 2001 International Conference on Databases Theory (2001)*

# Multi-way Association Clustering Analysis on Adaptive Real-Time Multicast Data

Sheneela Naz<sup>1</sup>, Sohail Asghar<sup>1</sup>, Simon Fong<sup>2</sup>, and Amir Qayyum<sup>3</sup>

<sup>1</sup> Center of Research in Data Engineering (CORDE),  
Mohammad Ali Jinnah University, Islamabad, Pakistan  
shahneela.cs@gmail.com, sohail.asghar@jinnah.edu.pk

<sup>2</sup> Department of Computer and Information Science,  
University of Macau, Macau SAR  
ccfong@umac.mo

<sup>3</sup> Center of Research in Networks & Telecommunication (CoReNeT),  
Mohammad Ali Jinnah University, Islamabad, Pakistan  
aqayyum@ieee.org

**Abstract.** Classification of real time multicast data using payload-based analysis is becoming increasingly difficult with many applications that a network supports. In this paper, we set our goal to identify the recurrent patterns and classification of transport layer data, as an effective measure of anomaly-based intrusion detection. These patterns are identified by using association rules techniques such as Apriori and clustering algorithms. A simulation experiment was configured to verify the efficacy of the algorithms. We are able to find an association between flow parameters for network traffic from the simulated data. This paper contributes a possible approach of analyzing behavior patterns for building a network traffic intrusion detection system and firewall at Transport layer, by using unsupervised association rule mining and clustering techniques.

**Keywords:** Clustering, association rules, real-time multicast, network security.

## 1 Introduction

A large variety of malicious attacks against computer network communication can be generally categorized into three aspects [1]: attacks against confidentiality, attack against integrity and attack against availability. The last two aspects of attacks (against confidentiality and against integrity) can be protected by manipulating the data with secrecy such as data encryption and data digestion methods; whereas attacks against the availability of a vulnerable computer network can be detected through the use of intrusion detection systems. Intrusion detection systems mainly function by two approaches on recognizing the users' behaviors, such as misuse detection and anomaly detection. Misuse detection tries to detect previously known attacks and flag the matching patterns. It assumes history of the attack is already known. In anomaly detection it checks on the network traffic behavior and measures how much it deviates



away from the normal network behavior. Anomaly detection is useful at detecting abnormal usage and it requires no prior knowledge on this new attack.

Multi-way Association Clustering Analysis on Adaptive Real-Time Multicast Data (MACAA) is an anomaly-based intrusion detection system (IDS) that uses a combination of association rules and clustering methods to identify malicious computer network activity from the traffic data. There are number of association rules techniques available in literature; e.g. they are Apriori, filtered associations and predictive Apriori, and clustering techniques are K-means, Y-means, DBSCAN etc.

Several techniques have been used in the past to classify network traffic flow. Jeffrey Erman et al. [2] classified network traffic by inspecting a list of Transport layer characteristics through the implementation of unsupervised approach such as clustering. The variables of Transport layer characteristics consist of duration of connection, total number of packets sent, size of packets and number of bytes sent. For network traffic classification (segmentation), K-Means and DBSCAN clustering algorithms are used in this paper. The results of these two algorithms are also compared with those by another clustering algorithm AutoClass. The performance of these three algorithms are studied together for identifying the pros and cons. Accuracy wise AutoClass algorithm performs very well, better than the other two clustering algorithms. K-Means and DBSCAN algorithms run quicker than AutoClass. As observed from the results of this performance evaluation, for network traffic classification K-Means algorithm seems to be more suitable other than the other two algorithms because its accuracy is relative high and its model building time is efficient. It takes approximately only one minute for model building where as DBSCAN takes approximately three minute and AutoClass takes approximately four and half hours in our experiments.

## 2 Background

There are a number of various techniques used to detect the anomalies in network traffic. According to Animesh Patcha et al. [4] mainly there are three types of techniques: statistical techniques, data-mining based methods, and machine learning based techniques. They are summarized in Fig 1.

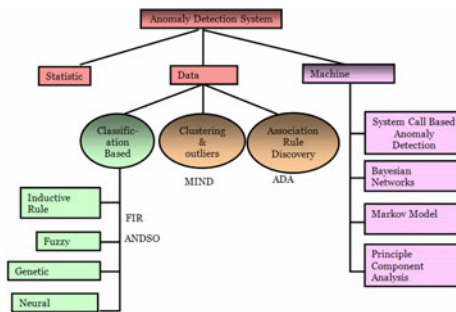


Fig. 1. Anomaly detection techniques

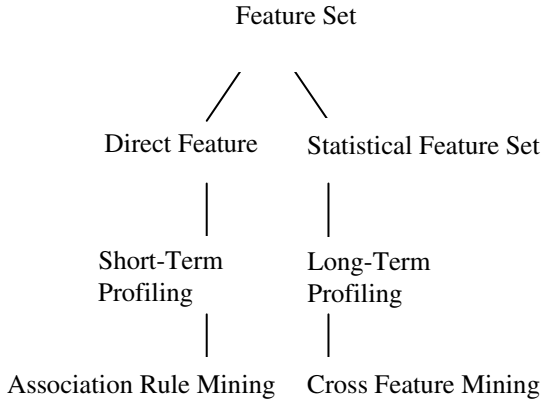
In statistical techniques, statistical measurements are used to reveal anomalies. Some well-known statistical anomaly detection techniques are Haystack [5], NIDES [6], Staniford et al. [7], and Ye et al. [8]. Data mining-based anomaly detection systems are largely built upon the following categories of techniques: Classification-based Clustering plus outlier detection and association rule discovery. Some examples of data mining-based anomaly detection systems are MINDS [9], ADAM [10] and FIRE [11] etc. Machine Learning category can be further branched into four sub-categories which are system-call based, anomaly detection, Bayesian networks, and Markove model and principle component analysis. Some machine learning based anomaly detection techniques are PHAD [12], ALAD [13] and Valdes et al. [14] etc.

In [1], Umang et al. proposed a network flow classification framework that performs two main tasks. First, it performs classification of network traffic, and second, it conducts the application behaviors profiling. Supported network traffic types by this framework include TCP, UDP and ICMP for wired or wireless data classification that was enabled by unsupervised clustering algorithms. This machine learning classification model contains three processes. They are Clustering, Transductive Classification technique and Association Rule Classification. These machine learning processes accept the flow data as input and perform the clustering on input flow data using K-Means and Modeling based clustering algorithms. After the clustering process is done, the next process Transductive Classification technique follows. The labels of the clusters are assigned. Then Association Rule Classification is applied for each cluster, then it proceeds to the final classification of given data flow as output. In this process Association Rule algorithm such as Apriori is applied. Under this framework, performance comparison between K-Means and Model based clustering algorithms with association rule techniques was model. The comparison shows that performance of K-Means is inefficient while Model based clustering performs efficiently and it also supports detection of new network traffic patterns.

The study exposed some short-comings of K-Means clustering algorithm which are the dependency and degeneracy of the number of required clusters. These two short comings are overcome in the work of Yu Guan et al [3]. Yu Guan et al proposed an intrusion detection clustering algorithm called Y-Means. The following steps describe the functional flow of Y-Means algorithm: partition the input data of total size  $n$  into  $k$  clusters where  $k$  lies between  $1 < k < n$ . After that, check a condition whether there is any empty cluster or not. If there are empty clusters then replace them out with newly created clusters. This process is repeated until there is no empty cluster remains. At the end the clusters are labeled according to the ratio of the instances. If the ratio is above a predefined threshold value, then these instances are labeled as normal; otherwise they are intrusive. Two major advantages can be found in Y-Means clustering algorithm; first is that it creates automatically an appropriate number of clusters, and second is raw log data can be used directly as training data without the need of labeling.

Founded on the Y-Means algorithm and its merits, Yu Liu et al proposed a hybrid technique which is used to detect the node based anomaly for ad-hoc communication networks [15]. This anomaly detection is considered as a hybrid approach that combines two data mining techniques. Associations rule mining techniques and cross-feature mining techniques are used together in action. This hybrid method takes two feature sets which are direct feature set and statistical feature set of MAC layer data

and network layer data respectively. Direct feature set targets on short-term node behavior profiling and statistical feature set targets on long-term node behavior profiling. For short-term profiling, this method applies associations rule mining techniques and for long-term profiling it uses cross-feature mining techniques. Fig 2 shows the feature set taxonomy according to this paper [15].



**Fig. 2.** Feature set taxonomy

Direct feature set of MAC layer is used to locate the source of attack within one hop perimeter. In that paper, multiple attack sources are evaluated through the Bayesian networks. The result analysis of both data mining techniques proves that the proposed IDS are effective because association rule provides precise detection performance whereas cross-feature approach is energy-efficient and effective in monitoring the network behavior.

Some classification based anomaly detection methods used the network traffic to detect the anomaly [11, 16, 17]. J. E. Dickerson et al. uses fuzzy logic to detect the malicious activity on network traffic such as TCP, UDP and ICMP data [11]. This technique is anomaly based intrusion detection system which is called Fuzzy Intrusion Recognition Engine (FIRE). This technique uses network input data features as fuzzy sets. These sets are used to define fuzzy rules. So, these rules can help to detect the individual attacks.

Several machine learning techniques are also used for anomaly detection in network traffic. Nong Ye proposed an anomaly diction technique which is used Markov chain model to detect intrusions attempting to hack into network systems. This technique represents the normal profile of temporal behavior. Probability is used to infer the normal behaviors and anomalous behaviors. If the probability is low then it implies the pattern is of anomalous behavior otherwise it is normal.

Another machine learning technique is used for host based anomaly intrusion detection. It processes sequences of system calls which form a multi layer intrusion detection model [18]. Their results indicate that this approach performs better in terms of accuracy and response time to detecting anomalous behavior of the software programs. This method is thus suitable for online intrusion detection.

For evaluation purposes we have used two protocols, one is Adaptive Smooth Multicast Protocol (ASMP) and second one is Packet-pair receiver-driven cumulative Layered Multicast (PLM). These protocols are multicast congestion control protocols and which are commonly applied over the transport layer. The trace files of these protocols were collected for analysis. Some descriptions of these protocols are defined below.

### 3 Multicast Congestion Control Protocols

Congestion control protocols play a pivoted role in reliable transfer of data in computer network. There are number of multicast congestion control protocols. We have selected two multicast congestion control protocols for simulation. We shall elaborate both of them briefly.

#### 3.1 Adaptive Smooth Multicast Protocol (ASMP)

ASMP was initially coined by [19]. It is a single-rate multicast transport protocol which is used for multimedia data transmission. It runs on top of UDP/RTP/RTCP protocols. In ASMP, sender and receiver share current information about network conditions through the use of RTCP sender and receiver's reports. In sender driven congestion control protocols, sender adjusts its transmission rate. ASMP is the sender driven protocol, so ASMP sender adjusts its sending rate according to the receiver's feedback reports. Receiver's feedback reports contain the receiving rate, as described in [20] which is calculated at each receiver according to the TCP analytical model.

Each receiver measures the following values such as packet loss rate, Round Trip Time, Delay Jitter and Congestion Indicators (CI), using the early congestion indication algorithm before the calculation of new TCP-friendly transmission rate. After calculating the current transmission rate, each receiver sends it to the sender by using the RTP/RTCP extensions. So, the sender receives the newly calculated receiving rate through receiver's feedback report and then it adjusts the sending rate keeping in consideration to the slowest receivers in the session. The main features of this protocol are: Smooth transmission rates, TCP-friendly behavior, and High bandwidth utilization. An advantage of this protocol is that it does not require any additional support from the routers or the underlying IP-multicast protocols. A disadvantage of this protocol is that, it does not show a very responsive behavior in varying network conditions because the gap between two successive RTCP feedback reports is very long.

#### 3.2 Packet-Pair Receiver-Driven Cumulative Layered Multicast (PLM)

Packet-pair receiver-driven cumulative Layered Multicast (PLM) was proposed by [21]. It is meant to address some deficiencies of Receiver Driven Layered Congestion Control multicast protocol (RLC). It is the multirate multicast congestion control protocol which is used for multimedia data transmission, such as, live audio/video. It runs on top of UDP/RTP protocols. In receiver driven congestion control protocols, the receiver is responsible for adapting the video transmission rate by subscribing and unsubscribing through various protocol layers. Therefore Receiver has an active role

while the sender has a passive role in adapting receiver based rate control. PLM is the receiver based congestion control mechanism, so the congestion control algorithm is implemented at the receiver side. Whereas at the sender side data is transmitted via cumulative layers and each layer packets are sent out in pairs. PLM defines two basic mechanisms: Receiver-side Packet-pair Probe (PP) and Fair Queuing (FQ).

Receiver-side Packet-pair Probe (PP) mechanism is used to estimate the currently available bandwidth and Fair Queuing (FQ) mechanism is used at each router. PLM assumes fair scheduler network and deploys a fair queuing mechanism at routers. It relies on a fair scheduler to ensure fairness, including intra-protocol fairness, inter-protocol fairness and TCP friendliness. PLM has some advantages over RLM and RLC. RLM and RLC produce losses at joint attempts, whereas PLM does not suffer any loss in discovering the available bandwidth. It has a fast convergence for rate adaptation.

## 4 Performance Evaluation

The input log files that are to be used in the performance evaluation experiments are generated by a simulation program NS2. The simulation is configured with multicast congestion protocols ASMP and PLM.

### 4.1 Simulation Setup and Topology

For the sake of the simulation-based experiments, ns simulator version 2.33 was used and integrated with the available code of PLM (built-in) and ASMP (asmp V1.1). Simulation topology was created in the NsWorkBench (nsBench v1.0) environment.

Fig. 3 shows the topology which was designed for checking the fairness and responsiveness of these two protocols. The bottleneck link from R1 to R2 has bandwidth of 600 Kbps and time delay of 8ms. Interior links from R2 to R3, R2 to R4 and R2 to R5 have 10Mbps bandwidth and delay of 8ms each. All exterior links have 10Mbps bandwidth and delay of 8ms. Each simulation trial was run for 250 seconds. Data rate of TCP and UDP connection is set at 500Kbps. Initially data rate for multicast protocol is 500Kbps. There is one multicast session (ASMP/PLM) and three TCP/UDP connections in total.

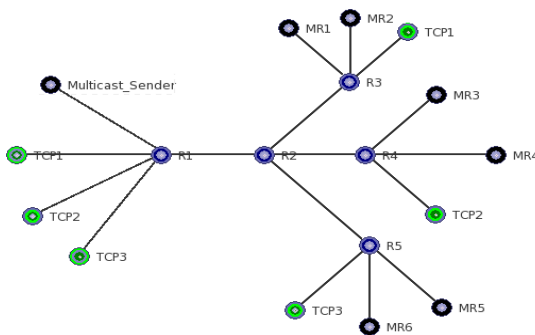


Fig. 3. Simulation Topology

## 4.2 Feature of Interest

Transport layer data is rich in a variety of intrinsic features. Interested features from the perspective of the experiment are extracted from the Transport layer protocols which are used to multicast the multimedia data such as Adaptive Smooth Multicast Protocol (ASMP) and Packet-pair receiver-driven cumulative Layered Multicast (PLM) which are run over the UDP protocol. In PLM/ASMP trace file, there are three packet types such as data, prune and graft. Data packet contains the multimedia data. The proposed feature set and its value space is illustrated in Table 1.

**Table 1.** Feature Set

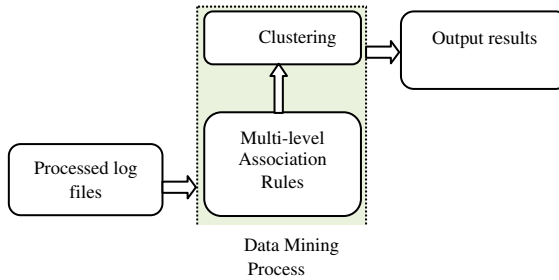
Feature	Feature Value
Flow Direction	Send, Receive, Drop
Source Address (SA)	One/many
Destination Address (DS)	One/many
Traffic Type	CBR/UDP, RTP/UDP,TCP
Packet Type	DATA, Prune, Graft
Sending Rate	Bytes
Packet Id	Number
Sequence #	Number
Time	Seconds

Simulation experiment yields trace file of large number of recordset. Data mining community is well aware of the fact that piles of data are prone to yield useful information. However in order to extract useful patterns, it is a mandatory requirement to perform data preprocessing activity. Researchers have argued that preprocessing activity normally consumes much of the time of overall experiment and the same was true in our case. The trace files essentially contain records of data packets that traversed across the specific links in the network topology and the patterns of these accesses are shaped by the specified protocols along those links. This processed trace file/log file is used as input to the data mining process unit in which the data mining algorithms are implemented.

Fig. 4 outlines the flow of data mining process model. It consists of multilevel association rule mining and clustering technique. Association rule is a supervised learning technique while clustering is an unsupervised learning technique. Association rule mining discovers the hidden relationships among the data which are sometimes known as casual relations. Multilevel association rule mining unit is driven by Apriori algorithm. The data mining unit reads in the input trace files and creates a number of association rules by using Apriori algorithm. Data of the features of interest that

occurred frequently together in the access records are sorted out. This sorted data is identified as association rules.

The next step is extracting the association rules from the trace files; the process classifies these extracted rules into groups according to their semantic meanings by applying K-Means clustering algorithm. This unsupervised clustering technique groups together association rules with the criteria of similar characteristics.



**Fig. 4.** Data Mining Process Model

## 5 Multilevel Association Rules

Multilevel association rule mining is aimed towards discovering the relations between data items at multiple levels in a given dataset. The data is subjected to normalization process with formats of RTP, UDP and TCP protocols. Multilevel association rules have the provision to recognize the pattern of computer network traffic while building a set of heuristic rules for prediction. According to the features of interest given in table 1, a transaction record is an artificially synthetic instantiation of the following feature set:

{Flow direction, SA, DA, traffic type, packet type, sending rate, pktid, seq#, time}

The rules shown in the table 2 have been generated using the orange tool. Where SA, DA stands for *Sender Address* and *Destination Address*

An example association rule looks like this:

HEAD (rec=multicast) → BODY (Type=cbr, packetsize=500, fid=1)

For cbr, packet is received on the multicast address. This packet size is of 500 units so is quantified as 500 only. cbr packet type and flow id is 1 with support of 18 and confidence of 1.0. When we track back this rule, we are accurately able to classify this rule as the cbr packet.

**Table 2.** Apriori Based Association Rules

Packet Type	Rule	C.	Lift	Str.	Cov.	Leverage
TCP	Type=tcp → pktsize=540	1.0	4.414	1.000	0.227	0.175
RTP	Type=rtp → pktsize=1000	1.0	4.412	1.000	0.225	0.175
ACK	Type=ack → pktsize=40	1.0	4.430	1.000	0.226	0.175
CBR	rec=multicast → Type=cbr, pktsize=500, fid=1	1.0	2.238	1.184	0.377	0.209

Each rule is associated with the following performance parameters [22] that might indicate how interesting or significant the rules are: Lift indicates the strength of the rule because it defines the ratio of the probability that antecedent and consequent occur together. Confidence is the number of cases in which the rule is correct relative to the number of cases in which it is applicable.

Rules with lower value of confidence, lift, leverage, strength and coverage are filtered out. This leaves only the most important rules which are ranked and appeared on the top in descending order. Fig. 5 shows the taxonomy of packet in a hierarchical representation. This representation which is a rooted tree represents the type, packet size, receiving status and flow id at first level. In the next level packet type is split into cbr, tcp, rtp and ack whereas packet sizes is grouped into 500, 540 and 40. Flow id is branched out into numeral values 1, 2 and 3. Packet receiving status is further branched out into multicast etc. These branches can be further divided into number of branches. Fig. 6 represents the multi-way association rules. Conceptually both of them are same in implantations.

## 6 Clustering Analysis

The clustering in data mining is a well renowned technique for grouping similar objects. In literature various algorithms and techniques have been discussed and proposed. In our proposed architecture we have used K-Means clustering. K mean clusters are based on a variety of measures including manhattan distance, euclidean distance etc. In each iteration, means all of the observations is calculated and then centroid is re calculated until the centroid becomes stable. The mean values of randomly selected centers converge to the appropriate cluster.

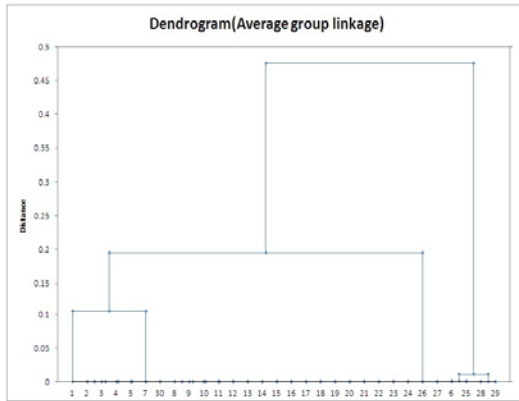
The reason to choose this technique is its simplicity and superior performance over its peer methods with reasonable accuracy. However choosing the initial seed is tricky and difficult for achieving best results with high precision.

After obtaining the multi-way association rules, we applied the hierarchical clustering technique. This technique is infact an extended version of K-Means. The end product of this technique resulted in four hierarchical clusters. Among various



similarity measures, we chose Euclidean distance whereas the clustering method is average group linkage. Fig 7 shows the dendrogram of these hierarchical clusters.

The clustering analysis is used in this paper is K-Means clustering because of its speed and reasonable accuracy. It is however difficult to estimate the  $k$  value for each new dataset as well as to maintain high accuracy and precision given the chosen  $k$  value. It is due to the fact that the resulting K-Means clusters are partitioned by mean values iteratively. The mean values of randomly selected centers converge to the appropriate cluster.



**Fig. 7.** Dendrogram of Four Clusters

After obtaining the multi-way association rules, we apply the hierarchical clustering technique (which is an extended version of K-Means) on these rules and obtained the four hierarchical clusters. Similarity measure applied was Euclidean distance and clustering method is average group linkage. Fig 7 shows the dendrogram of these heretical clusters.

Heretical clustering produces strong packet type clusters in the data. The Apriori based association rule classifier finds stronger association between flow parameters. The rules with high lift and confidence values represent strong relation to the application. Hence, the rules set help us derive behavior pattern for a particular packet type by looking at their cause-and-effect relations or casual relations. We trace back the flows to the main trace file and we observe a strong probability that those flows belong to a particular type of packet class.

Cluster 1: This cluster shows the transitions which have the multicast addresses. Cluster1 also contains the three sub-clusters.

Cluster 2: This cluster contains the CBR traffic. CBR traffic is also a multicast traffic. This is one-to-many correspondence between number of senders and receivers

Cluster 3: This cluster shows the TCP type of traffic with acknowledge information.

Cluster 4: This cluster shows the RTP transitions. These transactions also contain the multicast addresses.

## 7 Conclusion

In this paper, we presented the analysis of computer network traffic behavioral pattern for the Transport layer of TCP/IP protocol stack, using unsupervised association rule mining and clustering techniques. K-Means clustering with association rule mining techniques were shown to achieve high accuracy. We have show that our model is able to detect new behavior patterns for multicast traffic. For the future work, we are planning to extend this model to be scalable for a very large trace of dataset which are normally generated in the complex scenario based simulation experiment while soliciting behavior patterns for a wider range of network traffic.

## References

1. Chaudhary, U.K., Papapanagiotou, I., Devetsikiotis, M.: Flow Classification Using Clustering and Association Rule Mining (2010)
2. Erman, J., Arlitt, M., Mahanti, A.: Traffic Classification Using Clustering Algorithms. In: MineNet 2006 Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data (2006)
3. Guan, Y., Ghorbani, A.A., Belacel, N.: Y-MEANS: A Clustering Method for Intrusion Detection. In: Canadian Conference on Electrical and Computer Engineering CCECE, vol. 2, pp. 1083–1086 (2003)
4. Patcha, A., Park, J.-M.: An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends. *Computer Networks* (2007)
5. Smaha, S.E., Haystack.: An Intrusion Detection System. In: Proceedings of the IEEE Fourth Aerospace Computer Security Applications Conference, Orlando, FL, pp. 37–44 (1988)
6. Anderson, D., Frivold, T., Tamaru, A., Valdes, A.: Next Generation Intrusion Detection Expert System (NIDES). Software Users Manual, Beta-Update release, Computer Science Laboratory, SRI International, Menlo Park, CA, USA, Technical Report SRI-CSL-95-0 (May 1994)
7. Staniford, S., Hoagland, J.A., McAlerney, J.M.: Practical Automated Detection of Stealthy Portscans. *Journal of Computer Security* 10, 105–136 (2002)
8. Ye, N., Emran, S.M., Chen, Q., Vilbert, S.: Multivariate Statistical Analysis of Audit Trails For Host-Based Intrusion Detection. *IEEE Transactions on Computers* 51, 810–820 (2002)
9. Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P.-N., Kumar, V., Srivastava, J., Dokas, P.: The MINDS - Minnesota Intrusion Detection System. In: Next Generation Data Mining. MIT Press, Boston (2004)
10. Barbara´, D., Couto, J., Jajodia, S., Wu, N.: ADAM: a Testbed for Exploring the Use of Data Mining in Intrusion Detection. *ACM SIGMOD Record: SPECIAL ISSUE: Special Section on Data Mining for Intrusion Detection and Threat Analysis* 30, 15–24 (2001)
11. Dickerson, J.E., Dickerson, J.A.: Fuzzy Network Profiling for Intrusion Detection. In: Proceedings of the 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), Atlanta, GA, pp. 301–306 (2000)
12. Mahoney, M.V., Chan, P.K.: PHAD Packet Header Anomaly Detection for Identifying Hostile Network Traffic. Department of Computer Sciences, Florida Institute of Technology, Melbourne, FL, USA, Technical Report CS- 2001-4 (April 2001)

13. Mahoney, M.V., Chan, P.K.: Learning Non Stationary Models of Normal Network Traffic for Detecting Novel Attacks. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, pp. 376–385 (2002)
14. Valdes, A., Skinner, K.: Adaptive Model-Based Monitoring for Cyber Attack Detection. In: Recent Advances in Intrusion Detection Toulouse, France, pp. 80–92 (2000)
15. Liu, Y., Li, Y., Man, H.: A Hybrid Data Mining Anomaly Detection Technique in Ad Hoc Networks. *Int. J. Wireless and Mobile Computing* 2(1) (2007)
16. Lee, W., Stolfo, S.J.: Data Mining Approaches for Intrusion Detection. In: Proceedings of the 7th USENIX Security Symposium (SECURITY 1998), Berkeley, CA, USA, pp. 79–94 (1998)
17. Ramadas, M., Tjaden, S.O.B.: Detecting Anomalous Network Traffic with Self-Organizing Maps. In: Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection, Pittsburgh, PA, USA, pp. 36–54 (2003)
18. Hoang, X.D., Hu, J., Bertok, P.: A Multi-layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls. In: The 11th IEEE International Conference on Networks, ICON 2003, pp. 531–536 (2003)
19. Bouras, C., Gkamas, A., Kioumourtzis, G.: Adaptive Smooth Multicast Protocol for Multimedia Data Transmission. In: 2008 International Symposium on Performance Evaluation of Computer and Telecommunication Systems – SPECTS 2008, Edinburgh, UK, pp. 16–18 (June 2008)
20. Padhye, et al.: A model based TCP - friendly rate control protocol. In: Proc. International Workshop on Network (1999)
21. Legout, A., Biersack, E.W.: PLM: Fast Convergence for Cumulative Layered Multicast Transmission. In: Proceedings of ACM SIGMETRICS 2000, pp. 13–22 (2000)
22. Borgelt, C., Kruse, R.: Induction of association rules: Apriori implementation. In: Proceedings of the 15th Symposium on Computational Statistics, p. 395. Physica Verlag, Berlin (2002)

# Multi Level Mining of Warehouse Schema

Muhammad Usman and Russel Pears

School of Computing and Mathematical Science,  
Auckland University of Technology, Auckland, New Zealand  
{Muhammad.Usman, Russel.Pears}@aut.ac.nz

**Abstract.** The two mature disciplines, namely Data Mining and Data Warehousing have broadly the same set of objectives. Yet, they have developed largely separate from each other resulting in different techniques being used in each discipline. It has been recognized that mining techniques developed for pattern recognition such as Clustering and Visualization can assist in designing data warehouse schema. However, a suitable methodology is required for the seamless integration of mining methods in the design of warehouse schema. In previous work, we presented a methodology that employs hierarchical clustering to derive a tree structure that can be used by a data warehouse designer to build a schema. We believe that, in order to strengthen the decision making process, there is a strong need for a method that automatically extracts knowledge present at different levels of abstraction from a warehouse. We demonstrate with examples how mining at different levels of a hierarchical warehouse schema can give new insights about the underlying data cluster which not only helps in building more meaningful dimensions and facts for data warehouse design but can also improve the decision making process.

**Keywords:** Clustering, Data Warehouse, Data Mining, Warehouse Schema, Visualization.

## 1 Introduction

Data repositories are constantly growing in size due to the enormous use of information systems in various application domains [1]. Such repositories can be used to identify useful information and hidden patterns in the data by applying suitable data mining techniques along with the knowledge discovery mechanisms [2]. Efficient analysis of data using the modern analytical and mining techniques is an important step in the knowledge discovery process. One of the big hurdles to this efficient analysis of data is the presence of mixed numeric and nominal attributes in real world datasets. A vast majority of algorithms, methods and techniques have been proposed in the literature for analyzing numeric data but little attention has been given to overcome the problem of mixed data analysis. Traditional methodologies assume variables are numeric, but as application areas have grown from the scientific and engineering domains to the biological, engineering, and social domains, one has to deal with features, such as country, color, shape, and type of disease, that are nominal

valued [1]. In addition to the problem of efficient analysis of mixed data, high cardinality nominal variables with large number of distinct values such as product codes, country names and model types are not only difficult to analyze but also require effective visual exploration methods [2]. Visualization techniques are becoming increasingly important for the analysis and exploration of large multidimensional data sets [3]. However, the effectiveness of visualization techniques such as parallel coordinates [4, 5] is determined by the order in which attributes are displayed [6]. Moreover, accurate spacing among the attribute values is mandatory to recognize the semantic inter-relations that exist in the underlying data.

In our previous work [27], we focused on the seamless integration of data mining techniques into the design of data warehouses. Designing a data warehouse is a complex task, which involves knowledge of business processes in the domain of discourse, understanding the structural and behavioral system's conceptual model, and familiarity with data warehouse technologies [8]. The two disciplines, namely data warehousing and data mining are both mature in their own right but little research has been carried out in integrating these two strands of research for the purpose to support data warehouse schema design. Given the sheer volume of data normally involved in the building of a data warehouse, a case can be made for automated support in the construction of the warehouse schema in order to capture the patterns and trends that are needed in schema design. The human data warehouse designer, with his/her limited knowledge of the domain can supply some of these patterns, but there will always be cases where such knowledge needs to be augmented with automatic pattern generation methods. To overcome these problems, we proposed a methodology in our earlier work for the seamless integration of data mining techniques into data warehousing design. We employed hierarchical clustering and parallel coordinates techniques to aid the automatic design of the well-known STAR schema.

The focus of the previous work was the design of the schema with little consideration as to how the knowledge embedded in the schema could be extracted. In this research we propose a methodology that describes how knowledge is distributed across the different levels of a hierarchical schema structure. We develop an iterative method that explores the similarities and differences in information contained across consecutive levels in the hierarchy. The main motivation for such an analysis was that the presentation of information at different levels of abstraction provides the decision maker to get a better understanding of the patterns and trends present in the data. In the case study that we present in this research we observed that sharp differences in patterns emerged as the data was explored at lower levels of granularity, or abstraction. The rest of the paper is organized as follows: Section 2 presents an overview of prior work relevant to our research objectives. In Section 3, we present our methodology for knowledge extraction from a hierarchical schema. Section 4 presents the implementation details. Our case study results are discussed in Section 5 and in Section 6 we conclude the paper with a summary of the achievements of the research and discuss some possible directions for future research.

## 2 Related Work

Real world data sets consist of a mix of numeric and nominal data. While an abundance of algorithms exist for clustering for numerical data, little effort has been directed at clustering nominal data. For scalable clustering of mixed data, orthogonal partitioning clustering algorithm [13] was introduced which was later extended by the authors in [14] for the purpose of clustering large databases with numeric and nominal values using orthogonal projections. To achieve a similar objective, a fuzzy clustering algorithm [15] based on probabilistic distance feature, an agglomerative algorithm based on distinctness heuristics as well as the Evidence Based Spectral Clustering (EBSC) algorithm [16] based on evidence accumulation were introduced in the recent past. On the other hand, authors in [17] introduced three different distance measure functions based on *Mahalanobis-type* distance measure for the efficient analysis of mixed data. Another distance measure, using the cost function based on co-occurrence of categorical values was offered to overcome the limitations of the traditional k-means algorithm and to support mixed data analysis [13]. Hierarchical clustering has also been employed by the authors in [18] for mixed data based on distance hierarchy.

Our work is similar in terms of efficient analysis of mixed data but the proposed work is neither proposing a new algorithm nor a fresh distance measure to conquer mixed data analysis problem. Instead, we employ a hierarchical clustering technique as an initial step in the proposed methodology to produce natural clusters from the data based on numeric variables. For nominal data, we use a visualization technique known as parallel coordinates to identify and analyze the nominal variables within each data cluster. The Parallel coordinate technique has been used by in the past for the effective visualization of data. Authors in [4] used this technique with the help of some extensions for the effective exploration of complex data sets. Similarly, another efficient approach to construct frequency and density plots from parallel coordinates was introduced [19]. In parallel coordinates technique, order and spacing among the variables on the coordinates play a vital role for the extraction of useful information. Meaningful spacing among high cardinality values helps in interpretation of results and in recognition of meaningful patterns from the underlying data values. A major limitation of parallel coordinates technique is that it is suitable for small number of dimensions or variables. Its effectiveness is inversely proportional to the number of variables. In our proposed methodology, we use this technique only to visualize the nominal data present in a data cluster. This is an effective strategy as there is less number of nominal variables in the data clusters as compared to numeric variables. Furthermore, we do not visualize ordinal variables because the basic reason for using this visualization technique is to identify natural grouping among the categorical values for each nominal variable. For instance, we visualize variables such as Country which has more distinct values as compared to an ordinal variable such as Sex which has only two distinct values, namely male or female.

Another source of related work is the automatic generation of warehouse schema. A Model-transformation architecture has been proposed in order to facilitate the automatic schema generation process [20]. Their implementation was based on an open source development platform to automatically generate schema from a conceptual multidimensional model. Likewise, an Object-process-based Data

Warehouse Construction (ODWC) method was suggested for the purpose of automatic schema construction [8]. The authors in [21], proposed a technique that takes a list of database schema in the form of entity-relationship(ER) model as input and produce a schema as output. Another similar approach using ER diagrams was presented a few years later using a prototype system known as SAMSTAR[22]. This research shares the same objective of automatic schema generation as we use hierarchical clustering and parallel coordinates techniques in conjunction with each other to identify the various dimensional groupings or hierarchies and measure within a data cluster. These natural groupings based on the underlying dataset can assist the schema design and generation process.

Finally, we discuss the work relevant to the utilization of hierarchical clustering technique in data warehouse environment. Recently, a conceptual model was proposed for combining enhanced OLAP and data mining system [23]. However, the proposed model was not integrated with automatic generation of schema and lacked experimental evaluation as well. Another architecture was proposed by the authors in [24] to extend the work of [25] and provided a way of integrated enhancement of warehouse schema using self-organizing neural networks.

### 3 Proposed Methodology

In this section, we present our methodology for knowledge extraction from a multi level warehouse schema. Figure 1 gives the overview of the proposed methodology.

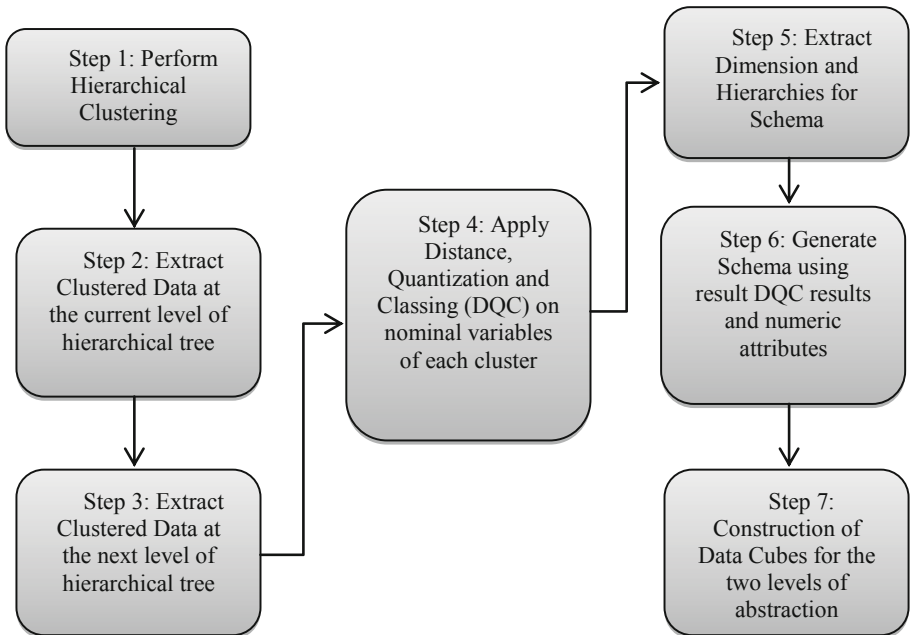


Fig. 1. Overview of the extended methodology

As Figure 1 shows, we construct cubes for consecutive levels of data abstraction. With the presence of data cubes at different levels, the analyst can explore the similarities and differences among the different attributes present in the data clusters at various levels of detail. We believe that without this two level exploration of cube data it is not easy to understand the underlying structure of the data and lack of understanding can lead to weak analysis and poor decision making. The description of each step involved in the methodology is explained as follows.

### **3.1 Perform Hierarchical Clustering**

This step helps to generate the warehouse schema at different levels of abstraction. In this step, the agglomerative hierarchical clustering algorithm is applied to the data set to generate clusters based on a similarity measure. The reason for choosing agglomerative hierarchical clustering is that it tends to produce natural clusters instead of performing unnecessary merges and splits like other clustering approaches. Furthermore, it allows users to set parameters to determine the proper number of clusters. As most of the clustering algorithms works well on numeric variables, in this step, we target the numeric variables in order to get optimal results.

### **3.2 Extract Data at the Current Level of the Hierarchical Tree**

In this step, we determine the number of clusters produced and extract the clustered data present in each of the clusters at the current level of abstraction. The nominal attributes present in each of these clusters are the possible dimensions of the dataset. Furthermore, the numerical variables present within each cluster at this level are the measures within the cluster. These measures become the input for the automatic schema generation process. In data warehouses the numerical attributes represent core potential measures which the analysts want to visualize from different dimensional perspectives.

### **3.3 Extract Data at the Next Level of the Hierarchical Tree**

Similar to the previous step, we go one level down in the hierarchy and extract the clusters present at the next level. This 2nd level clusters are generated by splitting the top level clusters. We mark the clusters in the 2nd level and extract the numeric and nominal variables from each of the clusters at this 2nd level of abstraction.

### **3.4 Application of Distance-Quantification-Classing (DQC) Technique**

After obtaining the hierarchical clusters and extracting the numerical values, the DQC approach is applied to each cluster for the mapping of nominal data into numeric data for effective visualization. The purpose of this application is that in real-world data sets there exists a number of nominal variables which have high cardinality. For example, country names and product codes are typical examples of high cardinality nominal variables. The DQC approach maps the nominal values in to numeric values for effective visualization. Additionally, the approach assigns order and spacing among the variable values in a manner that conveys relationships and associations in the data items. For instance, the DQC approach can group the product codes or



country names that are closer to each other based on the other variables in the underlying data set. This assignment of order and spacing is done in such a way that the distance between the two values in the nominal space is preserved in the numeric space.

### 3.5 Extraction of Dimensions and Hierarchies

Following step 4, the mapped nominal to numeric values are extracted for each cluster for both levels of abstraction. These values are responsible for defining the groups in each dimension and the dimensional hierarchy. These extracted values are fed into the automatic schema generator to model the dimensional hierarchy in the data warehouse schema.

### 3.6 Data Warehouse Schema Generation

In step 6, the extracted values from the previous steps, become the input to the automatic schema generator. Automatic schema generator first reads the input dimensions and measures. Secondly, handles the dimensional hierarchies. Schema generator module identify the natural groupings of the values within each dimension and name the group i.e. Group 1, Group 2 to Group N. Each of the groups created by the schema generator is then assigned the values which are closer to each other in the mapped numeric space (details are discussed in the case study). Thirdly, creates a fact table and manages the relationships among the fact and dimension tables. As an output, this step gives a star schema and also populates the data in the corresponding dimension and fact tables using automatically generated queries. The extended methodology repeats all the above mentioned steps of this section for the clusters at both levels of abstraction. It means that the warehouse schema builder will run repeatedly to produce schema for clusters at each level in the hierarchy. For instance, it will generate the schema for Cluster1 for Level-1 and later it will generate Cluster1 for Level 2 in the hierarchy.

### 3.7 Data Cube Construction

Finally, when the schema has been generated, the data cube is being constructed to allow various data warehouse operations such as drill-down, roll-up, slicing and dicing and pivoting. Here, the cubes are constructed for both levels of detail. The construction of data cube allows the flexibility to add/remove the dimensions and to control the granularity of the warehouse analysis.

## 4 Implementation Details

In this section, based on our implementation, we discuss details of the implementation steps of the proposed extension of the methodology. We have performed case study on a real world data set from the UCI machine learning repository, namely *Adult* dataset to validate the results of our proposed methodology.

#### 4.1 Case Study – Adult Data Set

We performed a case study on a relatively large Adult data set. The data set contain 48842 records with 9 nominal and 6 numeric variables. More details can be found on the machine learning website [28]. The distribution of the high cardinality variables present in the data set is given in Table 1. We have used the HCE tool to generate hierarchical clusters from this data set using the numeric attributes. The hierarchical clustering parameters of Euclidean distance and complete linkage were defined to guide the clustering process.

**Table 1.** High cardinality nominal variables in Adult data set

Attribute names	Categorical Values
Race	5
Relationship	6
Marital-Status	7
Work-Class	8
Occupation	14
Education	16
Country	41

At the minimum similarity value of 0.302, two clusters were produced, each having their individual hierarchy. We call it the first level of abstraction called Level-1. As per our new extension, we go one level down in the hierarchy and call it Level-2 which is the 2nd level of abstraction. We change the minimum similarity value to get more clusters at a level down. At the value of 0.351 we get three distinct clusters in the hierarchical tree. Basically, cluster 1 of Level-1 is being split into two clusters. We follow a naming convention and name the clusters at the two levels as C1, C2, C1-1 and C1-2.

Figure 2 shows the hierarchical tree of the two levels of abstraction with the cluster naming convention. In the next step, we extract the numeric values from each cluster and stored them in an Excel files. After the extraction of numeric facts, we apply DQC technique to the nominal attributes present in each cluster at each level of abstraction. Using *Xmdvtool*, we visualize the nominal variables along with the assigned grouping and ordering with the help of parallel coordinates visualization technique. Figure 3 displays the resultant of cluster C1-1 nominal data mapped to numeric. Due to lack of space in the paper we only show 4 dimensions present in the cluster C1-1. Using the parallel coordinate technique the groupings of high cardinality variables such as Occupation, Relationship, Work-Class and Education can be easily visualized. One example of such grouping can be seen in the Relationship variable depicted in Figure 3. The value set {Not-in-family, Own-child, Other-relative} has similar values as compared to the values of Husband and Wife in the same variable. The *Xmdvtool* uses mapped numeric values for the nominal variables.

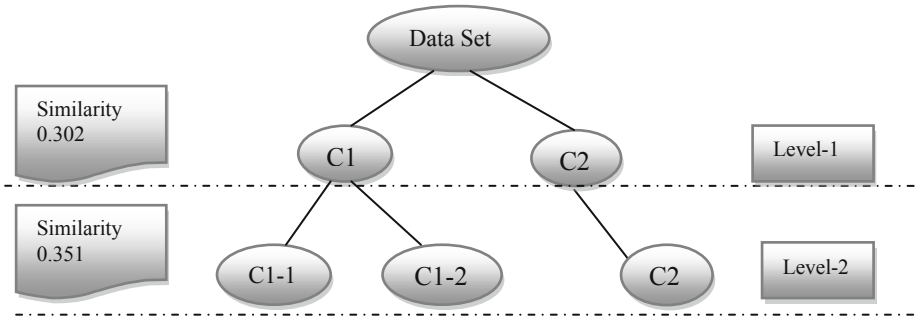


Fig. 2. Hierarchical tree representation of Adult dataset at two levels of abstraction

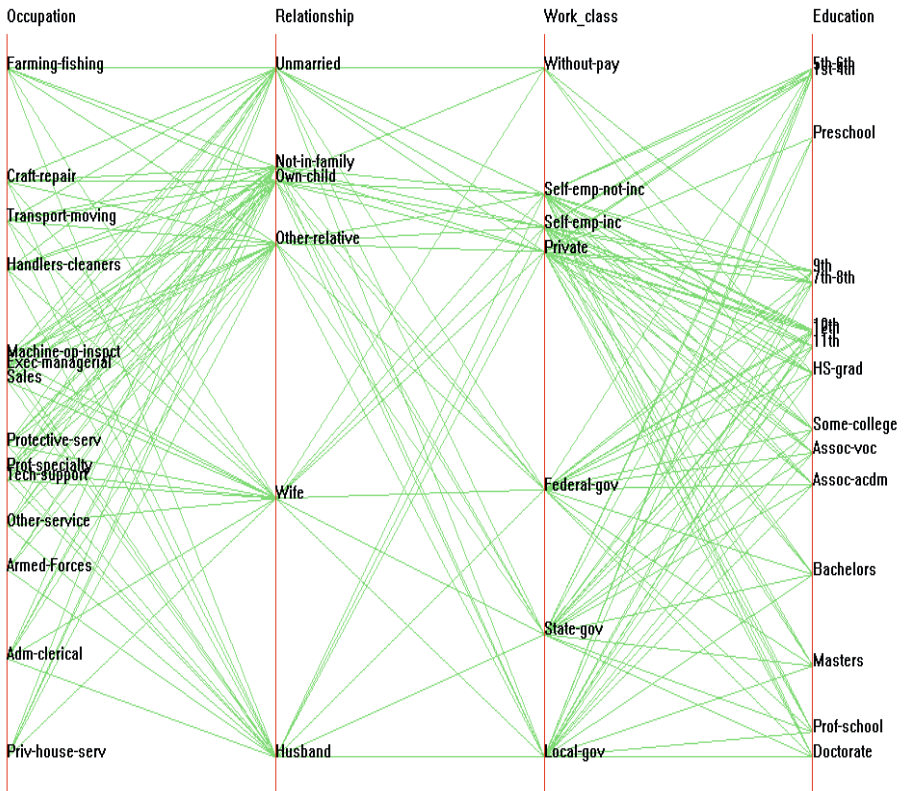


Fig. 3. Parallel Coordinate display of Level-2 Cluster named C1-1

These values numeric values along with the attribute names are exported in XML format. This exported XML file can be utilized for the purpose of identifying possible dimensions and dimensional grouping. The names within each variable which are close to each other are actually the ones who have very close numeric values. This

exported XML file along with the Excel file containing the numeric facts of a cluster at a specific level become the input for the automatic schema builder to generate data warehouse schema. The schema generator reads each file and separates the dimensions, groupings and measures present at a specific level of the hierarchical tree. In addition to this, automated set of queries create fact and dimension tables and later populate the data in the relevant tables using the SQL commands. Figure 4 shows the design of such an automatically generated schema for the Cluster C1-2 at Level-2 at the hierarchy. It can be seen that the automatically generated schema for C1-2 has six dimensions. The central fact table which consists of the measures contains *Age*, *Fnl\_Weight*, *Edu\_Num*, *Capital\_Gain*, *Capital\_Loss* and *Hrs\_Per\_Week*. From these dimensions and measures, in the final step of our proposed methodology, we construct the data cube for each Level in the hierarchical tree.

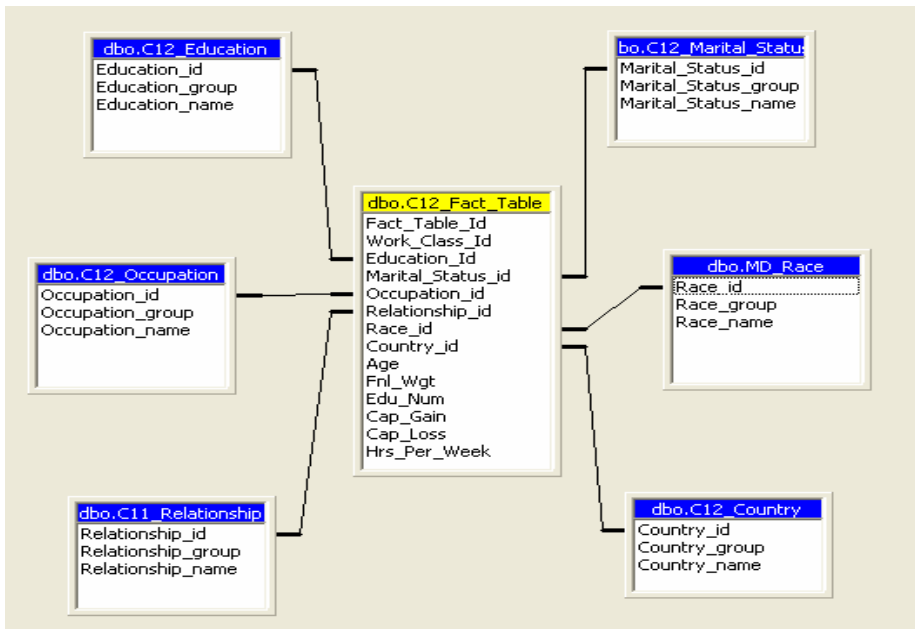


Fig. 4. Schema generated for Cluster C1-2 at Level-2 of the hierarchical tree

## 5 Results and Discussion

This section is divided into four subdivisions. We discuss the results with respect to each of our objectives that we aimed to achieve in our proposed methodology. One of the important aspects of the proposed methodology is the efficient analysis of mixed numeric and nominal data. Most of the clustering algorithms work well on the numeric data only but fail to produce meaningful clusters when mixed data is provided. In order to achieve efficient analysis of mixed data we applied a series of

different methods. We first cluster the data set with respect to the numeric attributes and obtain hierarchical clusters. On the basis of numeric attribute clustering, the HCE tool provided us 2 clusters at the top level for *Adult* data set. The resulting clusters were based on the distance measure and which works well for numeric data.

**Table 2.** Variable and grouping comparison of Cluster C1 and C1-2

Cluster C1						Cluster C1-2	
Education						Education	
Group 1	Group 2	Group 3	Group 4	Group 5	Group-Others	Group 1	Group-Others
5th-6th	9th	10th	Some-college	Prof-school	Preschool	Prof-school	HS-grad
1st-4th	7th-8th	12th	Assoc-voc	Doctorate	Bachelors	Assoc-voc	9th
		11th	Assoc-acdm		Masters	Doctorate	10th
		HS-grad				Masters	
						Occupation	
Occupation						Group1	Group-Others
Group1	Group2	Group3	Group-Others			Prof-specialty	Craft-repair
Craft-repair	Machine-op-inspct	Protective-serv	Farming-fishing			Exec-managerial	Handlers-cleaners
Transport-moving	Exec-managerial	Prof-specialty	Handlers-cleaners			Tech-support	
	Sales	Tech-support	Adm-clerical			Sales	
		Other-service	Priv-house-serv			Adm-clerical	
		Armed-Forces				Protective-serv	
						Other-service	
						Transport-moving	
						Machine-op-inspct	

For the sake of discussion, we compare groupings among selected high cardinality nominal variables contained by the two clusters at two different levels. C1 and C1-2 are the two clusters produced after performing hierarchical clustering at 2 different levels of abstraction. We compare four variables namely *Education*, *Work\_Class*, *Occupation*, and *Relationship* of Level1 Cluster C1 with Level2 Cluster C1-2. Table 2 highlights the differences among the Education and Occupation variables and their groupings. It has been observed that the number of groups for a given variable in Cluster C1 and the values within each group are different when compared with Cluster C1-2. For instance, in Cluster C1, Education has a group that contains (10th, 11th, 12th and HS-grad). In Cluster C1-2 the same variable has no group containing (12th and

11th) and furthermore 5-6th, 1st-4th and many others are totally absent in the variable values at the 2nd level of abstraction. This shows that each cluster has its own groupings or relationships based on the underlying data. In our proposed methodology, we use these variables as dimensions and the groupings within each variable as possible dimensional hierarchy levels. The Parallel Coordinates technique helps the user to visualize these grouping among the nominal values which improves the effectiveness of the visual display. Furthermore, the numeric facts and the dimensions and hierarchy levels are fed into the automatic schema generator to give a star schema as an output. As discussed in the case study, the schema generator produced a schema on using the Adult data Cluster C1-2 with 6 dimension and 6 measures. All the numeric values extracted from each cluster become potential measures and the high cardinality nominal variables become the dimensions. With these dimensions and measures a multi-dimensional data cube is generated. We show a few results with the help of some OLAP operations that how the data cube in Cluster C1 and C1-2 can facilitate warehouse analysis for the end-user. Suppose the analyst is interested to see the average number of hours per week at two different level clusters with respect to different dimensions present in each cluster. Analyst can perform this analysis using the front end of the developed prototype for OLAP exploration. First, the desired cube named C1\_Cube is selected and *Avg\_Hrs\_per\_week* is being selected as a measure. Second, the dimensions can be selected to see the *Avg\_Hrs\_per\_week* measure. Using the comparative analysis on the same dimension and measure at two different levels, analyst can determine the information present at each level. In Figure 5, we present the Cluster C1 result of the lowest number of average hours per week with 2 dimensions namely Occupation and Education.

Similarly, the analyst perform OLAP operation on Level-2 cube of the hierarchy named C12\_Cube to find the same answer of lowest number of hours per week.

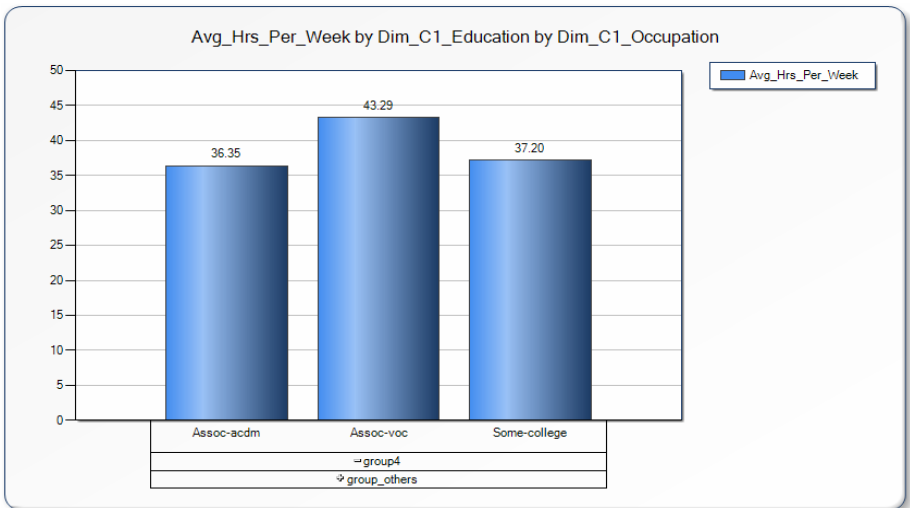


Fig. 5. Average hours per week by occupation and education dimension for Cluster C1

Figure 6 shows the result of such an OLAP query. Using the OLAP tool it can be easily identified by the analyst that in Cluster C1 the lowest average hours per week value belongs to *group\_others* which contains four types of occupation values namely *Adm-clerical*, *Farmer-fishing*, *Handler-cleaner* and *Pri-house-service* (also shown in Table 2). Furthermore, the education group which has the lowest value is also *group\_others* of education and the people whose education level is *Assoc-acdm* works the lowest in this cluster. It helps in determining a simple rule that for all occupation value equals to {*Adm-clerical OR Farmer-fishing OR Handler-cleaner OR Pri-house-service* } AND education value is *Assoc-acdm* the average hours per week are less than 36. In the same way, Cluster C1-2 analysis can help in determining another rule that if occupation is {*craft-repair OR handler-cleaner*} AND education is 9<sup>th</sup> it indicates the average hour per week value will not be more than 37.

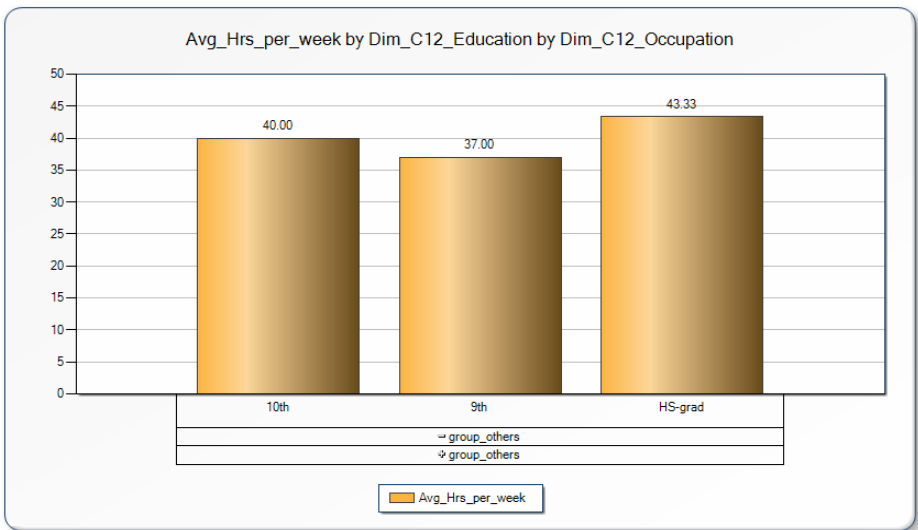


Fig. 6. Average hours per week by occupation and education dimension for Cluster C1-2

## 6 Conclusion and Future Work

In this paper, we proposed an extension to our previously reported methodology for the integration of data mining with data warehousing. The focus of this work was to extract knowledge at different levels of abstraction from a warehouse schema. We used hierarchical clustering and automated the warehouse schema generation process. Additionally, we demonstrated that efficient analysis and effective visualization of mixed nominal and numeric data at different levels of abstraction. We validated the methodology by performing case study on a real world dataset. We demonstrated that single level data analysis is insufficient to gain an adequate understanding the complexity of data. In our extended methodology, we present to the analyst information present at different levels in the hierarchical tree. In addition to this, we are construct cube for different levels of data abstraction using a semi-automatic

technique. With the presence of data cubes at different levels, the analyst can explore the similarities and differences among the different attributes present in the data clusters at various levels of detail. We demonstrated with examples that how two level analyses can help in forming simple rules for each cluster at a particular level. We believe that without this two level exploration of cube data it is not easy to understand the underlying structure of the data and lack of understanding can lead to weak analysis and poor decision making. Results show that with the extension of the two level analysis capability of the analyst can perform better mining and decision making. Further work will mainly focus on adding more sophisticated visualization techniques for the exploration of clustered data. In addition to this, we are working a more efficient implementation that will make it scale better for large and complex data sets.

## References

1. Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering* 14(4), 673–690 (2002)
2. Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering* 63(2), 503–527 (2007)
3. Rosario, G.E., Rundensteiner, E.A., Brown, D.C., et al.: Mapping nominal values to numbers for effective visualization. *Information Visualization* 3(2), 80–95 (2004)
4. Ankerst, M., Berchtold, S., Keim, D.A.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: *Proceedings of the IEEE Symposium on Information Visualization(InfoVis)*, p. 52 (1998)
5. Fua, Y.H., Ward, M.O., Rundensteiner, E.A.: Hierarchical parallel coordinates for exploration of large datasets, pp. 43–50
6. Chen, J.X., Wang, S.: Data visualization: parallel coordinates and dimension reduction. *Computing in Science & Engineering* 3(5), 110–112 (2001)
7. Artero, A.O., de Oliveira, M.C.F., Levkowitz, H.: Enhanced high dimensional data visualization through dimension reduction and attribute arrangement, pp. 707–712
8. Dori, D., Feldman, R., Sturm, A.: From conceptual models to schemata: An object-process-based data warehouse construction method. *Information Systems* 33(6), 567–593 (2008)
9. Kohavi. R., Becker. B.: UCI repository of machine learning databases, (January 20, 2011), <http://archive.ics.uci.edu/ml/datasets/Adult>, <http://archive.ics.uci.edu/ml/datasets/Adult>
10. Seo, J., Bakay, M., Zhao, P., et al.: Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis, pp. 461–464
11. Ward, M.O.: Xmdvtool: Integrating multiple methods for visualizing multivariate data, pp. 326–333
12. Soni, S., Kurtz, W.: Analysis Services: optimizing cube performance using Microsoft SQL server 2000 Analysis Services. *Microsoft SQL Server 2000 Technical Articles* (2001)
13. Milenova, B.L., Campos, M.M.: O-cluster: scalable clustering of large high dimensional data sets, pp. 290–297
14. Milenova, B.L., Campos, M.M.: Clustering large databases with numeric and nominal values using orthogonal projections
15. Doring, C., Borgelt, C., Kruse, R.: Fuzzy clustering of quantitative and qualitative data, pp. 84–89



16. Luo, H., Kong, F., Li, Y.: Clustering mixed data based on evidence accumulation. *Advanced Data Mining and Applications* 4093, 348–355 (2006)
17. McCane, B., Albert, M.: Distance functions for categorical and mixed variables. *Pattern Recognition Letters* 29(7), 986–993 (2008)
18. Hsu, C.C., Chen, C.L., Su, Y.W.: Hierarchical clustering of mixed data based on distance hierarchy. *Information Sciences* 177(20), 4474–4492 (2007)
19. Artero, A.O., de Oliveira, M.C.F., Levkowitz, H.: Uncovering clusters in crowded parallel coordinates visualizations. In: *Proceedings of the IEEE Symposium on Information Visualization(InfoVis)*, pp. 81–88 (2004)
20. Pardillo, J., Mazón, J.N.: Designing OLAP schemata for data warehouses from conceptual models with MDA. *Decision Support Systems* (2010)
21. Palopoli, L., Pontieri, L., Terracina, G., et al.: A novel three-level architecture for large data warehouses\* 1. *Journal of Systems Architecture* 47(11), 937–958 (2002)
22. Song, I.Y., Khare, R., An, Y., et al.: Samstar: An automatic tool for generating star schemas from an entity-relationship diagram, pp. 522–523
23. Usman, M., Asghar, S., Fong, S.: A Conceptual Model for Combining Enhanced OLAP and Data Mining Systems. In: *2009 Fifth International Joint Conference on INC, IMS and IDC*, pp. 1958–1963 (2009)
24. Usman, M., Asghar, S., Fong, S.: Integrated Performance and Visualization Enhancement of OLAP Using Growing Self Organizing Neural Networks. *Journal of Advances in Information Technology* 1(1), 26–37 (2010)
25. Asghar, S., Alahakoon, D., Hsu, A.: Enhancing OLAP functionality using self-organizing neural networks. *Neural, Parallel & Scientific Computations* 12(1), 1–20 (2004)
26. Goil, S., Choudhary, A.: PARSIMONY: An infrastructure for parallel multidimensional analysis and data mining. *Journal of parallel and distributed computing* 61(3), 285–321 (2001)
27. Usman, M., Pears, R.: A methodology for integrating and exploiting data mining techniques in the design of data warehouses. In: *Proceedings of ICMIA2010 2nd International Conference on Data Mining and Intelligent Information Technology Applications*, Seoul (November 2010)
28. Kohavi, R., Becker, B.: Adult dataset (1996),  
<http://archive.ics.uci.edu/ml/datasets/Adult>

# On Inserting Bulk Data for Linear Hash Files

Satoshi Narata and Takao Miura

HOSEI University, Dept. of Elect. & Elect. Engr.  
3-7-2 KajinoCho, Koganei, Tokyo, 184-8584 Japan

**Abstract.** In this investigation, we propose a new approach for bulk insert under linear hash organization, which is one of the known problems of dynamic hash techniques. Generally *Dynamic Hash* allows us to adjust the size of hash space dynamically according to the amount of data so that we obtain the nice time/space efficiency of the hash space. One of the typical techniques is *linear hash* where we can keep the hash space size linearly in terms of the amount of data. However, the technique doesn't always provide us with suitable properties, especially we face to severe deficiencies at bulk insert/delete operations, because the successive operations cause heavy manipulation (one extension at each operation) so that huge amount of I/O access happen.

Here in this work, we propose a new approach for bulk insert (delete) to relieve the thrashing situation and to reduce total I/O access. We discuss some experimental results and show how well the approach works.

**Keywords:** Dynamic Hash, Linear Hash, Bulk Insert, I/O Thrashing.

## 1 Motivation

Almost all of the modern computer processing assume some foundation of data management mechanism which utilize powerful and indispensable techniques directly or indirectly through database systems and other sophisticated vehicles. One of these techniques is certainly *hash approach* by which we can access every record in  $O(1)$ , though there exist fundamental deficiencies such as *data collision*, *spill-out* and *space limitation* caused by *static size* of hash spaces, as described in every textbook.

*Dynamic hash* has been proposed so far for purpose of the improvement of space efficiency. It provides us with dynamic growth of hash space according to the amount of records by which we keep *density*<sup>1</sup> constant. *Linear Hash* (LH) is one of the techniques applied widely to several applications<sup>[1]</sup>. By this technique, hash space grows *smoothly* (i.e., we will split buckets and append new ones in an one by one manner) for relieving overflow situation, and we expect good efficiency (small overhead of I/O).

On the other hand, in LH there is no feature for collective updates such as bulk inserts<sup>[4,5]</sup>. For instance, we are forced to split a bucket whenever *insert* operation happens to maintain the density close to our specified density factor

---

<sup>1</sup> *Density* means the ratio of the practical number of records to the space size.

$\sigma$ , called *load factor*. Since there is no explicit condition like load-factor against bucket-splitting, we can't improve this issue within a general framework. More serious is that bucket-splitting doesn't always relieve overflow buckets and the splitting may not improve efficiency. In fact, a new bucket is appended to the end of the space because of *linear* property. Especially we face to severe deficiencies to bulk insert/delete operations, because the successive operations cause heavy manipulation (one extension at each operation) so that huge amount of I/O access happen to the secondary storage devices, called *thrashing*.

Rafiei et al. examines the situation and proposes how to allocate LH space distribution in advance only for LH creation[5]. Yasuda et al. has proposed *Tree Hash* to recover overflow problem[4] where they don't manage hash spaces linearly but extend the space over tree-structure spaces so that some additional management mechanism has to be introduced.

In this investigation, we propose a new approach for bulk insert (delete) to relieve the thrashing situation and to reduce total I/O access. The basic idea comes from bulk extension of the space, we accept many records (for insert) at once, reorganize hash-space as well as new records and generate new *image* of the space. We show how to do that and how well the approach works. In fact, we show excellent improvement for this problem.

In section 2 we review Linear Hash and describe outstanding aspects of this technique. In section 3, we discuss a new approach for bulk insert under linear file organization. Section 4 contains some experiments, several analysis and the comparison with other approach. We conclude our investigation in section 5.

## 2 Linear Hashing

In this section, we review Linear Hash very quickly. For more detail, see the literatures[1,2].

Given a linear hash (LH) space consisting of a set of buckets, a possible domain  $C$  of keys, we assume a record  $d$ , key  $c \in C$  of  $d$  and a hash function  $h : C \rightarrow H$ , then we can identify a bucket position of  $d$  which contains  $d$  in the hash space by an address  $h(c)$ . So we can go to the position by  $h(c)$  directly (on the secondary storage). Each bucket in a primary hash space may contain several records.

Given two non-negative integers  $L$  (called a *level* of LH space) and  $p$  (called a *growth position*) where  $0 \leq p < 2^L$ , we assume a hash function  $h_L$  where  $h_L(c) \leq 2^L$  for any key  $c$  in such a way that  $h_{k+1}(c)$  is identical to  $h_k(c)$  on the low  $k$  bits for any  $k$  and  $c$  so that  $h_{k+1}(c)$  might be different from  $h_k(c)$  only at  $(k+1)$ -th bit position. A trivial example is  $h_k(c) = c \bmod 2^k$  considered  $c$  as bit sequence.

For given a key  $c$ , we can retrieve a record containing  $c$  as a key by the algorithm illustrated below:

$$\begin{aligned} a &\leftarrow h_L(c); \\ \text{if } a < p \text{ then } a &\leftarrow h_{L+1}(c); \quad (A1) \end{aligned}$$

That is, we obtain the value  $a = h_L(c)$  first, then if  $h_L(c) < p$ , we calculate  $a$  again by  $a = h_{L+1}(c)$ .

Once we get the position  $a$ , we access the bucket and examine whether this bucket and the overflows contain  $c$  or not.

To insert a record  $d$  with a key  $c$ , we apply the algorithm illustrated above to obtain  $a$ .

In this case, when we insert  $d$ , we might put it into an overflow area if  $a$ -th bucket is full. Each overflow bucket may contain several records.

After inserting  $d$ , we examine the status of the LH space whether there exist too many records or not, by checking density, for example. If the density goes beyond  $\sigma$ , we should split some bucket and add new buckets to relieve the status. In LH, we select the bucket of the growth point  $n$  and distribute all the records into the two buckets of  $p$  and of  $p + 2^L$  according to their  $h_{L+1}$  values<sup>2</sup>. Finally we increment  $p$ . Note that we should generate  $p + 2^L$ -th bucket as a new one and this means the LH space grows smoothly.

Also note that the growth point  $p$  should be  $p \leq 2^L$ . We adjust this condition by the following algorithm:

```

 $p \leftarrow p + 1;$ 
 $if\ p \geq 2^L\ then$ 
 $p \leftarrow 0$ 
 $L \leftarrow L + 1;$ 
    
```

Let us illustrate the insert operation in a figure 1 (b). Once  $n$  comes to  $2^L$ , we increment the level  $L$  and put  $p = 0$  as in 1 (c).

### 3 Bulk Insert to Linear Hash Files

#### 3.1 Thrashing at Bulk Insert

By using Linear Hash technique, a hash space grows *smoothly* (i.e., in an incremental manner) so that we can keep reasonable density and good efficiency (small overhead of I/O). On the other hand, when it is hard to estimate how many records are inserted in advance, there may happen bulk insert suddenly. For example, when there happen huge amount of newly registered students at entrance season and thus huge relevant formalities, we must have bulk inserts in very short period.

As pointed out previously, LH has the practical drawbacks that comes from the linearity property of the algorithm described [45]. One issue is *scalability*, i.e., it is hard to manage a huge amount of records by LH and many works have been made to improve the drawback [26][74]. Also we can't improve overflow status since  $n$ -th bucket is split but not the full bucket. Another problem is *thrashing*. By splitting one bucket, we can reduce only small amount of density

---

<sup>2</sup> For any key  $c'$  in the bucket, we have  $n = h_L(c')$ , and  $h_{L+1}(c') = p$  or  $h_{L+1}(c') = p + 2^L$  by definition.

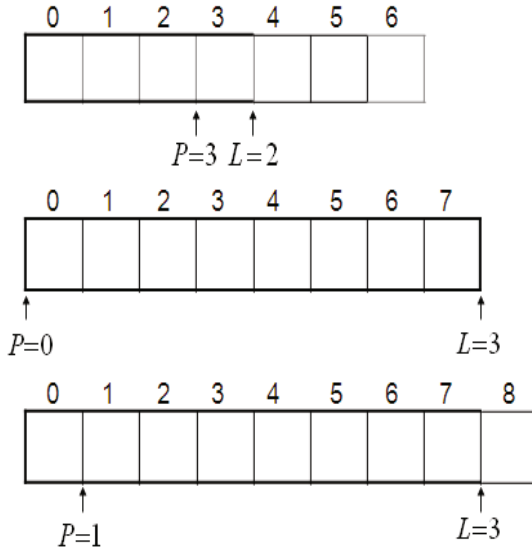


Fig. 1. Linear Hash Space

value so that there may happen many I/O to each splitting, which damages total throughput dramatically. In short, we can't avoid to have thrashing by many inserts.

For example, LH space has 1028 buckets (each bucket may contain at most one record) with the density threshold  $\sigma = 0.8$  and contains 822 records thus the density =  $822/1028 = 0.7996$ . When inserting one record, the density becomes  $823/1028 = 0.8007$  and we should have splitting. The new density is  $823/1029 = 0.7998$  and relieves the LH space. But when we insert one more records, the density becomes  $824/1029 = 0.8008$  and the splitting arises again. Generally, once we get splitting situation, we may have successive splitting very often because the splitting reduces very small amount of the density.

### 3.2 Algorithm for Bulk Inserts

In this section, we propose sophisticated algorithms for bulk inserts to LH space. To do that, we put all the records in order of hash-values in advance and then we write the bucket contents to the space at once. To simplify the discussion we assume each bucket in a primary hash space may contain only one record, though one bucket may contain several records in practice.

**Creating LH Space.** First of all, we assume there are huge amount of records to be inserted and we like to build LH space from scratch. Since we have all the initial data ( $N$  records) at hand, we can determine LH space containing all the records. Assuming the level  $L = \log_2 N$ , we obtain all hash values  $v_1, \dots, v_N$  by the hash function  $h_L$  where  $v_1 \leq \dots \leq v_N < 2^L$  and write them into the LH

space of  $2^L$  buckets. Since each bucket contains at most one record, we assume  $v_1 < \dots < v_N < 2^i$  otherwise we send them to an overflow area. Thus we must have the density  $0.5 \leq N/2^L < 1.0$  at most. Let us illustrate the situation by an example shown in a figure 2. Given 10 records ( $N = 10$ ), we have the level  $L = \lfloor \log_2 10 \rfloor = 4$  and the growth position  $p = 0$ . Assume we have the hash values 0, 2, 2, 3, 9, 9, 10, 11, 14, 15 and 8 records (0, 2, 3, 9, 10, 11, 14, 15) in the primary area of LH space of  $2^L = 16$  buckets. Note each bucket can keep single record and additional data are stored as overflow during the usual hash process.

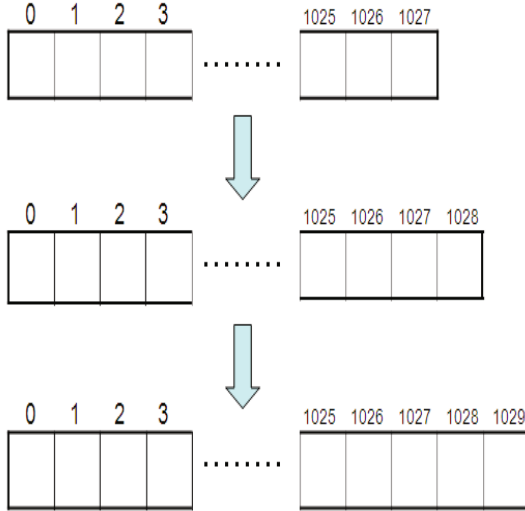


Fig. 2. LH Creation

Basically the process takes time of  $O(N)$ . The procedure can be summarized as follows:

- (1) Given  $N$  records, we estimate level  $L = \lfloor \log_2 N \rfloor$  and  $p = 0$ .
- (2) Sort the hash values of the  $N$  records by using  $h_L, v_1, \dots, v_N, 0 \leq v_j < 2^L$
- (3) Remove the duplicated values and write them to LH space of  $2^L$  buckets.
- (4) Put duplicated values to an overflow area according to each hash value and link them from the primary bucket. The new parameters are  $L, p = 0$ .

**Bulk Insert to LH Space (1).** Here we insert  $N$  records to LH space and assume  $N \geq K$  where  $K > 0$  means the total number of records contained in the primary and overflow areas of LH space. Let  $L'$  be  $\lfloor \log_2 N \rfloor$  and the new growth position  $p' = 0$ . Note the new level is  $L' + 1$  because  $2^{L'-1} < N \leq 2^{L'}$  and  $2^{L'} < N + K \leq 2^{L'+1}$ .

Similar to the previous case, we obtain the sorted sequence  $v_1, \dots, v_N$  by the hash function  $h_{L'+1}$  to the  $N$  records. Using  $K$  records in the LH space (the primary and overflow areas) of the level  $L$ , we also generate a sorted sequence  $u_1, \dots, u_K$  by the hash function  $h_{L'+1}$  to the  $K$  records. For each hash value  $0, \dots, 2^{L'+1} - 1$ , we combine the two sequence and remove the duplicated values, then write one of the values to the primary area and others to the overflow area from scratch, whatever the primary area and the overflow area contain. Note we write NULL to the both primary and overflow areas if no record is given.

Let us show an example of this case in Figure 3. Assume we have  $K=3$  records (1,6,14) in LH space and let us insert  $N=5$  records (0,2,2,6,15). Since  $L' = \lceil \log_2 5 \rceil + 1 = 4$ , we assume 5 records by  $h_4$  and 3 records in the LH space as above. The primary area of new LH space contains 6 records (0,1,2,6,14,15) as shown in figure 3. The density is  $6/16 = 0.375$ .

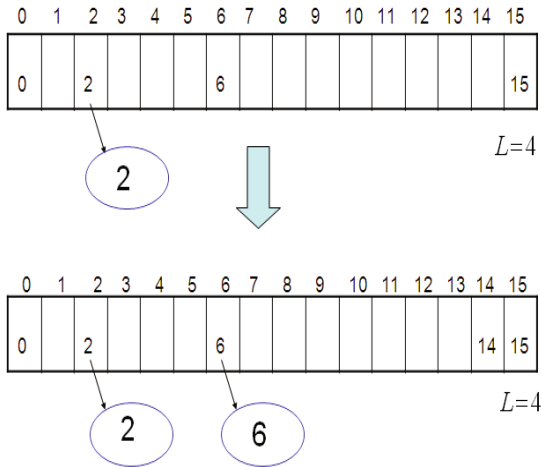


Fig. 3. Bulk Insert(1)

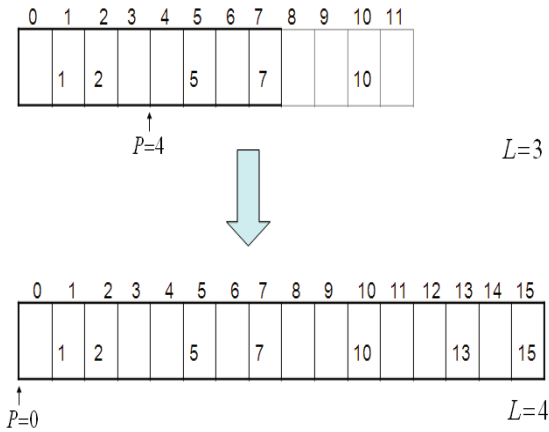
Let us show the general procedure of this case:

- (1) Given  $N$  records to be inserted and  $K$  records contained in the LH space, calculate the new level  $L' = \lceil \log_2 N \rceil + 1$  and  $p' = 0$ .
- (2) For each of  $N$  records and  $K$  records, by using using  $h_{L'}$ , we obtain the sorted sequence of hash values,  $v_1, \dots, v_N$  and  $u_1, \dots, u_K, 0 \leq v_j, u_j < 2^{L'}$ . Then remove duplicate values.
- (3) Initialize a new LH space of size  $2^{L'}$  and we write the values in (3) to the primary space.
- (4) Finally put all the duplicated values to the new overflow area and link them to the primary bucket. The new parameters are  $L', p' = 0$ .

**Bulk Insert to LH Space (2).** Here we insert  $N$  records to LH space and assume  $N \geq K$  where  $K > 0$  means the number of records contained in the LH primary space. There exist two special cases to be considered. Here we examine the special case of  $N + p \leq 2^L$ . This is special because  $N$  records may not increase LH level so that we may save space efficiency.

First we get a sorted sequence of hash values  $v_1, \dots, v_N$  of  $N$  records to be inserted to the LH space by the function  $h_{L+1}$ . Also we obtain new hash values of all the records in  $n$ -th, ...,  $(2^L - 1)$ -th buckets and the ones in each overflow bucket in the LH space by the function  $h_{L+1}$ . After removing duplicate values, we put them to the extended LH primary area and to the overflow area. Finally we put  $L + 1$  as a new level and  $p = 0$ .

Let us discuss some example of this case. We insert 2 records (5,15 in this case) to LH space of  $K = 5$  records with  $L = 3, p = 4$ . Assume the LH space contains the records of hash values 1,2,5,7,10. We calculate the new hash values to all the records (5,7 in this case) in 4,5,6,7 buckets and assume we get 7,13 for these records. Thus we will write 7 records (1,2,5,7,10,13,15) to the primary area of LH space with the new level  $L = 4$  as illustrated in figure 4.



**Fig. 4.** Bulk Insert(2)

Let us summarize the general procedure of  $N$  records insertion to LH space of  $L, p$  containing  $K$  records.

- (1) Get a sorted sequence of hash values  $v_1, \dots, v_N$  of  $N$  records to be inserted to the LH space by the function  $h_{L+1}$ .
- (2) Obtain new hash values of all the records in  $p$ -th, ...,  $(2^L - 1)$ -th buckets and the ones in each overflow bucket in the LH space by the function  $h_{L+1}$ .
- (3) Write (1) and (2) back to the primary and the overflow areas of LH space. The new parametera are  $L + 1, p = 0$



**Bulk Insert to LH Space (3).** Here we insert  $N$  records to LH space and assume  $N \geq K$  where  $K > 0$  means the number of records contained in the LH primary space, and we examine another case of  $2^{L+1} \geq N + p > 2^L$ . This is special because  $N$  records insertion causes double size of the current LH space and increase LH level so that we should take care of space efficiency.

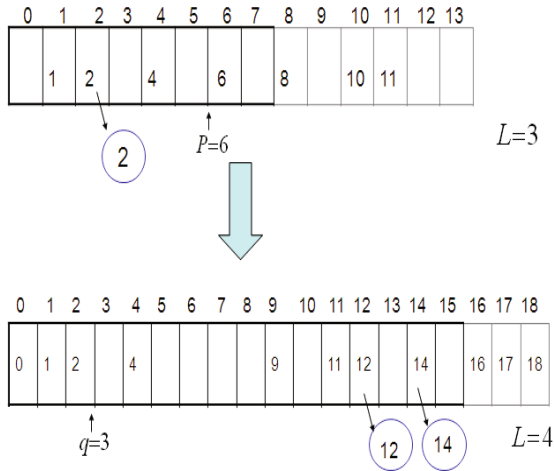
Let  $m$  be a new growth position where  $m = (N + p + 2^L) \bmod 2^{L+1}$ .

Because many records could be affected by this bulk insert, we should recalculate all the hash values of 0-th, ...,  $(m - 1)$ -th buckets by  $h_{L+2}$  and  $m$ -th, ...,  $(2^L - 1)$ -th buckets by  $h_{L+1}$  in the current LH space. We also obtain hash values  $v_1, \dots, v_N$  in sorted order by two hash functions  $h_{L+1}, h_{L+2}$  to given  $N$  records to be inserted where we obtain new hash values by  $h_{L+2}$  if hash value by  $h_{L+1} < m$ . Thus we should have  $0 \leq v_j < 2^{L+1} + m$ .

Figure 5 shows the situation where LH space contains 8 records with  $L = 3, p = 6$ . We add 5 records to the space. We have a new growth position  $m = (6 + 5 + 8) \bmod 2^3 = 3$ . Assume 8 records in the LH space have hash values 1, 2, 2, 4, 6, 8, 10, 11. Then we recalculate 1, 2, 2 by  $h_5$  and we assume to obtain new values 1, 2, 18 of the records respectively.

We hash 5 records in 3rd, 4-th, ..., 15-th buckets by  $h_4$  (new level 4) and assume we get 4, 10, 11, 14, 16 respectively.

As for  $N = 5$  records insertion, assume we get hash values 0, 1, 9, 12, 14 by  $h_4$  (level 4). We calculate again 0, 1 records by  $h_5$  since  $0, 1 < m$ , and we assume to get 0, 17 respectively. Eventually we have new LH space where the primary area contains 11 records, 0, 1, 2, 4, 9, 11, 12, 14, 16, 17, 18. The density is  $11/16 = 0.688$ .



**Fig. 5.** Bulk Insert(3)

Let us describe our general procedure where  $N$  records insertion to LH space of  $L, p$  containing  $K$  records.

- (1) Get a new growth position  $m = (N + p + 2^L) \bmod 2^{L+1}$ .
- (2) Recalculate all the hash values of 0-th, ...,  $(m - 1)$ -th buckets by  $h_{L+2}$  and  $m$ -th, ...,  $(2^L - 1)$ -th buckets by  $h_{L+1}$  in the current LH space.
- (3) Get hash values  $v_1, \dots, v_N$  in sorted order by two hash functions  $h_{L+1}, h_{L+2}$  to given  $N$  records to be inserted where we obtain new hash values by  $h_{L+2}$  if hash value by  $h_{L+1} < m$ .
- (4) Write (2) and (3) back to the primary and the overflow areas of LH space. The new parameters are  $L + 1, m$ .

## 4 Experimental Results

### 4.1 Preliminaries

Let us discuss the experimental results to see how well the proposed algorithms work.

We prepare two kinds of experimental data for our experiments. One kind consists of 5 files containing 5000, 10000, 50000 and 60000 records respectively which are generated randomly and uniformly distributed by means of MersenneTwister [3], abbreviated by "MT". Another kind consists of 5 files coming from 123593 postal addresses (points) which represent three metropolitan areas (New York, Philadelphia and Boston) in the US, abbreviated by "ZIP". Basically there are three clusters in the form of non-uniformly distributed rural areas and smaller population centers [3].

During our experiments we assume one bucket can keep single record in a primary area of LH space and 3 records in an overflow area. We examine all the frequencies of I/O requests within the primary area and the overflow area as well as LH levels, densities and load factors.

We examine several aspects of the proposed algorithms. For LH space creation (bulk insert operation from scratch), we examine practical time and space complexities by inspecting I/O counts on the primary area and the overflow area.

For bulk insert to LH space (containing some records), we also examine practical time and space complexities in the same way.

To compare bulk inserts with conventional ones, we examine several sizes of creation of LH space and bulk inserts.

First, we create LH space (bulk insertion from scratch) of 5000, 10000, 50000 and 600000 records.

Second, we add 5000, 10000 and 600000 records to LH space containing 50000 records. Also we add 50000 records to newly created LH space of 10000 records for comparison purpose.

To these LH spaces, we examine 10000 queries where half of them have no answers and take counts of bucket access.

As the baseline of our experiments, we examine conventional LH files: we add 50000 records to empty LH space, and we add 10000 records to newly created

---

<sup>3</sup> [www.rtreportal.org](http://www.rtreportal.org)

LH space (linear inserts) of 50000 records, and 50000 records to newly created LH space of 10000 records.

In this case, we examine several threshold values 0.5, 0.6, 0.7, 0.8 and 0.9 by which splitting arises<sup>4</sup>.

## 4.2 Results

Let us show the results of our experiments to both kinds of "MT" and "ZIP". We denote the number of inserted records by "Data", the number of bucket access in primary space by "PR", the number of buckets modified in primary space by "PW", the number of bucket access in overflow space by "OR", the number of buckets modified in overflow space by "OW", their summary by "TOTAL", the level of LH space by "Level", the ratio (%) of the number of records to primary hash space by "Density" and the average number of whole I/O per record by "AVG".

As for LH space creation we show all the results for bulk inserts from scratch in a table [1](#).

**Table 1.** LH space creation (Bulk Inserts)

Data	PW	OW	TOTAL	Density(%)	Level	AVG
5000						
MT	3741	1005	4746	45.67	13	0.95
ZIP	3690	1040	4730	45.04	13	0.95
10000						
MT	7401	2093	9494	45.17	14	0.95
ZIP	7171	2154	9325	43.77	14	0.93
50000						
MT	33999	12048	46047	51.88	15	0.92
ZIP	31757	12641	44398	48.46	15	0.89
60000						
MT	37975	15815	53790	57.95	16	0.90
ZIP	34567	16305	50872	52.75	16	0.85

A table [2](#) contains the results of creation with 50000 records by conventional LH technique with several load factors. Since conventional splitting requires "Write" operations many times as well as "Read", we show the both frequencies.

A table [3](#) contains the results for bulk inserts to LH space containing 50000 records in advance. Here we show cases of adding 5000, 10000 and 60000 records. We also show the case of (normal) LH in a table [4](#) where we insert 10000 records to a LH file of 50000 records with several load factors (LF).

A table [5](#) shows the results for 50000 inserts to LH space containing 10000 records in advance. We show a case of Bulk Insert and several cases of LH Inserts with several load factors.

After completing insertion, we examine the efficiency of query performance. Here we take counts of bucket access requested to 10000 queries of which 5000 queries fail (no answer).

<sup>4</sup> Note, in our bulk insert process, there is no threshold since we always split the LH space.

**Table 2.** LH space creation of 50000 records (LH Inserts)

LoadFactor	PR	PW	OR	OW	TOTAL	AVG	Level
0.5							
MT	135006	67274	78981	43073	324334	6.49	15
ZIP	134692	65015	81292	44402	325401	6.51	15
0.6							
MT	134160	66913	78848	43103	323024	6.46	15
ZIP	127581	62340	78978	44168	313067	6.26	15
0.7							
MT	122858	60862	93064	49595	326379	6.53	15
ZIP	120729	58191	97529	51341	327790	6.56	15
0.8							
MT	112682	51486	111769	57101	333038	6.66	14
ZIP	109361	48404	112345	57696	327806	6.56	14
0.9							
MT	103618	38964	140554	65136	348272	6.97	14
ZIP	102982	37565	145712	65982	352241	7.04	14

**Table 3.** Bulk Inserts to LH space of 50000 records (Bulk Inserts)

Data	PR	PW	OR	OW	TOTAL	AVG	Density	Level
5000								
MT	27634	12454	36123	9268	85479	17.10	79.93	15
ZIP	28464	12710	38870	9654	89698	17.94	78.02	15
10000								
MT	30260	14082	37205	11888	93435	9.34	82.63	15
ZIP	30672	14316	39854	12138	96980	9.70	81.02	15
60000								
MT	49910	53041	25311	39849	168111	2.80	35.25	17
ZIP	49909	53041	25311	39849	168110	2.80	35.25	17

**Table 4.** 10000 Inserts to LH space of 50000 records (LH Inserts)

LoadFactor	PR	PW	OR	OW	TOTAL	AVG	Level
0.5							
MT	28362	14341	18796	9444	70943	7.09	15
ZIP	29334	13889	21577	11020	75820	7.58	15
0.6							
MT	28408	14308	18959	9518	71193	7.12	15
ZIP	23359	10925	17522	9916	61722	6.17	15
0.7							
MT	21303	10137	15546	9221	56207	5.62	15
ZIP	18852	8618	14454	9303	51227	5.12	15
0.8							
MT	26505	12415	29251	12736	80907	8.09	15
ZIP	29654	13424	37021	14562	94661	9.47	15
0.9							
MT	17440	5958	22351	11668	57417	5.74	14
ZIP	16889	5424	22609	11658	56580	5.66	14

First, in a table 6, let us illustrate the number of access buckets for query just after bulk inserts where "Access" means the the total number of bucket access.

In two tables 7 and 8, we show the access counts under 2 different situations respectively, insertions of 10000 records to 50000 file and of 50000 records to 10000 files.

**Table 5.** 50000 Inserts to LH space of 10000 records

Insert	PR	PW	OR	OW	TOTAL	Avg	Level	Density
Bulk Insert								
MT	13473	43055	5035	16270	77833	1.56	17	30.54
ZIP	12274	31711	6044	20184	70213	1.40	17	22.26
LH, LF=0.5								
MT	136303	67566	82359	44179	330407	6.61	15	
ZIP	130698	60520	85243	47863	324324	6.49	15	
LH, LF=0.6								
MT	134653	66817	81828	44163	327461	6.55	15	
ZIP	113007	54125	77314	45856	290302	5.81	15	
LH, LF=0.7								
MT	117776	57402	89170	48859	313207	6.26	15	
ZIP	110765	52167	92962	51044	306938	6.14	15	
LH, LF=0.8								
MT	116703	52967	120895	59086	349651	6.99	15	
ZIP	116293	50704	131155	61405	359557	7.19	15	
LH, LF=0.9								
MT	97659	35516	131007	63004	327186	6.54	14	
ZIP	96534	33394	139464	63858	333250	6.67	14	

### 4.3 Discussion

First of all let us note that, in almost all cases, we can't distinguish MT results sharply from ZIP results so that we say LH mechanism works independent of data distribution.

**Table 6.** Query Efficiency after Creation

Data	Access
5000	Bulk
MT	11898
ZIP	11992
10000	Bulk
MT	11425
ZIP	11546
50000	Bulk
MT	11354
ZIP	11805
60000	Bulk
MT	11704
ZIP	12160
50000	LH, LF=0.5
MT	13793
ZIP	13971
50000	LH, LF=0.6
MT	13803
ZIP	14248
50000	LH, LF=0.7
MT	15068
ZIP	15234
50000	LH, LF=0.8
MT	16887
ZIP	17236
50000	LH, LF=0.9
MT	19611
ZIP	20061

**Table 7.** Query Efficiency after 10000 addition

Data	Access
5000	Bulk
MT	16230
ZIP	15891
10000	Bulk
MT	16624
ZIP	16215
60000	Bulk
MT	17686
ZIP	17686
10000	LH, LF=0.5
MT	13933
ZIP	14272
10000	LH, LF=0.6
MT	13986
ZIP	14645
10000	LH, LF=0.7
MT	15274
ZIP	15646
10000	LH, LF=0.8
MT	16602
ZIP	16636
10000	LH, LF=0.9
MT	20063
ZIP	20559

**Table 8.** Query Efficiency after 50000 addition

Data	Access
Bulk	
MT	14140
ZIP	15358
LH	LF=0.5
MT	14053
ZIP	13971
LH	LF=0.6
MT	14160
ZIP	15241
LH	LF=0.7
MT	15434
ZIP	15847
LH	LF=0.8
MT	16754
ZIP	16468
LH	LF=0.9
MT	20361
ZIP	20632

As for LH creation (tables 1 and 2), Bulk Insert works poor from the view point of space utilization, say 50 % in 50000 case. However it works quite fast, in fact, we need 0.9 access per record in 50000 case, while LH Inserts (50000 case, load factor 0.5) takes 7 times slower (6.49 in MT and 6.51 in ZIP).

When we insert records to large file, we can't say Bulk Insert approach is much superior. In our case of 10000 insert to LH file of 50000 records in tables 3 and 4, the average access per record of Bulk Insert goes down to 9.34 (MT) and 9.70 (ZIP) which is 1.3 times slower compared to LH cases (load factor 0.5) of 7.09 (MT) and 7.58 (ZIP).

On the other hand, LH creation of 60000 by Bulk Insert (in a table 1) takes the access counts of 53790 (MT) and 50872 (ZIP) which are 1.7 times and 1.9 times faster respectively compared to the addition by Bulk Insert, 93435 (MT) and 96980 (ZIP). In other words, we'd better insert 60000 file from scratch.

When we insert 50000 records to LH file of 10000 records as in a table 5, we apply two steps of creation and addition. But we put the two steps into one in an efficient manner. Our experiment says about 77833 access in MT case while we need both of 46047 (creation) and 93435 (insert) access in the separate steps, 1.79 times slower. But the space utilization goes worse to 30.54 % while 2 steps approach takes more than 80 %. Note we have the level 17. LH Inserts (load factor 0.5) takes more than 4.5 times slower while the levels keep 15 in both cases. Compared to LH creation of 60000 records by Bulk Insert (53790 access in MT), this addition by Bulk Insert take 1.45 times slower (77833 access). In other words, the more records we insert, the less difference we see between the creation by Bulk Insert and the addition by Bulk Insert.

Finally, let us discuss the query efficiency. After LH creation by Bulk Insert, we examine our queries and we need 11735 access (average) with small variance independent of MT/ZIP and number of records, as shown in a table 6. We get similar observation after inserting records in tables 7 and 8. However, after LH Insert of 50000 case, we need 15991 access in average for querying, but variance of more than 15% as in a table 6. In fact, we have the average 15832 (MT) and 16157 (ZIP) but 15.6% of the variance in both cases. We get similar observation after inserting records as in tables 7 and 8.

Our discussion of queries shows that Bulk Insert provides us with uniformly distributed overflow by which we may keep small amount of variance for query access.

## 5 Conclusion

By examining our experimental results, we can say that our Bulk Insert algorithms provide us with quite fast creation and quite fast insertion of large records to small LH files. But it doesn't seem suitable to insert small amount of records. From the view point of space utilization, we sometimes get in trouble with space efficiency. This means, if there are enough space to keep records, Bulk Insert approach is quite efficient. Certainly we can avoid thrashing in this case. To make full use of our approach, we recommend Bulk Insert algorithms to create LH

space and to insert large amount of records to small LH files, and conventional LH approach otherwise.

## References

1. Litwin, W.: Linear hashing - A New Tool for File and Table Addressing. In: VLDB (1980)
2. Litwin, W., Neimat, M.-A., Schneider, D.: LH\* - Linear Hashing for Distributed Files. ACM SIGMOD, 327–336 (1993)
3. Matsumoto, M., Nishimura, T.: Mersenne twister – a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulations 8(1), 3–30 (1998)
4. Yasuda, K., Miura, T.: Distributed Processes on Tree Hash, In: COMPSAC (2006)
5. Rafei, D., Hu, C.: Bulk Loading a Linear Hash File, DaWaK, pp.23-32 (2006)
6. Rathi, A., Lu, H., Hedrick, G.E.: Performance Comparison of Extendible Hashing and Linear Hashing Techniques. In: ACM SIGSMALL/PC Symposium on Small Systems (1990)
7. Severance, C., Paramanik, S., Wolberg, P.: Distributed Linear Hashing and Parallel Projection in Main Memory Databases, In: VLDB (1990)
8. Shaffer, C.A.: Practical Introduction to Data Structures and Algorithms (Java Edition). Prentice Hall, Englewood Cliffs (1997)

# Cloud Data Storage with Group Collaboration Supports

Jyh-Shyan Lin

Department of Information Management, Yuanpei University,  
No.306, Yuanpei St., HsinChu, Taiwan, R.O.C.  
jslin@mail.ypu.edu.tw

**Abstract.** Cloud computing is now an important development trend in information technology. With virtualization technologies and centralized controls of equipments and computer facilities, cloud computing allows users to obtain required services speedier than ever and easily to expand the services they need, with cheaper software acquisition and hardware maintenance costs. Moreover, cloud computing offers more flexible and convenient access to data and services. Cloud computing also allows people to create potential brand-new applications, such as automatic data backup and cross-regional group collaborations. Many researches on cloud computing have been proposed in the literature. However, how these methods could be used for group collaboration is still unclear. In this paper we develop a cloud data storage scheme which supports group collaborations.

**Keywords:** Cloud computing, Cloud data storage, Information security.

## 1 Introduction

In recent years, mobile communication and mobile computing devices become more and more popular. The computing power of lightweight netbooks and smart phones is growing stronger. Many electronic devices, such as medical diagnosis devices, healthcare instruments, electrical facilities, automobile equipments, and home appliances are developing network interconnectivity. At the same time, network communication environment is also improved day by day. Broadband networks and wireless communication networks become more and more speedy and popular. All of these facts, in addition to the demand of people who desire flexible, convenient, geographical location independent and low-cost access to data and services, bring forth the era of cloud computing. In cloud computing, data and application software are moved from traditional local hosts to remote data centers and application servers, providing on-demand service, heterogeneous and ubiquitous network access, location independent resource pooling, rapid resource elasticity, and usage-based pricing [15]. With virtualization technologies and centralized controls of equipments and computer facilities, cloud computing allows users to obtain required services speedier than ever and easily to expand the services they need, with cheaper software acquisition and hardware maintenance costs. Moreover, cloud computing offers more flexible and convenient access to data and services. Users can use various client devices, no matter



desktop PCs or lightweight thin client devices such as netbooks and smart phones, to subscribe services from cloud server providers with relatively cheaper software and hardware costs, and relief from the complexity of direct hardware maintenance. Furthermore, cloud computing also allows people to create potential brand-new applications, such as automatic data backup and cross-regional group collaborations.

Although cloud computing possesses so many advantages as stated above, security is a seriously concerned issue. There must be some ways to convince users that the data stored in the cloud are intact, confidential, and retrievable. There are many researches that address the issues of integrity and retrievability of cloud computing, for example [1], [2], [3], [5], and [6]. However, how these methods could be used for group collaboration is still unclear. We claim that group collaboration is an important application in cloud computing. On one hand, the client-server essence of cloud computing makes it suitable to support group collaboration. On the other hand, the requirement of group collaboration may be the main reason for an organization or a corporation deciding to subscribe a cloud service, especially for transnational enterprises. In this paper, we propose a cloud data storage scheme which supports group collaborations. This is the main contribution of this paper.

## 2 Bilinear Pairings

Let  $G_1$  and  $G_2$  are two additive group,  $G_T$  is a multiplicative group, a bilinear pairing is a function

$$e : G_1 \times G_2 \rightarrow G_T$$

satisfying the following properties:

- **Bilinearity:** For all  $P, P_1,$  and  $P_2 \in G_1$  and  $Q, Q_1,$  and  $Q_2 \in G_2,$ 

$$e(P_1 + P_2, Q) = e(P_1, Q) e(P_2, Q),$$
and
$$e(P, Q_1 + Q_2) = e(P, Q_1) e(P, Q_2).$$
- **Non-degeneracy:** If for all  $P \in G_1, e(P, Q) = 1$  (identity of  $G_T$ ), then  $Q = \mathcal{O}$  (identity of  $G_2$ ); Likewise, if for all  $Q \in G_2, e(P, Q) = 1,$  then  $P = \mathcal{O}$  (identity of  $G_1$ ).

In practice,  $G_1$  and  $G_2$  are two subgroups of the points on an elliptic curve  $E$  over finite field  $F_q,$  notated by  $G_1, G_2 \subseteq E(\overline{F}_q),$  where  $q$  is a large prime power and  $\overline{F}_q$  is the algebraic closure of  $F_q;$   $G_T$  is equivalent to a subgroup of a finite field.  $G_1, G_2$  and  $G_T$  all have the same order. In some application, we can take  $G_1 = G_2,$  in this case, pairing is symmetric. In this paper we use the *reduced Tate pairing*

$$e_t : E(F_{q^k})[r] \times E(F_{q^k})/rE(F_{q^k}) \rightarrow \mu_r,$$

where  $k$  is the *embedding degree* of the elliptic curve  $E, E(F_{q^k})[r] = \{P \in E(F_{q^k}) \mid rP = \mathcal{O}\},$   $rE(F_{q^k}) = \{rP \mid P \in E(F_{q^k})\}, E(F_{q^k})/rE(F_{q^k})$  is a quotient group, and  $\mu_r = \{x \in \overline{F}_q \mid x^r = 1\}$  is the  $r$ -th root of unity.

## 2 Related Works

Early related works focused on peer-to-peer (P2P) network data storage problems. Lillibridge et al. proposed a scheme for P2P data backup by using  $(m+k, m)$ -erase codes to distribute file blocks to  $m + k$  peer hosts [13]. Their method can successfully detect data losses, but cannot ensure that all data are unchanged. Filho et al. used RSA-based hash functions to verify data integrity, achieving undeceivable data authentication in P2P networks [10]. However, the time complexity of their method is exponential to data size and therefore unpractical.

Ateniese et al. introduced the model of *provable data possession* (PDP) [1]. The main intention of PDP is to confirm the accuracy of data stored in untrusted storage servers. Their model uses public-key encryptions for data validation, thus allowing public verification. But their method requires a lot of calculations. It could be a very heavy burden for the encoding and verification of a large file. In the following research, Ateniese et al. used traditional symmetric key encryptions to construct their PDPs [3], providing more efficiency than the previous scheme, and supporting dynamic data file block appending and modification. However, this scheme can only work on a single storage server, and cannot deal with small amounts of data corruption. Curtmola et al. extended the PDP model to multiple data replicas across distributed storage systems [7]. Their scheme can ensure the integrity of data without encoding each replica separately.

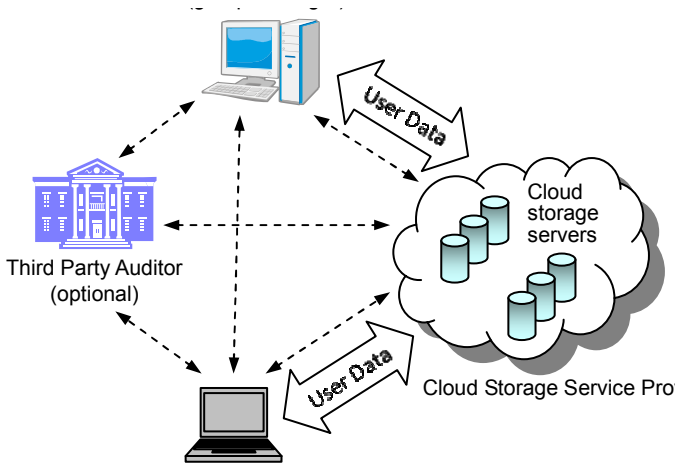
Juels et al. introduced the *proof of retrievability* (PoR) model to ensure the integrity of remote data [12]. Their scheme utilized error-correcting codes and pseudo-random dispersed checking blocks to ensure both possession and retrievability of data. However, a checking block is turned into invalid after it is used for verification, thus restricting the number of verifications. Shacham et al. extended this PoR model with a random linear function, called *homomorphism authenticator*, and presented a PoR scheme without limitation on the number of verifications [16]. Bowers et al. generalized Juels's model and Shacham's model and presented an improved PoR scheme [5]. Latter, Bowers et al. expanded their scheme to a distributed architecture [6]. Dodis et al. linked PoR with the well studied topic *hardness amplification* in complexity theory and defined a purely information-theoretic notion of PoR codes [8]. Some improved PoR codes were also introduced.

Using erasure correction codes and homomorphism symbols, Wang et al. proposed a decentralized scheme which supports dynamic data modification, deletion, and appending (but no data insertion supported) [18]. Data errors can be detected and trying to find the location of the error blocks, recovering errors efficiently. However, their scheme stores the homomorphism symbols in local sites for data validation, thus cannot support public verification. In their follow-up study, Wang et al. utilized the *Merkle hash tree* (MHT) data structure to improve the PoR model, supporting both public verification and fully dynamic data update [17]. Around the same time, Erway et al. extended the PDP model to a fully dynamically updatable scheme [9]. They utilized the *skip list* data structure to support dynamic data updates, in particular for data insertions. Wang et al. extended their scheme to a public-key encryption-based scheme which supports public verification and fully dynamic data update, and ensures no privacy leakage during public verifications [19]. To increase efficiency, this scheme also

utilized aggregate signature to combine multiple verifications into one verification. Ateniese recently proposed a general construction of public-key *homomorphic linear authenticator* (HLA) [2]. Any identification protocol can be transformed into a public-key HLA as long as the protocol satisfies appropriate conditions. Around the same time, Itani et al. presented some protocols to ensure the privacy of individual users in cloud data storage [11]. Lin described a cloud data storage model for group collaboration [14]. This is the first paper in literature that considered group collaboration in cloud data storage, but its structure is still incomplete.

## 4 Architecture

A cloud data storage architecture for group collaboration is illustrated in Figure 1. In the architecture, messages between entities are transmitted by secured channels. The architecture consists of the following entities:



**Fig. 1.** Cloud data storage architecture for group collaboration

- **User:** A user is an individual who stores data in the cloud storage system. A user has full access right to his/her data and can verify the integrity and the retrievability of the data at any time. A user may belong to one or more groups. In each group there is a group manager who is able to trace a group signature to a member. All members of a group may work on a set of data on behalf of the group.
- **Cloud Storage Service Provider (CSSP):** A CSSP is an organization which possesses abundance of resources and expertise in order to construct and maintain a cloud data storage system.
- **Third Party Auditor (TPA):** A TPA is an agency authorized by users or groups to verify the integrity and the retrievability of their data. In the cloud data storage architecture, TPA is an optional entity in the architecture.

A CSSP provides a vast amount of storage space that shared by all users. The storage space is usually constructed by multiple storage servers in a distributed manner. All storage servers work simultaneously and collaboratively. In order to ensure the accuracy of stored data, appropriate redundancies may be stored in the storage servers for the usage of error correction codes or erasure correction codes to prevent data loss due to accident or deliberate destruction. Users can access their private data through the interfaces provided by the CSSP and manipulate the data with appending, insertion, modification, and deletion operations according to their will. The redundancies in the storage servers must be adjusted immediately corresponding to the changes made by these operations. In order to let users feel relieved to store their data on the cloud storage system, there must be some ways to convince users that the data stored in the cloud are intact, confidential, and retrievable. For this purpose, we need an efficient method by which users can verify their data, and the verification will cause little computational load to the storage servers. Furthermore, the amount of the message transmitted between the users and the CSSP for the verification is as small as possible. When data are shared by a group of users, i.e. under group collaborations, each member can read, modify, and verify the shared data independently. The behavior of a group member should be concealed from outside of the group. That is, an entity outside a group cannot distinguish a modification is made by which group member, or identify two modifications are made by the same group member. However, the behavior of a group member should be traceable inside the group, i.e., group manager can disclose a modification was made by which group member.

A TPA is an institution trusted by users and has capability and expertise that users may not have. Users may not have sufficient capacity and resources (including time, computing power and network bandwidth) to verify the accuracy of the data stored in the cloud. In such a situation, a TPA can be authorized by users or groups to verify their data immediately or periodically, and report the results to the corresponding users.

## 5 Proposed Scheme

As described in the previous section, users and TPAs must have an efficient way to verify specific data stored in the cloud. In the proposed scheme, verifications are carried out by a challenge-response interaction. A user or a TPA can submit a request to the CSSP as a challenge. The CSSP then computes a value corresponding to the challenge and sends it back to the user or the TPA as a response. If the response coincides with the knowledge about the data, then it has proved that the data stored in the storage servers are intact and retrievable. The complete scheme contains the following procedures:

- **Setup ( $l$ ):** This procedure generates global parameters and a master key used in the scheme according to the security parameter  $l$ . The global parameters include a large prime power  $q$  of length  $l$ , an elliptic curve  $E$  over finite field  $F_q$ , the group order  $r$  of length  $l$ , two group  $G_1$  and  $G_2$  of order  $r$ , a generator  $g$  of  $G_2$ , a computable isomorphism  $\psi: G_2 \rightarrow G_1$ , a hash function  $H = \{0,1\}^* \rightarrow Z_r^*$ , and a cryptographic hash function  $H_2: \{0,1\}^* \rightarrow G_2$ . The master key is a random number  $msk \in Z_r^*$ .

- **GrpSetup** ( $msk, n$ ): On input the master key  $msk$  and the number of group member  $n$ , this procedure generates the group public key  $gpk$ , the group master key  $gmsk$ , the group secret key  $gsk$ , and each member's secret key  $usk_i, i = 1, 2, \dots, n$ .  $gmsk$  is given to the group manager,  $gsk$  is given to each user, and  $usk_i$  is given to designate group member.
- **Enc** ( $gsk, b$ ): Executed by a user to encrypt a data block  $b$  using group secret key  $gsk$ .
- **Dec** ( $gsk, c$ ): Executed by a user to decrypt an encrypted data block  $c$  using group secret key  $gsk$ .
- **Sign** ( $gpk, usk_i, m$ ): Executed by a group member with secret key  $usk_i$  to create a group signature  $\sigma$  for a message  $m \in G_1$ .
- **Verify** ( $gpk, \sigma, m$ ): Executed by an entity to verify a group signature  $\sigma$  for a message  $m \in G_1$ .
- **Open** ( $gpk, gmsk, \sigma, m$ ): Executed by the group manager to trace a message  $m$  from its signature  $\sigma$  to a signer.
- **SigGen** ( $usk_i, gpk, gsk, F$ ): A user utilizes this procedure to generate the verification metadata for a file  $F = \{b_i\}$ . The user first executes **Enc**( $gsk, b_i$ ) to encrypt each file block  $b_i$  to  $b_i'$ , then obtains  $\sigma_i = \mathbf{Sign}(gpk, usk_i, b_i')$  with group public key  $gpk$  and his private key  $usk_i$ , and then sends the file tag, the encrypted file  $F' = \{b_i'\}$ , the signature  $\Phi = \{\sigma_i\}$ , and a signed MHT root corresponding to the file. Note that one file needs only one execution of this procedure.
- **ReqProof**( $F', chal$ ): A user or a TPA uses this procedure to send a request to the CSSP for the verification of the file  $F'$ . The parameter  $chal$  includes index of file blocks which indicates the CSSP to generate a proof for this verification.
- **GenProof** ( $F', \Phi, chal$ ): The CSSP applies this procedure to generate a proof for the verification of a file. When the CSSP has received a request  $chal$  for the verification of  $F'$ , the CSSP utilizes  $F', \Phi$ , and the block indices included in  $chal$  to compute a proof  $P$ , then transfers  $P$  to the verifier.
- **ChkProof** ( $F', gpk, chal, P$ ): A user or a TPA uses this procedure to validate the proof generated by the CSSP for the verification of the file  $F'$ .
- **ReqUpdate**( $F', inst$ ): A group member uses this procedure to update the file  $F'$ . The parameter  $inst$  consists of an instruction, a block index, and a new block for the update and its signature (optional). The instruction is either modification, insertion, or deletion. Data appending can be accomplished by insertion.
- **ExecUpdate**( $F', \Phi, inst$ ): The CSSP utilizes this procedure to accomplish the update of the file  $F'$  and return an update proof  $P$  to the user.
- **ChkUpdate**( $F', gpk, inst, P$ ): A user uses this procedure to check whether the CSSP has updated the file  $F'$  correctly as  $inst$  designated or not. If the answer is YES, the user will send the updated signature of the MHT root to the CSSP.

We combine the methods in [4] and [17] to accomplish the Setup, GrpSetup, Sign, Verify, Open, GenSign, ReqProof, GenProof, ChkProof, ReqUpdate, ExecUpdate, and ChkProof procedures. For Enc and Dec procedures, we can use traditional cryptography primitives, such as AES or ECIES. Group data are encrypted and only members of the group can correctly decrypt and read the data. A group user or a TPA can verify the accuracy and retrievability of the data by using the public key of the group. Therefore, our scheme is possible for public verification and users can authorize a TPA to verify their data. Fully dynamic data updates, i.e. appending, insertion, modification, and deletion, are supported since our scheme is inherited from [17].

## 6 Discussions

In order to provide sufficient security (approximately the same as the standard 1024-bit RSA signature), the elliptic curve  $E$  over  $F_q$  used in this paper has embedding degree  $k = 6$  where  $q$  is approximately 170 bits long.  $G_1$  of prime order  $r$  is a subgroup of  $E(F_q)$  where  $r$  is also of length approximately 170 bits. Therefore, the discrete logarithm problem in  $G_1$  is as hard as the discrete logarithm problem in finite field  $F_{q^k}$  where  $q^k$  is of length approximately 1020 bits. For a file block, the group signature is a  $(t_1, t_2, t_3, c_1, c_2, c_3, c_4, c_5, c_6)$ -tuple where  $t_1, t_2$ , and  $t_3 \in G_1$  and  $c_1, c_2, c_3, c_4, c_5$ , and  $c_6 \in \mathbb{Z}_r$ , totally about 1530 bits, or 192 bytes. For a 4MB file with file block of size 4KB, there are totally 1025 signatures, resulting about 192KB overhead to be stored in the cloud storage servers for this file. For a 1GB file with the same block size, the overhead is about  $49152\text{KB} = 48\text{MB}$ .

**Table 1.** Comparison of our scheme with state-of-the-art under equivalent security strength

	Our scheme	[9]	[17] BLS-based	[17] RSA-based	[19]
Group collaboration support	Yes	No	No	No	No
Fully data dynamic	Yes	Yes	Yes	Yes	Yes
Public verifiable	Yes	No*	Yes	Yes	Yes
Storage overhead per file block (bytes)	192	128	22	128	22

\* No public verification mechanism mentioned.

Table 1 makes a comparison of our scheme with state-of-the-art under equivalent security strength. All these schemes support fully data dynamic, and all schemes except [9] support public verification. However, only our scheme supports group collaboration. The storage overhead per file block of our scheme is 192 bytes. The scheme in [9] utilized RSA-based encryption to generate the metadata for each file block. It needs 1024-bit, or 128 bytes, RSA product number for equivalent security, resulting 128 bytes overhead per file block. Likewise, the RSA-based scheme in [17] causes 128 bytes overhead for each block. The overhead of the BLS-based scheme in [17] and [19] is 22 bytes, assuming they use the same setting as our scheme.

In our scheme we adapt a encrypt-and-sign mechanism. However, it is better to combine these two steps into one step as signcryption for efficiency. We are currently devoting to this topic.

## 7 Conclusions

Cloud data storage has been an important application in cloud computing. Technology giants including Microsoft, Amazon, Google, IBM, Cisco, and Dell have been devoted to develop cloud data storage technologies and provide services. However, they offer few support to group collaboration. We believe that group collaboration is a huge inducement for an entity using a cloud storage service, especially for a cross-regional enterprise. In this paper we proposed a cloud data storage scheme with group collaboration supports. A group of users can operate on a set of data collaboratively and dynamically with appending, insertion, modification, and deletion operations. Every member of the group can access and verify the data independently. The verification can also be authorized to a TPA for convenience.

## Acknowledgments

We would like to thank the anonymous referees for their useful comments on this paper. This work was supported in part by the National Science Council of the Republic of China under contracts NSC 99-2221-E-264-014-.

## References

1. Ateniese, G., Burns, R., Curtmola, R., Herring, J., Kissner, L., Peterson, Z., Song, D.: Provable Data Possession at Untrusted Stores. In: Proc. of CCS 2007, pp. 598–609 (2007)
2. Ateniese, G., Kamara, S., Katz, J.: Proofs of Storage from Homomorphic Identification Protocols. In: Matsui, M. (ed.) ASIACRYPT 2009. LNCS, vol. 5912, pp. 319–333. Springer, Heidelberg (2009)
3. Ateniese, G., Pietro, R.D., Mancini, L.V., Tsudik, G.: Scalable and Efficient Provable Data Possession. In: Proc. of SecureComm 2008, pp. 1–10 (2008)
4. Boneh, D., Boyen, X., Shacham, H.: Short group signatures. In: Proc. of the Advance in Cryptology (2004)
5. Bowers, K.D., Juels, A., Oprea, A.: Proofs of Retrievability: Theory and Implementation. Cryptology ePrint Archive, <http://eprint.iacr.org/2008/175>
6. Bowers, K.D., Juels, A., Oprea, A.: HAIL: A High-Availability and Integrity Layer for Cloud Storage. Cryptology ePrint Archive, <http://eprint.iacr.org/2008/489>
7. Curtmola, R., Khan, O., Burns, R., Ateniese, G.: MR-PDP: Multiple-Replica Provable Data Possession. In: Proc. of ICDCS 2008, pp. 411–420 (2008)
8. Dodis, Y., Vadhan, S., Wichs, D.: Proofs of retrievability via hardness amplification. In: Reingold, O. (ed.) TCC 2009. LNCS, vol. 5444, pp. 109–127. Springer, Heidelberg (2009)
9. Erway, C., Kupcu, A., Papamanthou, C., Tamassia, R.: Dynamic provable data possession. Cryptology ePrint Archive, <http://eprint.iacr.org/2008/432>

10. Filho, D.L.G., Barreto, P.S.L.M.: Demonstrating Data Possession and Uncheatable Data Transfer. Cryptology ePrint Archive, <http://eprint.iacr.org/2006/150>
11. Itani, W., Kayssi, A., Chehab, A.: Privacy as a service: privacy-aware data storage and processing in cloud computing architectures. In: Proceedings of the 2009 International Conference on Dependable, Autonomic and Secure Computing (DASC 2009), pp. 711–716 (2009)
12. Juels, A., Burton, J., Kaliski, S.: PORs: Proofs of Retrievability for Large Files. In: Proc. of CCS 2007, pp. 584–597 (2007)
13. Lillibridge, M., Elnikety, S., Birrell, A., Burrows, M., Isard, M.: A Cooperative Internet Backup Scheme. In: Proc. of the 2003 USENIX Annual Technical Conference (General Track), pp. 29–41 (2003)
14. Lin, J.-S.: Cloud Data Storage for Group Collaborations. In: Proceedings of the World Congress on Engineering 2010 (WCE 2010), London, U.K, pp. 485–486 (2010)
15. Mell, P., Grance, T.: Draft nist working definition of cloud computing (2009), <http://csrc.nist.gov/groups/SNS/cloud-computing/index.html>
16. Shacham, H., Waters, B.: Compact Proofs of Retrievability. In: Asiacrypt 2008 (December 2008)
17. Wang, Q., Wang, C., Li, J., Ren, K., Lou, W.: Enabling public verifiability and data dynamics for storage security in cloud computing. In: ESORICS 2009, Saint Malo, France (September 2009)
18. Wang, C., Wang, Q., Ren, K., Lou, W.: Ensuring data storage security in cloud computing. In: Proc. of IWQoS 2009, Charleston, South Carolina, USA (2009)
19. Wang, C., Wang, Q., Ren, K., Lou, W.: Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing. In: INFOCOM 2010, pp. 525–533 (2010)



# A Negotiation Mechanism That Facilitates the Price-Timeslot-QoS Negotiation for Establishing SLAs of Cloud Service Reservation

Seokho Son and Kwang Mong Sim \*

Multiagent and Cloud Computing Systems Lab.,  
Department of Information and Communication,  
Gwangju Institute of Science and Technology (GIST), Gwangju, Korea  
{shson, kmsim}@gist.ac.kr  
<http://mas.gist.ac.kr/>

**Abstract.** A negotiation mechanism is essential to establish a service level agreement between Cloud participants who need to resolve different preferences of a Cloud service. Whereas there are some mechanisms for supporting service level agreement negotiation, there is little or no negotiation support of price, time slot, and QoS issues concurrently for a Cloud service reservation. The contribution of this work is designing a multi-issue negotiation mechanism to facilitate 1) concurrent price, time slot, and QoS negotiations between agents representing Cloud participants and 2) trade-off proposals for price, time slots, and level of QoS issues. The ideas of the negotiation mechanism are implemented in an agent-based Cloud testbed, and the empirical results obtained from simulations carried out using the testbed suggest that using the concurrent negotiation mechanism, (i) a consumer and a provider agent have a mutually satisfying agreement on price, time slot, and QoS issues in terms of the aggregated utility, and (ii) both agents achieved the highest negotiation speed among related approaches.

**Keywords:** Agent-based Cloud Computing, Cloud Business Models, Service Level Agreement, Negotiation Agents, Multi-issue Negotiation, Cloud Service Reservation, QoS Negotiation.

## 1 Introduction

For a Cloud service reservation, Cloud participants have to consider the price of the service and when to use a service (i.e., position of timeslot). Also QoS requirements of the service are important issues for Cloud participants. Both a consumer and a provider have to reach agreements on these issues for establishing the service-level agreement (SLA) when a consumer is leasing a Cloud service. However, since Cloud participants that consist of consumers and providers are individual bodies, some

---

\* Corresponding author.

mechanisms are necessary to resolve different preferences on leasing or lending services. A negotiation mechanism is one of key methods to resolve different preferences between participants who need an agreement. Whereas previous works have reported on advance reservations considering bandwidth or time constraints [1–3] and adopting a negotiation mechanism for SLA [4], as yet there is no definitive service reservation system that concurrently considers both the price and timeslot negotiation together except [5]. However, the design of [5] does not include QoS as a negotiation issue.

Since negotiation issues (i.e., price, timeslot, and QoS) are in a tradeoff relationship—a consumer who pays a great deal of money for a service can demand to use the service at a more desirable timeslot or QoS—price, timeslot, and QoS have to be negotiated not individually but concurrently. Accordingly, a multi-issue (i.e., multi-attribute) negotiation mechanism also has to be considered in this work. Even though there are several negotiation mechanisms for Grid resource negotiation (see [6] for a survey), these negotiation mechanisms are designed for Grid resource management and thus may not be appropriate for Cloud service reservation. Whereas [7] designed multi-issue SLA negotiations for Web service, these mechanisms are not specifically designed for negotiation issues in Cloud service reservation. There has been little work to date on Cloud service negotiation except for [8]; [8] proposes a complex Cloud negotiation mechanism for supporting concurrent negotiation activities in interrelated markets in which the negotiation outcomes between Cloud brokers and Cloud resource providers in one market can potentially affect and influence the negotiation outcomes of Cloud brokers and Cloud consumers in another market. The difference between this work and [8], however, is that whereas [8] focuses on a complex concurrent negotiations in multiple interrelated Cloud markets, in which the outcomes in one market can potentially influence another, this work is the earliest work to consider a Cloud service reservation system supporting a concurrent negotiation mechanism for price, timeslot, and QoS level including a tradeoff algorithm.

Finally, it should be noted that the earlier work of this paper were presented in [5]. In [5], the concurrent price and timeslot negotiation mechanism (CPTN) was designed. The design of CPTN includes a novel tradeoff algorithm, referred to as the “burst mode” proposal, designed to enhance both the negotiation speed and aggregated utility of the price and timeslot in a multi-issue negotiation. With burst mode, agents are allowed to make a concurrent set of proposals, in which each proposal consists of a different pair of price and timeslot that generate the same aggregated utility. Increasing the number of proposals in concurrent set lets agents have an enhanced negotiation speed and aggregated utility. However, in [5], the negotiation agent is not sufficient for a Cloud service reservation since there is no consideration for Cloud QoS issues in the multi-issue negotiation mechanism. Since there have been many unconsidered and unspecified Cloud QoS attributes in multi-issue negotiations, it can be expected that extending [5] by considering and specifying other negotiation issues will contribute to facilitating not only multi-issue negotiations but also Cloud service reservations.

As such, the purpose of this work is to: 1) design the Cloud service reservation system supporting concurrent Price-Timeslot-QoS negotiation (Section 2); 2) describe

the concurrent price, timeslot, and Cloud QoS negotiation agent including the design of Cloud QoS utility functions and a tradeoff algorithm (Section 3); 3) evaluate performances of the proposed negotiation mechanism in terms of the negotiation outcomes (i.e., total utility and negotiation speed) (Section 4). Finally, Section 5 concludes this paper with a list of future works.

## 2 Overview of Cloud Service Reservation System Supporting Concurrent Price-Timeslot-QoS Negotiation

This section introduces the framework of the Cloud service reservation system. The proposed Cloud service reservation system supports concurrent negotiation for price, time slot, and QoS level. The initiative of the system is that price, position of time slot, and QoS level are most important issues for a Cloud service reservation and both time slot and QoS level are closely related with price of the service. Therefore, a concurrent negotiation for those issues can be key issues in Cloud service reservation.

For consumers, the benefit of concurrent negotiation for price, time slot, and QoS level is that consumers can adjust level of satisfaction on timeliness of using service and the quality of service on their budget because the concurrent negotiation mechanism can specify tradeoff relationship among price, time slot, and QoS level. A consumer who has not sufficient budget but have sufficient time to utilize a Cloud service is willing to reserve their service on negotiated time slots with a provider. Likewise, the concurrent negotiation can increase resource utilization of providers since providers can schedule their resource utilization by controlling the price of services for each period. For example, provider can guide a consumer to reserve their job on available time slots by pricing the service at that time in a low cost. Therefore, the proposed reservation system is useful to manage their resources to providers and gives flexibility for selecting suitable services to service consumers. Even though one of the properties of Cloud is elasticity, managing resource utilization overall Clouds is important because there are limited resources in a Cloud eventually, and the reservation scheme is useful to efficiently share the Cloud resources for both consumer and provider.

Fig.1. shows the framework of the Cloud service reservation system. The Cloud service reservation system consists of consumer's site and provider's site. For each side, the main component of the system is concurrent multi-issue negotiation agent. The negotiation protocol of the concurrent negotiation agents follows the Rubinstein's Alternating Offers protocol [9], in which agents make counter offers to their opponent in alternate rounds. Both agents generate counter offers and evaluate their opponent's offers until either an agreement is made or one of the agents' deadlines is reached. Counter proposals are generated according to the negotiation strategy designed in Section 3.2, and proposals are evaluated by utility functions designed in Section 3.1. If a counter-proposal is accepted, both agents can eventually reach a mutually acceptable price, timeslot, and QoS level, and consumer can reserve the service in agreed price, timeslot, and QoS level. Conversely, if one of the agents' deadlines expires before they reach an agreement, their reservation fails.

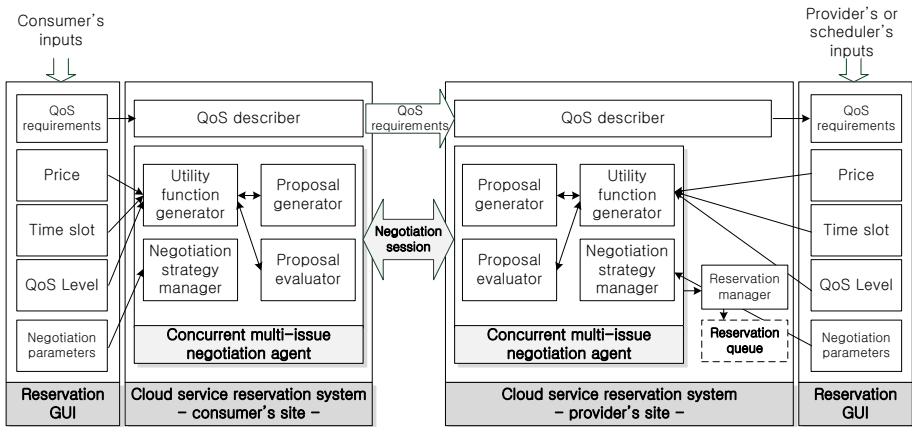


Fig. 1. The framework of the Cloud service reservation system

With Cloud service reservation system, a consumer and a provider can specify preferences on price, time slot, QoS requirements, and QoS level through the reservation GUI. A provider also specifies preferences on price and time slot; determines preference on QoS level for negotiation according to QoS requirements from the consumer and a system situation. Also, negotiation parameters such as negotiation strategy and negotiation deadline can be specified through the reservation GUI. In case of QoS requirements, it can be diverse according to its service. Therefore, it is hard to specify all QoS utility functions for negotiation. To avoid complicated negotiation, this paper classifies QoS requirements into three kinds of QoS level: 1) service performance, 2) service availability, and 3) service response time.

In addition, the reservation manager in the provider site manages reservation queue. In reservation queue, the services agreed with consumers are listed in the timeline. The number of services that can be served concurrently can be estimated by each system of Cloud providers. If there is enough resource to execute the requested service, the time slots are available time slots. Otherwise, time slots are reserved time slots. Reservation manager records indices of available time slots and reserved time slots, and it is especially used for determining time slot utility function for provider.

### 3 Concurrent Price-Timeslot-QoS Negotiation Agent

This section introduces the design of a concurrent negotiation mechanism for price, timeslot, and Cloud QoS. The proposed mechanism includes the design of utility functions and negotiation strategies.

#### 3.1 Utility Functions

**Aggregated Total Utility Function.** The utility function  $U(x)$  represents an agent's level of satisfaction of a negotiation outcome  $x$ . Since, each user has different

preferences for the price, position of the timeslot, and QoS issues, a price utility function  $U_p(P)$ , timeslot utility function  $U_T(T)$ , Cloud QoS utility functions are defined in this chapter. In [5], a price utility function  $U_p(P)$ , timeslot utility function  $U_T(T)$ , and an aggregated utility function were defined. Also, we newly define and classify Cloud QoS issues into 3 categories: 1) service performance, 2) service availability, and 3) service response time.

Let  $w_p$ ,  $w_T$ ,  $w_{Qp}$ ,  $w_{Qa}$  and  $w_{Qr}$  be the weights for the price utility, the time slot utility, the service performance utility, Service availability utility, and service response time utility respectively; the weights satisfy  $w_p + w_T + w_{Qp} + w_{Qa} + w_{Qr} = 1$ . The aggregated price, timeslot, and QoS utility  $U_{total}$  at each price, time slot, and QoS issues is given as

$$\begin{aligned}
 &U_{total}(P, T, Q_{performance}, Q_{availability}, Q_{response\ time}) \\
 &= \begin{cases} 0, & \text{if either } U_p(P)=0, U_T(T)=0, U_{Qp}(Qp)=0, U_{Qa}(Qa)=0, \text{ or } U_{Qr}(Qr)=0 \\ w_p \cdot U_p(P) + w_T \cdot U_T(T) + w_{Qp} \cdot U_{Qp}(Qp) + w_{Qa} \cdot U_{Qa}(Qa) + w_{Qr} \cdot U_{Qr}(Qr), & \text{otherwise.} \end{cases} \quad (1)
 \end{aligned}$$

By varying weights, users can place different combinations of emphases on the price, time slot, and Cloud QoS issues for negotiation. If either  $U_p(P)=0, U_T(T)=0, U_{Qp}(Qp)=0, U_{Qa}(Qa)=0, \text{ or } U_{Qr}(Qr)=0$ , the aggregated total utility is because all utilities should be within the acceptable range of each utility.

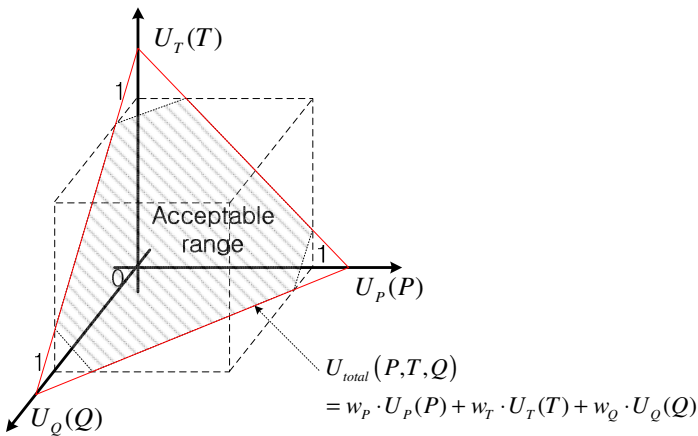


Fig. 2. Utility plane of the aggregated total utility function

Since there are 5 issues to negotiate in the proposed reservation system,  $U_{total}(P, T, Q_{performance}, Q_{availability}, Q_{response\ time})$  can be expressed in 5 dimensions. For a simple expression, the total utility function is expressed as a plain

(3 dimensions for simplicity), as shown in Fig. 2. All combinations of price, timeslot, and QoS levels in this plain give the same aggregated total utility.

**Cloud QoS Utility Function.** In this work, three Cloud QoS issues are discussed. There are many QoS issues such as performance metrics, security attributes, (transactional) integrity, response time, scalability, and availability discussed in service-oriented computing [10]. Also, QoS issues for web-service such as availability, security properties, response time, reliability, and throughput are introduced in [11] and [12]. Whereas there are various QoS issues, some issues may not be essential and negotiable for Cloud services. Cloud QoS issues discussed in this work as follow: 1) service performance, 2) service availability, and 3) service response time, since these issues are not only essential non-functional requirements but also negotiable requirements for a Cloud service.

The proposal and negotiation outcome for Cloud QoS is a satisfaction level presented as a percentage. For the purpose of simplicity, the QoS utility functions are modeled as linear and monotone functions.

**1) Service performance utility.** The service performance is defined as resource constraints or throughput (the rate at which a service can process requests) of a Cloud service in this work. The performance utility represents that the level of satisfaction about the provisioned resource constraints or throughput is guaranteed. The actual parameters for service performance can be specified by the consumer who requests a service. For example, in case of a virtual machine (VM) instance service which is an Infrastructure as a Service (IaaS), resource constraints such as CPU speed, RAM size, and disk I/O bandwidth of the VM instance can be described by a consumer. The consumer and the provider who provides the VM service can negotiate level of the guarantee for the performance resource constraints. Likewise, in case of a Service as a Service (SaaS), the service performance can be described by a throughput. Intuitively, consumers prefer the highest guarantee for the service performance, and providers want to sell their services with the lowest guarantee for the service performance with a given price. Let  $IQp_C$  and  $RQp_C$  ( $IQp_P$  and  $RQp_P$ ) be the most preferred (initial) performance quality and the least preferred (reserve) performance quality for a consumer agent (a provider agent), and let  $Qp$  be the performance quality at which a consensus is reached by both parties. For the consumer, the performance utility  $U_{Qp}^C(Qp)$  for reaching a consensus at  $Qp$  is given as

$$U_{Qp}^C(Qp) = \begin{cases} u_{min}^{Qp} + (1 - u_{min}^{Qp}) \cdot \left| \frac{Qp - RQp_C}{IQp_C - RQp_C} \right|, & RQp_C \leq Qp \leq IQp_C \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

For the provider, the performance utility  $U_{Qp}^P(Qp)$  is

$$U_{Qp}^P(Qp) = \begin{cases} u_{min}^{Qp} + (1 - u_{min}^{Qp}) \cdot \left| \frac{RQp_P - Qp}{RQp_P - IQp_P} \right|, & IQp_P \leq Qp \leq RQp_P \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

In (2) and (3),  $u_{min}^{Qp}$  is the minimum utility that the consumer and the provider receive for reaching a deal at its reserve performance. For the purpose of experimentation,  $u_{min}^{Qp}$  is defined as 0.01.

**2) Service availability utility.** Availability is expressed as a percentage of uptime of the service in reserved timeslots. Intuitively, consumers prefer the highest guarantee for availability, and providers want to sell their services with the lowest guarantee for service availability with a given price. Let  $IQa_c$  and  $RQa_c$  ( $IQa_p$  and  $RQa_p$ ) be the most preferred (initial) availability quality and the least preferred (reserve) availability quality for a consumer agent (a provider agent), and let  $Qa$  be the availability quality at which a consensus is reached by both parties. For the consumer, the availability utility  $U_{Qa}^C(P)$  for reaching a consensus at  $Qa$  is given as

$$U_{Qa}^C(Qa) = \begin{cases} u_{min}^{Qa} + (1 - u_{min}^{Qa}) \cdot \left| \frac{Qa - RQa_c}{IQa_c - RQa_c} \right|, & RQa_c \leq Qa \leq IQa_c \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For the provider, the availability utility  $U_{Qa}^P(P)$  is

$$U_{Qa}^P(Qa) = \begin{cases} u_{min}^{Qa} + (1 - u_{min}^{Qa}) \cdot \left| \frac{RQa_p - Qa}{RQa_p - IQa_p} \right|, & IQa_p \leq Qa \leq RQa_p \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

In (4) and (5),  $u_{min}^{Qa}$  is the minimum utility that the consumer and the provider receive for reaching a deal at its reserve availability. For the purpose of experimentation,  $u_{min}^{Qa}$  is defined as 0.01.

**3) Service response time utility.** response time is the time a service takes to respond to various types of request. The actual parameters for service response time can be specified by the consumer who requests a service. Intuitively, consumers prefer the highest guarantee for service response time (i.e., the fastest response time), and providers want to sell their services with the lowest guarantee for service response time (i.e., the slowest response time) with a given price. Let  $IQr_c$  and  $RQr_c$  ( $IQr_p$  and  $RQr_p$ ) be the most preferred (initial) response time and the least preferred (reserve) response time for a consumer agent (provider agent), and let  $Qr$  be the response time at which a consensus is reached by both parties. For the consumer, the response time utility  $U_{Qr}^C(Qr)$  for reaching a consensus at  $Qr$  is given as

$$U_{Qr}^C(Qr) = \begin{cases} u_{min}^{Qr} + (1 - u_{min}^{Qr}) \cdot \left| \frac{Qr - RQr_c}{IQr_c - RQr_c} \right|, & RQr_c \leq Qr \leq IQr_c \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

For the provider, the response time utility  $U_{Qr}^P(Qr)$  is

$$U_{Qr}^P(Qr) = \begin{cases} u_{min}^{Qr} + (1 - u_{min}^{Qr}) \cdot \left| \frac{RQr_P - Qr}{RQr_P - IQr_P} \right|, & IQr_P \leq Qr \leq RQr_P \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

In (6) and (7),  $u_{min}^{Qr}$  is the minimum utility that the consumer and the provider receive for reaching a deal at its reserve price. For the purpose of experimentation,  $u_{min}^{Qr}$  is defined as 0.01.

### 3.2 Negotiation Strategy

**Concession-making Algorithm.** The concession-making algorithm determines the amount of concession  $\Delta U_{total}$  for each negotiation round, and corresponds to the reduction in an agent’s expectation based on its total utility. Agents in this work adopt the time-dependent strategies in [13] to determine the amount of concession required for the next proposal and the corresponding expectation.

Let  $t$ ,  $\tau$ , and  $\lambda$  be the negotiation round, the negotiation deadline, and negotiation strategy respectively. Based on (8), the negotiation agent determines the amount of concession  $\Delta U_{total}$  and then (9) determines its expectation of the total utility in the next round.

$$\Delta U_{total} = U_{total}^t \cdot \left( \frac{t}{\tau} \right)^\lambda \quad (8)$$

$$U_{total}^{t+1} = U_{total}^t - \Delta U_{total} \quad (9)$$

**Tradeoff Algorithm.** This section introduces a general idea of “burst mode”, which is designed to enhance both the negotiation speed and the aggregated utility. Burst mode is originally designed for concurrent negotiation for price and time slot in [5]. In this work, we extend application of it for concurrent negotiation of 5 negotiation issues. In general, a multi-attribute proposal  $P$  from agent A to agent B during negotiation round  $t$  can be represented as  $P^{A \rightarrow B,t} = (x^t, y^t, \dots, k^t)$  where  $x, y, k$  are elements of negotiation attribute  $x, y, k$  respectively. Hence, a negotiation agent can make only one multi-attribute proposal in a negotiation round.

With burst mode, agents are allowed to make a concurrent set of proposals, in which each proposal consists of a different pair of price, timeslot, QoS levels that generate the same aggregated utility, but differ in terms of individual price, timeslot, and QoS utility.

A burst multi-attribute proposal  $x$  from agent A to B during negotiation round  $t$  can be represented as  $BP^{A \rightarrow B,t} = [(x_1^t, y_1^t, \dots, k_1^t), (x_2^t, y_2^t, \dots, k_2^t), \dots, (x_n^t, y_n^t, \dots, k_n^t)]$  where  $x, y, \dots, k$  are a concurrent set of proposals; these concurrent proposals are uniformly



selected from the utility line to generate a burst proposal. Since there are 5 negotiation issues in the proposed reservation system (i.e., price, timeslot, performance, availability, and response time), a burst proposal can be represented as

$$BPA \rightarrow B,t = \left[ (P_1^t, T_1^t, Q_{p1}^t, Q_{a1}^t, Q_{r1}^t), (P_2^t, T_2^t, Q_{p2}^t, Q_{a2}^t, Q_{r2}^t), \dots, (P_n^t, T_n^t, Q_{pn}^t, Q_{an}^t, Q_{rn}^t) \right] \tag{10}$$

The opponent who receives a concurrent set of proposals evaluates all sets, and then it can select the best proposal that gives the highest utility among concurrent proposals in a negotiation round. Therefore, with burst mode, a negotiating agent can provide more options for its opponent agent without having to make concession.

### 4 Simulations and Empirical Results

A series of experiments was carried out using an agent-based Cloud testbed to evaluate the performance of the negotiation mechanism in terms of the outcomes for negotiating for price, timeslot and QoS for Cloud service reservation.

#### 4.1 Performance Measure

To evaluate the performance of the proposed burst mode, we used 1) negotiation speed and 2) average total utility of the negotiation pair as the performance measures. The negotiation speed  $S$  is a function of the negotiation round  $R$  spent in the negotiation.  $S \rightarrow 0$  means the negotiation has a lower speed and  $S \rightarrow 1$  means the negotiation has a higher speed. The average total utility of the negotiation pair shows the level of satisfaction in terms of price, timeslot, and QoS with the agreed upon service. A more detailed expression of the performance measures is given in Table 1.

**Table 1.** Performance measure

Measures	Annotation
Negotiation speed (0–1)	$S = 1 - R / \text{Min}(\tau_c, \tau_p)$
Average total utility of negotiating pair (0–1)	$U_{total}^{aver.} = (U_{total}^P (P, T, Q_p, Q_a, Q_r) + U_{total}^C (P, T, Q_p, Q_a, Q_r)) / 2$

#### 4.2 Experimental Setting

Tables 2 and 3 show the input data sources for this experiment, and include the experimental settings for a Cloud market (Table 2) and user preference (Table 3) for a Cloud service. The input data sources of the Cloud market are parameters for the Cloud simulation controller. In the experiments, a Cloud market consists of 200 provider agents and 200 consumer agents to examine the performance of the reservation system in a balanced Cloud market. They are automatically generated by

the controller, and the controller randomly invokes a consumer 300 times for each simulation to start a service reservation. We note that market dynamics are not considered in this experiment.

**Table 2.** Input data source: Cloud market

Input Data	Settings
Cloud Load( $CL = N_{res} / N_{tot}$ )	$0 \leq CL \leq 1$
No. of provider agents	100 service provider agents
No. of consumer agents	200 consumer agents
Cloud services a provider lends	200 services/provider (randomly selected)
No. of negotiation sessions per each simulation	300 negotiation sessions

The Cloud load ( $0 \leq CL \leq 1$ ) in Table 2 represents and simulates different levels of utilization of the Cloud service in the Cloud environment. CL is defined here as the ratio of: 1)  $N_{res}$  —the total number of timeslots in the reservation queues of all service providers, and 2)  $N_{tot}$  —the number of timeslots already reserved. To simulate the uniformly distributed load to all service providers, each provider agent automatically fills their reservation queue with uniformly distributed virtual reservations from the simulation settings, up to a given CL.

For user preference values, settings are given in Table 3 for each consumer agent and provider agent. In this experiment, some variables (e.g., IP, RP, FT, LT, and job size) were controlled as extraneous variables to clearly observe the effects of independent variables such as CL, negotiation strategy, and negotiation deadline, because it is hard to simulate all possible combinations of input negotiation parameters due to space limitations. Every agent randomly selects the weights that satisfy  $w_p + w_r + w_{qp} + w_{qa} + w_{qr} = 1$  in the simulations to emulate diverse preferences on non-functional requirements of Cloud consumers and providers.

**Table 3.** Input data source: user inputs for service reservation

Input Data	Annotation	Settings	
		Consumer	Provider
Initial price (IP)	Integer(Cloud \$)	10–60	200–250
Reserve price (RP)	Integer(Cloud \$)	200–250	10–60
First timeslot (FT)	Integer, $FT < LT$	10–60	10–60
Last timeslot (LT)	Integer, $FT < LT$	300–350	300–350

**Table 3.** (continued)

Initial performance level (IPL)	Min=0, Max=100	0–49	0–49
Reserve performance level (RPL)	Min=0, Max=100	50–100	50–100
Initial availability level (IAL)	Min=0, Max=100	0–49	0–49
Reserve availability level (RAL)	Min=0, Max=100	50–100	50–100
Initial response time level (IRL)	Min=0, Max=100	0–49	0–49
Reserve response time level (RRL)	Min=0, Max=100	50–100	50–100
Job size	Integer (Cloud time unit)	2–8	2–8
Negotiation Strategy( $\lambda$ )	Conciliatory( $< 1$ ) Linear( $= 1$ ) Conservative( $> 1$ )	1/3( $< 1$ ), 3.0( $> 1$ )	1/3( $< 1$ ), 3.0( $> 1$ )
Negotiation deadline ( $\tau$ )	Integer (Round unit)	50 rounds	50 rounds

**4.3 Simulations**

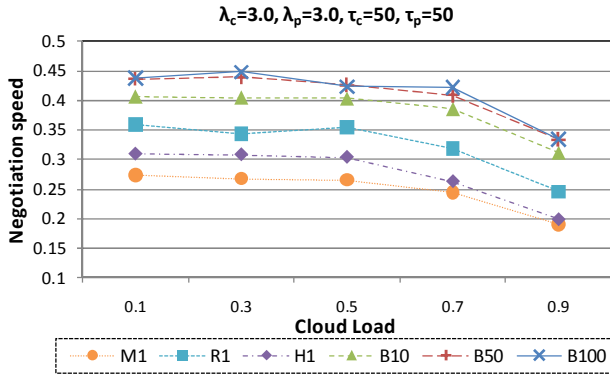
Empirical results were obtained for some representative combinations of the input data (i.e.,  $\{\tau_c : \tau_p\} = \{50:50\}$ ,  $CL = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ , and negotiation agents adopting  $\lambda_c = \{1/3\}$  and  $\lambda_p = \{1/3\}$ , or  $\lambda_c = \{3.0\}$  and  $\lambda_p = \{3.0\}$ ). The performance measures (i.e., negotiation speed and average total utility of the negotiating pair) were then simulated for all burst modes while changing the number of proposals in each burst proposal (i.e., B10, B50, and B100) and related schemes (i.e., M1, R1, and H1). M1 refers to a tradeoff scheme that selects a middle point of total utility line to generate a tradeoff proposal, and R1 refers to a random selection of total utility line to generate a tradeoff proposal. Finally, H1 refers to a heuristic selection based on the similarity of proposals between negotiation agents [7]. We considered M1, R1, and H1 as representative schemes that can generate and evaluate only one proposal in each negotiation round. To interpret the simulations, several graphs for each performance measure were plotted. Fig. 3 shows the performance results of the negotiation speed and Fig. 4 shows the performance results of the total utility.

**4.4 Observations and Results**

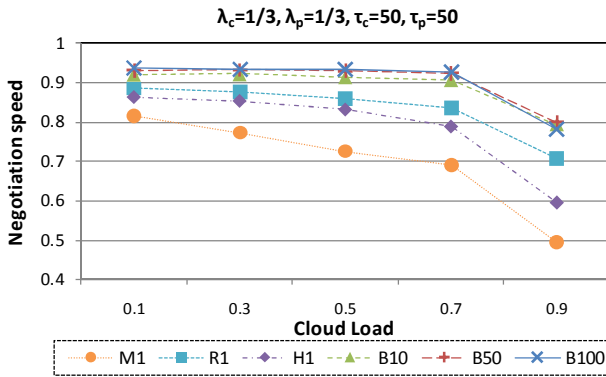
The following observations were made from the experiment results.

- 1) Whereas R1 achieved the fastest negotiation speed among all types of the related schemes (M1, R1, and H1) that can propose only one proposal in each negotiation round, all burst modes (B10, B50, and B100) achieved a faster negotiation speed than the related schemes. Among all types of the burst modes (B10, B50, and B100), B100 achieved the fastest negotiation speed. When both negotiation agents adopted a conservative strategy ( $\lambda_c = \{3.0\}$  and  $\lambda_p = \{3.0\}$ ), the average negotiation speed of B100 is 9% faster than R1. Also, when both

negotiation agents adopted conciliatory strategy ( $\lambda_c = \{1/3\}$  and  $\lambda_p = \{1/3\}$ )), the average negotiation speed of B100 is 7% faster than R1. The reason why the proposed tradeoff algorithm achieves a higher negotiation speed is that the concurrent proposals can give many options to choose to negotiation opponent. With the concurrent proposals, the opponent may find a satisfied proposal, and it can increase the negotiation speed.



(a) Conservative strategy



(b) Conciliatory strategy

Fig. 3. Negotiation speed.

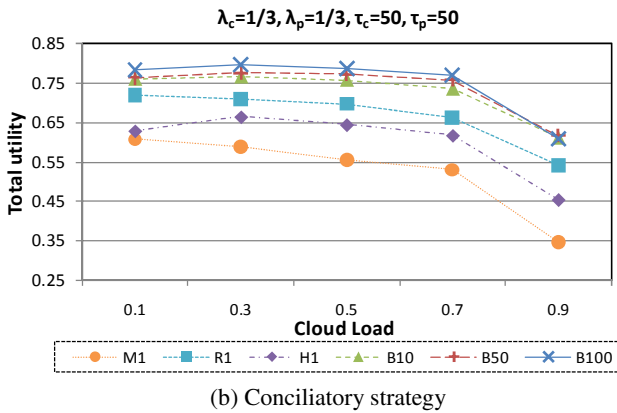
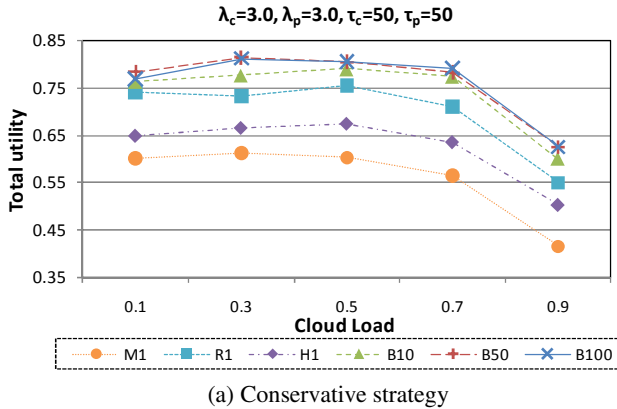


Fig. 4. Total utility

- 2) The burst mode (B10, B50, and B100) achieved a higher average total utility than the other trade off schemes (M1, R1, and H1). Among all types of the burst modes (B10, B50, and B100), B100 achieved the highest average total utility. Also, among all types of related schemes (M1, R1, and H1) that can propose only one proposal in each negotiation round, R1 achieved the highest average total utility. When both negotiation agents adopted a conservative strategy ( $\lambda_c = \{3.0\}$  and  $\lambda_p = \{3.0\}$ ), the average total utility of B100 is 7% higher than R1. Also, when both negotiation agents adopted conciliatory strategy ( $\lambda_c = \{1/3\}$  and  $\lambda_p = \{1/3\}$ ), the average total utility of B100 is 9% higher than R1. Even though [7] reported that the similarity on utilities between negotiation agents can increase the utility of the negotiation outcome, R1 gives better utility than H1 in this negotiation mechanism. It is because the similarity on utilities between negotiation agents is less effective with time slot utility function. The timeslot utility function designed in [5] is a nonlinear utility function and the timeslot

utility can be totally different between a consumer and a provider. Whereas R1 achieved the highest total utility among all types of the related schemes, all burst modes (B10, B50, and B100) achieved a higher total utility than the related schemes because the concurrent proposals can give many options to choose to negotiation opponent. Therefore, with the concurrent proposals, the opponent may find a more satisfied proposal, and the burst mode can increase the total utility.

- 3) Both performance measures with burst modes increase as the number of proposals encoded in the burst proposal increases, but the ratio of increments decreases as the number of proposals increases. For instance, when both negotiation agents adopted a conservative strategy ( $\lambda_c = \{3.0\}$  and  $\lambda_p = \{3.0\}$ ) and Cloud load is 0.3, the difference of negotiation speed between B100 and B50 is only 1% even though an agent increased the number of concurrent proposals 2 times.

## 5 Conclusion and Future Work

The novelty and the significance of this work are the design of the negotiation mechanism that facilitates the Price-Timeslot-QoS negotiation for establishing SLAs of Cloud Service Reservation. The contributions of this paper are detailed as follows:

- 1) Whereas [5] introduced a multi-issue negotiation mechanism for both price and timeslot, in this paper, a Cloud service reservation system supporting concurrent price, timeslot, and QoS level negotiation is designed. To design an automated negotiation, first, we classified Cloud QoS issues into 3 categories (i.e., performance, availability, and response time) and defined QoS utility functions. Also, this paper extends a novel tradeoff algorithm, referred to as the “burst mode” for concurrent price and time slot negotiation to concurrent price, timeslot, and QoS level negotiation. Then, we verified the performance of the proposed tradeoff algorithm is applicable for concurrent price, timeslot, and QoS level negotiation.
- 2) Empirical results obtained from simulations carried out using an agent-based testbed show that the use of the proposed tradeoff algorithm enables Cloud participants to quickly reach agreements and successfully acquire/lease desired Cloud services in terms of price, timeslot, and QoS level in a higher utility than related approaches.

Finally, with the proposed tradeoff algorithm, agents can specify only a definite number as the number of proposals in concurrent set. However, it requires more computational load even though the algorithm enhances negotiation speed and total utility. Accordingly, as a future work, this paper can be extended by considering an adaptive scheme to reduce computational load by adaptively selecting the number of proposals in concurrent set.

**Acknowledgments.** This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST 2009- 0065329) and DASAN (project code: 140309).

## References

1. Foster, A., Roy, V. S.: A Quality of Service architecture that combines resource reservation and application adaptation. In: 8th International Workshop on Quality of Service (2000)
2. Netto, M.A., Bubendorfer, K., Buyya, R.: SLA-Based Advance Reservations with Flexible and Adaptive Time QoS Parameters. In: 5th International Conference on Service-Oriented Computing, Vienna, Austria (September 2007)
3. Foster, Kesselman, C., Lee, C., Lindell, B., Nahrstedt, K., Roy, A.: A distributed resource management architecture that supports advance reservations and co-allocation. In: 7th International Workshop on Quality of Service (IWQo 1999), IEEE CS Press, London (1999)
4. Venugopal, S., Chu, X., Buyya, R.: A negotiation mechanism for advance resource reservation using the alternate offers protocol. In: 16th Int. Workshop on Quality of Service (IWQoS 2008), Twente, The Netherlands (June 2008)
5. Son, S., Sim, K.M.: A Multi-issue Negotiation Mechanism for Cloud Service Reservation. In: Annual International Conference on Cloud Computing and Virtualization (CCV 2010) (May 2010)
6. Sim, K.M.: Grid Resource Negotiation: Survey and New Directions. *IEEE Trans. Syst., Man, Cybern. C, Applications and Reviews* 40(3), 245–257 (2010)
7. Yan, J., Kowalczyk, R., Lin, J., Chhetri, M.B., Goh, S.K., Zhang, J.: Autonomous service level agreement negotiation for service composition provision. *Future Generation Computer Systems* 23(6), 748–759 (2007)
8. Sim, K.M.: Towards Complex Negotiation for Cloud Economy. In: Bellavista, P., Chang, R.-S., Chao, H.-C., Lin, S.-F., Sloot, P.M.A. (eds.) *GPC 2010. LNCS*, vol. 6104, pp. 395–406. Springer, Heidelberg (2010)
9. Rubinstein.: Perfect equilibrium in a bargaining model. *Econometrica* 50(1), 97–109 (1982)
10. Papazoglou, M.P., Traverso, P., Dustdar, S., Leymann, F.: Service-Oriented Computing: State of the Art and Research Challenges. *Computer* 40(11), 38–45 (2007)
11. Menasce, D.A.: Qos Issues in Web Services. *IEEE Internet Computing* 6(6), 72–75 (2002)
12. Jafarpour, N., Khayyambashi, M.R.: QoS-aware Selection of Web Service Compositions using Harmony Search Algorithm. *Journal of Digital Information Management* 8(3), 160–166 (2010)
13. Sim, K.M.: Equilibria, Prudent Compromises, and the “Waiting” Game. *IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics* 35(4), 712–724 (2005)

# Author Index

- Abednejad, Davood 93  
Afzal, Muhammad Khalil 120  
Agarwal, Nitin 224  
Aksoy, Cihan 319  
Alghazzawi, Daniyal M. 47  
Ali, Alwan A. 212  
Asghar, Sohail 373, 383  
Azmi, Masrah Azrifan 164
- Baba, Kensuke 195  
Barjini, Hassan 82  
Ben Ayed, Ghazi 105  
Bhukya, Sreedhar 203
- Caplat, Guy 147  
Challita, Khalil 13  
Cherifi, Chantal 319  
Chettibi, Saloua 128  
Cheung, Ronnie 292  
Chi, Chi-Hung 364  
Chikhi, Salim 128
- Deb, Suash 53  
Ding, Chen 364  
Dykimching, Alyssa Marie 67
- Farhat, Hikmat 13  
Fathi, Leila 37  
Fatimah, Sidi 212  
Fendley, Mary 251  
Fong, Simon 53, 373, 383
- Ghasemi, Safiye 269  
Ghernaouti-Hélie, Solange 105
- Hamidah, Ibrahim 212  
Hasan, Syed Hamid 47  
Huang, Yuan-Ko 240  
Hussain, Kashif 373
- Ibrahim, Hamidah 24, 37, 82, 164  
Ito, Eisuke 195  
Izura, Udzir Nur 212
- Javed, Muhammad Younas 120
- Kazemian, Hassan B. 292  
Kazmi, Madiha 120  
Kidambi, Phani 251  
Korica-Pehserl, Petra 334
- Labatut, Vincent 319  
Latif, Atif 334  
Lee, Jan Aaron Angelo 67  
Lee, Vincent C.S. 280  
Lim, Merlyna 224  
Lin, Jyh-Shyan 423  
Lin, Lien-Fa 240  
Lux Wigand, F. Dianne 307
- Ma, Xiuqin 357  
Mamat, Ali 24, 37  
Mirabi, Meghdad 24, 37  
Miura, Takao 409  
Mlýnková, Irena 179  
Momeni, Ladan 93  
Mori, Masao 195
- Naqvi, Mohsin 373  
Narata, Satoshi 409  
Narayanan, S. 251  
Naz, Sheneela 383
- Okba, Kazar 147  
Ong, Kok-Leong 280  
Othman, Mohamed 82
- Pears, Russel 395
- Qayyum, Amir 383  
Qin, Hongwu 357
- Rahmani, Amir Masoud 269  
Rezazadeh, Arshin 93
- Saber, Benharzallah 147  
Salim, Trigui Mohamed 47  
Santucci, Jean-François 319  
Sim, Kwang Mong 432



- Son, Seokho 432  
Sriharee, Gridaphat 345  
Sulaiman, Norrozila 357  
Svoboda, Martin 179  
  
Tohidi, Hossein 164  
  
Udzir, Nur Izura 24, 37  
Usman, Muhammad 395  
  
Wang, Hao 1  
Wigand, Rolf T. 224  
Wu, Dezhi 136  
  
Yang, Xin-She 53  
Yip, Tan Chik 212  
Yu, William Emmanuel 1, 67  
  
Zalaket, Joseph 13  
Zhao, Yun Wei 364