Volume 37 Number 2 June 2013 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Grid, Cloud and Sky Applications for Knowledge-based Industries and Businesses

Guest Editors:

Vlado Stankovski

Dana Petcu

Editorial Boards, Publishing Council

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editors' Introduction to the Special Issue on "Grid, Cloud and Sky Applications for Knowledge-based Industries and Businesses"

In the past decade, we have witnessed spurt of activity in the area of distributed computing. Several new distributed computing paradigms have emerged that promise to facilitate the delivery of software-based services at unprecedented scale. This includes the emergence of Grid, Cloud, Sky and Fog computing technologies.

These middleware technologies are increasingly used for the development of data and computationally intensive, Web-based applications, involving the use of geographically dispersed resources and potentially huge number of users. With our current experience, it may be forseen that in the future applications will be completely detached from the underlying infrastructures and will be able to elastically scale based on dynamically changing requirements. Middleware solutions facilitating such applications are currently of interest to many knowledge based industries and businesses, active in the areas of engineering, finance, medicine, biology, pharmacy, telecommunications and so on, since they are facing challenging scientific and engineering problems.

Important research and technology development areas at the moment include the investigation of industrial and scientific requirements for distributed computing applications, architectural considerations, the use of the Model Driven Architecture in the software services area, the integration of software services and the Internet of Things, development of new business models for software services, investigation of the possibilities for migration of legacy codes across Cloud and Grid environments, the evolution of standards related to software services and so on.

This Special Issue is based on an open Call for Papers, but, it also includes exended version of selected papers, which were presented at the 4^{th} Workshop on Software Services (WoSS 4) and at the 1^{st} International Conference on CLoud Assisted ServiceS (CLASS 2012) that took place from October 22-25, 2012 in Bled, Slovenia.

The Special Issue contains six papers presenting both application and technology oriented approaches.

The paper of Peter Peer *et al.* presents a Cloud-based fingerprint service which is integrated with the e-learning framework Moodle. The paper discusses the various issues that need to be considered when designing Cloud-based biometric services.

Pawel Czarnul's paper focuses on creation of an effective dynamic ranking service for Infrastructure as a Service, Platform as a Service and Software as a Service providers.

The paper of the authors Ivan Tomašić *et al.* describes the application of Hadoop modules for processing and analyzing large amounts of tabular data acquired from a computer simulation of heat transfer in bio tissues.

The paper of the authors Chengying Mao and Jifu Chen focuses on prediction of the Quality of Service of various software services available over the Internet.

The paper of Ravi Singh Pippal *et al.* deals with the possibility to improve the security aspects when using Cloud services.

The work of the authors Zahra Pooranian *et al.* focusses on optimisation and improvements of a grid scheduling algorithm.

At this point, we would like to thank professor Matjaž Gams for the opportunity to publish this Special Issue, the authors for sharing the results of their research and the members of the WoSS 4 Program Committee: Pawel Czarnul, Janis Grabis, Matjaž B. Jurič, Andras Micsik, Enn Õunapuu, Tomas Pitner for their contribution to the Workshop and for reviewing the papers submitted to this Special Issue.

Vlado Stankovski
Dana Petcu

Building Cloud-based Biometric Services

Peter Peer and Jernej Bule
Faculty of Computer and Information Science
University of Ljubljana, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
E-mail: {jernej.bule, peter.peer}@fri.uni-lj.si

Jerneja Žganec Gros and Vitomir Štruc[1]
Alpineon d.o.o., Ulica Iga Grudna 15, SI-1000, Slovenia
[1]Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
E-mail: {vitomir.struc, jerneja.gros}@alpineon.com

Keywords: biometrics, cloud computing, cloud integration, SaaS, fingerprint recognition

Received: December 4, 2012

Over the next few years the amount of biometric data being at the disposal of various agencies and authentication service providers is expected to grow significantly. Such quantities of data require not only enormous amounts of storage but unprecedented processing power as well. To be able to face this future challenges more and more people are looking towards cloud computing, which can address these challenges quite effectively with its seemingly unlimited storage capacity, rapid data distribution and parallel processing capabilities. Since the available literature on how to implement cloud-based biometric services is extremely scarce, this paper capitalizes on the most important challenges encountered during the development work on biometric services, presents the most important standards and recommendations pertaining to biometric services in the cloud and ultimately, elaborates on the potential value of cloud-based biometric solutions by presenting a few existing (commercial) examples. In the final part of the paper, a case study on fingerprint recognition in the cloud and its integration into the e-learning environment Moodle is presented.

Povzetek: Predstavljene so metode za biometrično razpoznavanje oseb, realizirane v oblaku.

1 Introduction

When talking about Internet authentication, in most cases, people are still talking about passwords. One of the biggest problems with current authentication approaches is the existence of too many password-account pairings for each user, which leads to forgetting or using the same username and password for multiple sites [1]. A possible solution to this problem can be found in the use of biometrics [2]. Biometric authentication techniques, which try to validate the identity of an user based on his/her physiological or behavioral traits, are already quite widely used for local authentication purposes (for private use), while their use on the Internet is still relatively modest. The main reason for this setting is open issues pertaining mainly to the accessibility and scalability of existing biometric technology.

Similar issues are also encountered in other deployment domains of biometric technology, such as forensics, law-enforcement and alike. For example, according to [3], the biometric databases of the Federal Bureau of Investigation, the US State Department, Department of Defense, or the Department of Homeland Security are expected to grow significantly over the next few yours to accommodate several hundred millions (or even billions) of identities. Such expectations make it necessary to devise highly scalable biometric technology, capable of operating on enormous amounts of data, which, in turn, induces the need for sufficient storage capacity and significant processing power.

The first solution that comes to mind with respect to the outlined issues is moving the existing biometric technology to a cloud platform that ensures appropriate scalability of the technology, sufficient amounts of storage, parallel processing capabilities, and with the widespread availability of mobile devices also provides an accessible entry point for various applications and services that rely on mobile clients. Hence, cloud computing is capable of addressing issues related to the next generation of biometric technology, but at the same time, offers new application possibilities for the existing generation of biometric systems [4], [5].

However, moving the existing biometric technology to the cloud is a nontrivial task. Developers attempting to tackle this task need to be aware of:

- the most common challenges and obstacles encountered, when moving the technology to a cloud platform,

- standards and recommendations pertaining to both cloud-based services as well as biometrics in general, and
- existing solutions that can be analysed for examples of *good practices*.

This paper tries to elaborate on the above listed issues and provide potential developers with some basic guidelines on how to move biometric technology to a cloud platform. It describes the most common pitfalls encountered in the development work and provides some directions for their avoidance. Additionally, it presents a case study on fingerprint recognition in the cloud, where the presented guidelines are put into action. The main motivation for the paper stems from our own work in the field of cloud-based biometric services[1] and the fact that the available literature on this field is extremely limited.

The rest of the paper is structured as follows. In Section 2 the existing literature pertaining to biometrics in the cloud is surveyed and differences with this paper are highlighted. In Section 3 some basic characteristics of cloud computing, biometrics, and cloud-based biometric services are presented. In Section 4 issues to consider when developing cloud-based biometrics are elaborated on. In Section 5 a case study on fingerprint recognition in the cloud is presented and, finally, the paper is concluded with some final comments and directions for future work in Section 6.

2 Related work

Cloud computing is a highly active field of research and development, which gained popularity only a few years ago. Since the field covers a wide range of areas relating to all levels of cloud computing (i.e. PaaS, IaaS, and SaaS), it is only natural that not all possible aspects of the field is appropriately covered in the available scientific literature. This is also true for cloud-based biometrics.

While there are some papers addressing this topic, they are commonly concerned with specific aspects of the technology and neglect the bigger picture. The work of Gonzales et. al [7], for example, addresses cloud-based biometrics, but focuses on how to protect biometric data from miss-use through a crypto-biometric system. A similar topic is also discussed by Vallabhu and Satyanarayana in [8]. Other researchers focus more on developing biometric technology for a certain biometric modality and present cloud computing as a possible use-case [9], [10]. This paper, on the other hand, tries to cover different aspects of cloud-based biometrics and is equally interested in legal (e.g., issues relating to data protection, data retention etc.) as well as technical issues. From this point of view, the topic of the paper is more closely related to the work of Senk and Dotzler [11] or Kohlwey et. al [12], where biometrics and cloud computing are also discussed in a broader context in addition to presenting a case study on a specific modality.

3 Biometrics and cloud computing

3.1 Cloud computing

Cloud computing is a computing model, where resources such as computing power, storage, network and software are abstracted and provided as services on the internet in a remotely accessible fashion [13].

NIST defines five key characteristics of cloud computing [14]:

- *Rapid elasticity* - elasticity is defined as the ability to scale resources both up and down as needed. To the consumer, the cloud appears to be infinite, and the consumer can purchase as much or as little computing as needed [14].
- *Measured services* – certain aspects of the cloud service are controlled and monitored by the cloud provider. This is crucial for billing, access control, resource optimization, capacity planning and other tasks [14].
- *On-demand self-service* - a consumer can use cloud services as needed without any human interaction with the cloud provider [14].
- *Ubiquitous network access* - the cloud provider's capabilities are available over the network and can be accessed by various clients through standard mechanisms [14].
- *Resource pooling* - allows a cloud provider to serve its consumers via a multi-tenant model. Physical and virtual resources are assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources, but may be able to specify location [14].

Clearly, cloud computing has several desirable characteristics, which make the cloud platform highly suitable for various applications, including biometrics.

3.2 Biometric systems

Biometric recognition systems represent pattern recognition systems, capable of recognizing individuals based on their physiological or behavioural traits [2]. These traits are considered to be unique to each individual and unlike knowledge or token-based security mechanisms cannot be forgotten, lost or stolen. The most common traits used for biometric recognition are: faces, fingerprints, irises, palm-prints, speech etc.

[1] Conducted in the scope of the KC CLASS (CLoud Assisted ServiceS) project. [6]

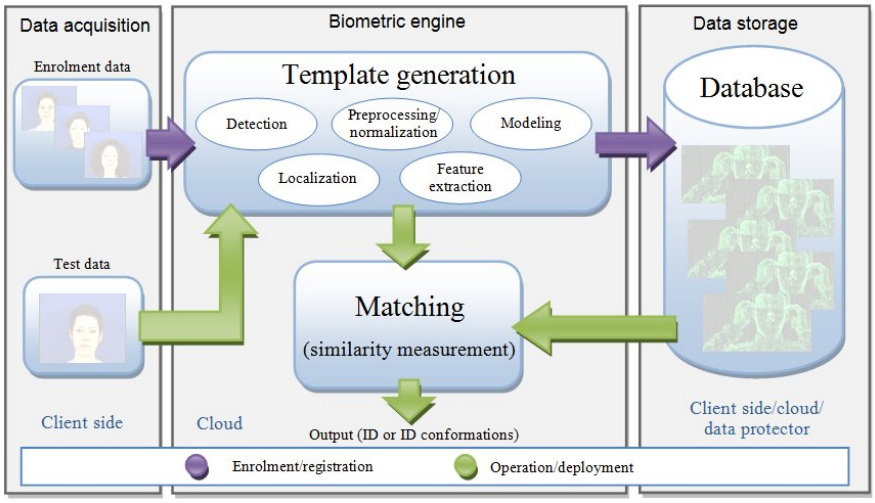


Figure 1: Block diagram of a typical biometric recognition system.

Biometric systems typically conduct one of two tasks: *identification* or *verification/authentication*. The verification/authentication task tries to validate the identity claim of the user currently presented to the system, while the identification task tries to determine, which of the registered user the acquired "live" biometric sample corresponds to. Hence, the identification problem is commonly considered to be a one-to-N matching problem, while the verification/authentication problem is considered to be a one-to-one matching problem.

Biometric systems always comprise the same basic components regardless of whether they are designed for the cloud or any other platform. These components, which are also shown in Fig. 1 for the case of a face recognition system, include [2], [4]:

i) *a data acquisition component* (or sensor) that captures a still image or video sequence of a user trying either to enrol into the system or to use the system for authentication/identification purposes,

ii) *a template generation component* that uses machine learning, computer vision and pattern recognition techniques to derive a biometric template from the input data,

iii) *a database of biometric templates* belonging to enrolled/registered users, and

iv) *a matching component* that compares the biometric template derived from the "live" image with the appropriate template(s) stored in the database of the system and based on the outcome makes a decision regarding the identity of the user currently presented to the system.

While the basic layout of a biometric recognition system is more or less the same on any platform (and biometric modality), there are, however, a number of aspects that are specific to the cloud. These aspects will be discussed in more detail in the next section.

3.3 Biometrics in the cloud

As emphasized in the previous section, there are certain aspects of biometric systems that are specific to cloud computing. First of all, the biometric engine[2] is located in the cloud and not on some local processing unit, as it is the case with traditional (e.g. access control) biometric recognition systems. This characteristic makes the cloud-based biometric technology broadly accessible and provides the necessary means for integration in other security and/or consumer applications. Second of all, storing biometric data in the cloud makes the system highly scalable and allows quick and reliable adaptation of the technology to an increasing user base [3].

On the other hand, storing biometric data in the cloud may raise privacy concerns and may not be in accordance with national legislation. Last but not least, a cloud implementation of biometric technology may harvest all merits of the cloud, such as real-time and parallel processing capabilities, billing by usage etc. [3]. All of the presented characteristics make cloud-based biometric recognition technology extremely appealing.

When developing biometric technology for the cloud, one needs to make a number of design choices. Probably the most important choice is, which components to move to the cloud and which to implement locally. A review of some existing market solutions ([15], [16], [17], [18], [19]) from the field of cloud-based biometrics reveals that most often both the biometric engine as well as the biometric database is moved to the cloud. The commercial solutions typically operate on the principle of the client-server model. The local client (e.g. on the user's computer) is responsible for capturing a biometric sample of the user and sending it to the server (hosted in the cloud), where the matching process is executed. For the safety of the network traffic between the client and the server designated security protocols are commonly used.

[2] We will refer to the template generation and matching components as the biometric engine in the remainder of the paper.

While the presented configuration makes full use of the merits of the cloud platform, it may not be conformant with the local legislation. Therefore, the possibility of using a locally hosted database needs to be considered when designing a cloud-based biometric system. Such a setting may limit the scalability of the technology to a certain extent, but is reasonable as it makes potential market-ready technology more easily adjustable to currently existing legislation. Another possible solution to the legislation problem could also be found in the use hybrid clouds.

4 Integrating biometrics in the cloud

4.1 Challenges and obstacles

When developing biometric technology for the cloud, one inevitably encounters a number of challenges and obstacles that need to be addressed. Next to meeting performance criteria and selecting the most suitable platform for the development work, current legislation pertaining to cloud computing and biometrics in general, privacy concerns and data protection issues all represent major challenges for the development process [4].

The challenges pointed out above are addressed in different ways. The performance of the biometric recognition technology can systematically be evaluated using established reproducible scientific methodology. Here, publicly available databases with predefined experimental protocols and performance criteria are typically employed to produce performance estimates that can be compared with performance estimates of previously assessed technology.

The platform used in the development work is commonly selected according to ones preferences or with respect to the planned characteristics of the final product (i.e. deployable in a private or public cloud etc.).

When it comes to legal, privacy and data protection concerns, there are usually no universal solutions, as they differ from country to country. In the case of Slovenia, for example, the information officer has composed several guidelines/recommendations both for the cloud as well as biometric technology. The recommendations relating to biometric technology, biometric data protection and template storage can be found in [20] and fall in the domain of ZVOP-1 (in Slovenian: *Zakon o varstvu osebnih podatkov*), while the guidelines for cloud computing are accessible from [21].

4.2 Standards and recommendations

There are several standards and recommendations that are relevant in the context of both biometric recognition as well as cloud computing. These include internet protocols, data formats, communication and security protocols, recommendations for cloud application design, recommendations for biometric technology design etc. Since this field is too broad to be covered completely, the focus of this paper is only on a small number of important standards related to biometric recognition technology in the cloud.

The first group of standards of interest for every developer working in the field of biometric recognition are standards that allow for interoperability among different vendors (e.g. [22], [23]). These standards define interchange formats for biometric data and (next to interoperability) also enable consolidation of different biometric databases. The standard in [23], for example, specifies interchange formats for face images and as such defines full-frontal and token face images (defined by the location of the eyes) and ensures that enrolled images meet a sufficient quality standard for arbitrary face recognition technology. Similar standards also exist for other biometric traits [24].

The second group of standards of relevance to cloud-based biometrics is the OASIS standard for Biometric Identity Assurance Services (BIAS) [25]. The open standard defines all specifications for SOAP-based biometric services and is conveniently supported by a reference implementation (for fingerprints) provided by NIST. The ISO/IEC JTC 001/SC 37 has just recently approved a project to internationalize the above mentioned BIAS standard.

4.3 Deployment possibilities and existing solutions

Cloud-based biometric technology offers attractive deployment possibilities, such as smart spaces, ambient intelligence environments, access control applications, mobile application, and alike. While traditional (locally deployed) technology has been around for some time now, cloud-based biometric recognition technology is relatively new. There are, however, a number of existing solutions already on the market, these include (among others) the solutions by Animetrics [15], BioID [16] and, of course, Face.com [17], which has recently been acquired by Facebook.

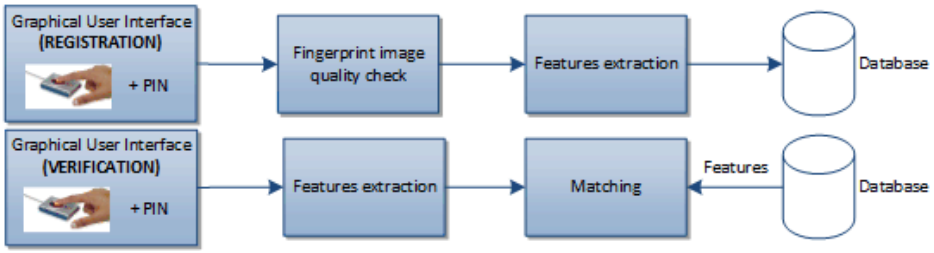


Figure 2: Simplified block diagram of biometric registration and verification.

5 A case study: fingerprint recognition in the cloud

5.1 Goal and setup

The goal of the case study presented in the remainder is to put the general guidelines presented in the previous sections into practice and provide more detailed (technical) information on the process of integrating biometric technology into a cloud platform. The basis of the case study represents a prototype fingerprint recognition systems, named FingerIdent [26]. A local test version of this prototype system is already installed at the Faculty of Computer and Information Science, University of Ljubljana, in front of the Computer Vision Laboratory.

The functionality of the existing local version of the FingerIdent system can be divided into two main categories:

i) *user registration (enrollment)*, during which a biometric template of a given user is constructed and stored in the system's database, and

ii) *user verification*, during which the identity claim of a given user is validated.

The registration process uses a fingerprint reader to capture the (biometric) fingerprint data. In the next phase the quality of the captured sample is evaluated and if it is found to be adequate, the system extracts features from it and stores them in the form of a biometric template in the database. During the verification process features from the captured "live" fingerprint are again extracted and compared to those stored in the database. The comparison is made based on pattern matching procedures, which form the foundation for the validation of the identity claim. An illustration of both functions is shown in Fig. 2.

To reach the goal of devising a cloud-based biometric service, one needs to migrate the presented functionality of the local FingerIdent system to the cloud and provide the necessary infrastructure for accessing the biometric service. Details on this procedure are given in the next section.

5.2 Designing cloud biometric services

It was emphasized in Section 3.3 that a decision has to be made with respect to which components of the biometric system should be moved to the cloud and which implemented locally. For our case study, we decided to move the biometric engine as well as the biometric database to the cloud. A block diagram of the complete cloud-based biometric service design is shown in Fig. 3.

Note that the verification process with the described design is conducted using the following scenario:

i) the fingerprint of a given user is first captured via a fingerprint scanner (here scanner libraries that allow capturing fingerprint images need to be integrated into the local (desktop or/and web) application);

ii) the application then communicates through a (REST) API with the biometric web service hosted in the cloud and sends an encoded image to the fingerprint processing library (i.e. FingerIdent library) that provides the functionality for the cloud service;

iii) the transmitted fingerprint image is processed in the cloud and finally the result is sent back to the local application.

The security of the presented solution is provided on different levels through:

- the use of the HTTPS protocol for data transfer,
- the use of certificates (the SSL protocol),
- the encryption of passwords and other data (such as biometric templates) in the database, and
- the protection of the access to the cloud-service with a complex 40-digit password.

The cloud-based service is designed modularly, which makes upgrading the service a relatively simple task. Equally important is the fact that the same design is also suitable for other biometric modalities and allows for devising multi-modal person authentication as well.

5.3 Moodle with fingerprint verification

To demonstrate the effectiveness of the presented solution and to provide a proof-of-concept, the e-learning environment Moodle [27] is augmented with biometric authentication capabilities by integrating it with the cloud-based fingerprint verification service.

Since Moodle is also designed modularly, the biometric authentication procedure is implemented as an additional (optional) authentication scheme, which can complement the existing procedures and provide an additional level of access security. A block diagram of the integration is shown in Fig. 4.

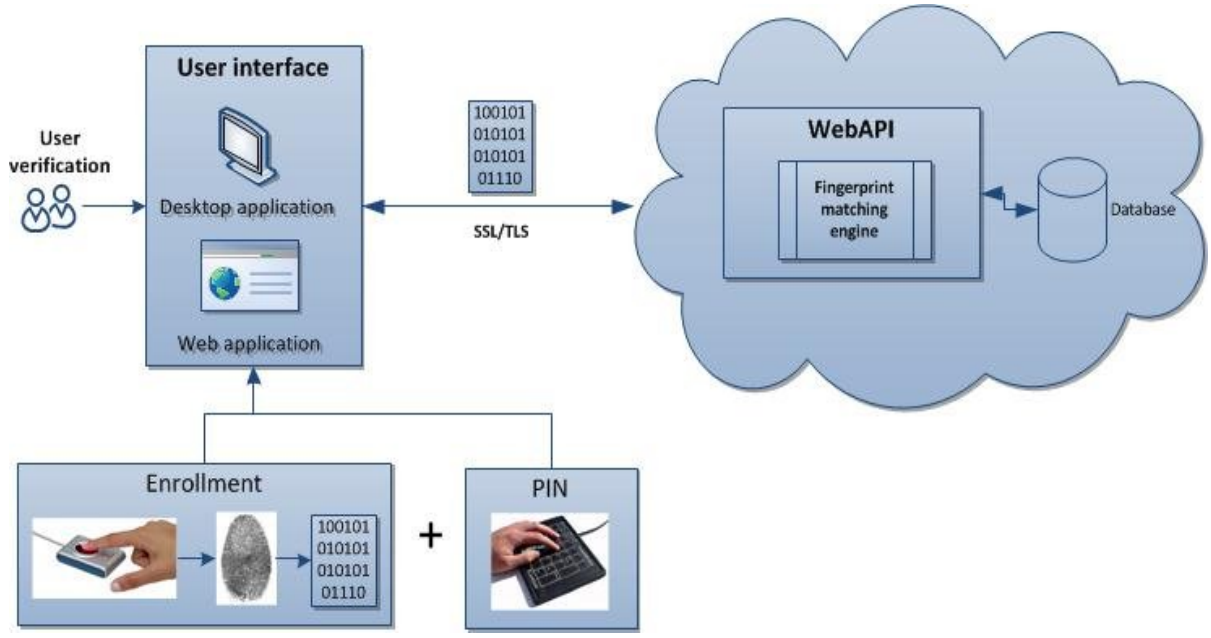


Figure 3: Scheme of the biometric verification system in the cloud.

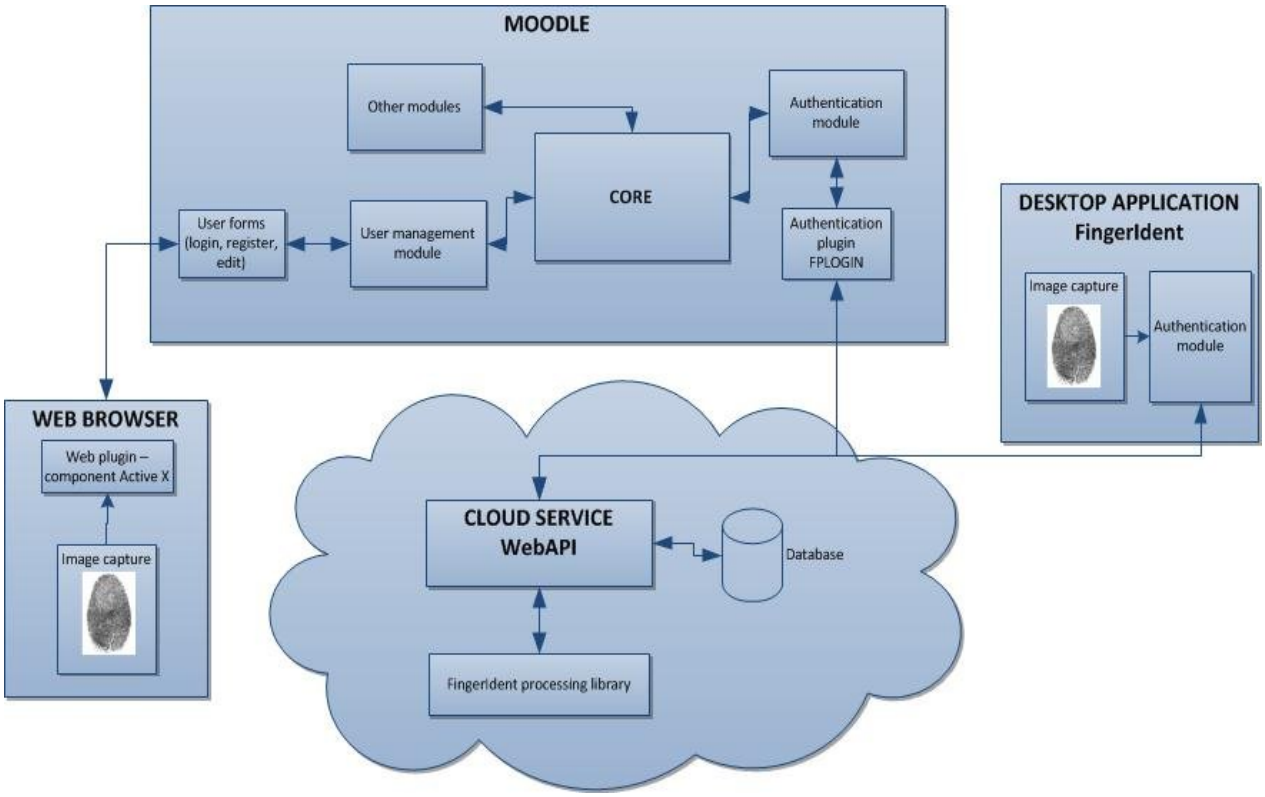


Figure 4: Cloud fingerprint verification in Moodle.

The main problem faced during integration is the compatibility of various fingerprint readers with different browsers. Each manufacturer of fingerprint readers offers their own protocols and libraries to access the corresponding hardware. A standard is not yet available.

The solution developed in the scope of this case study uses an ActiveX component to access the hardware. ActiveX components are officially supported only on Internet Explorer, which represents a weakness in the implementation. As future work, an extension of the presented solution is planned, so it can work with

other popular browsers, such as Firefox, Opera or Chrome too.

After the integration of the fingerprint authentication service into the Moodle framework, the Moodle login screen was modified to account for the added functionality. The result of this procedure is shown in Fig. 5. Note how the added biometric authentication functionality seamlessly integrates into the existing framework.

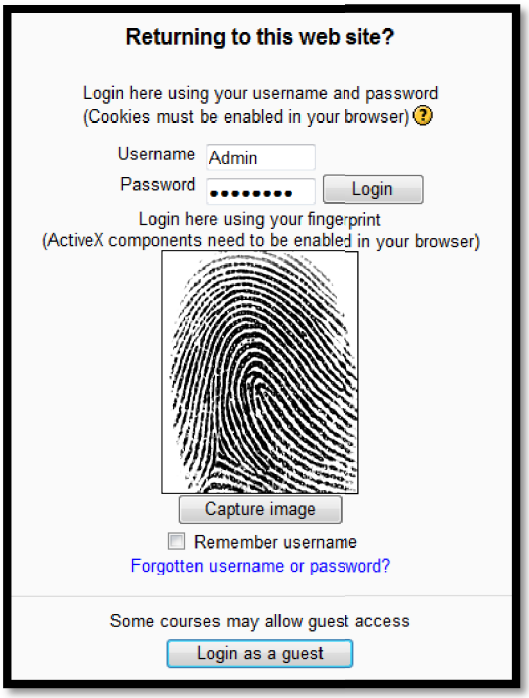


Figure 5: Customized Moodle login.

6 Conclusion

Cloud based biometric services have an enormous potential market value and as such attract the interest of research and development groups from all around the world. In this paper some directions on how to move existing biometric technology to a cloud platform were presented. Issues that need to be considered when designing cloud-based biometric services have been presented and a case study, where a cloud-based fingerprint service was developed and integrated with the e-learning framework Moodle was described as well. As part of our future work we plan to migrate more biometric modalities to the cloud and, if possible, devise a multi-modal cloud-based biometric solution

Acknowledgements

The work presented in this paper was supported by the European Union, European Regional Fund, within the scope of the framework of the Operational Programme for Strengthening Regional Development Potentials for the Period 2007-2013 contract No. 3211-10-000467 (KC Class), the postdoctoral project BAMBI with ARRS ID Z2-4214.

References

[1] D. Balfanz et al., "The future of authentication", *IEEE Security & Privacy*, vol. 10, pp. 22-27, 2012.

[2] A.K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Video Technology*, vol. 14, no. 1, pp. 4-20, 2004.

[3] E. Kohlwey, A. Sussman, J. Trost, and A. Maurer, "Leveraging the Cloud for Big Data Biometrics: Meeting the performance requirements of the Next Generation Biometric Systems," in *Proceeding of the IEEE World Congress on Services*, pp. 597-601, 2011.

[4] V. Štruc and J. Žganec-Gros, "Developing Face Recognition Technology for the KC Class Biometrics service," in: *CLASS Conference 2012*, pp. 68-75, 2012.

[5] J. Bule and P. Peer, "Fingerprint Verification as a Service in KC CLASS," in: *CLASS Conference 2012*, pp. 76-82, 2012.

[6] The KC Class project, available from: http://www.kc-class.eu/, last visited: 5.12.2012.

[7] D. Gonzales Martinez, F.J. Gonzels Castano, E. Argones Rua, J.L. Ala Castro, D.A. Rodriguez Silva, "Secure Crypto-Biometric System for Cloud Computing," in: *International Workshop on Securing Services on the Cloud*, pp. 38-45, 2011.

[8] H. Vallabhu and R.V. Satyanarayana, "Biometric Authentication as a Service on Cloud: Novel Solution," *International Journal of Soft Computing and Engineering*, vol. 2, no. 4, pp. 163-165, 2012.

[9] S. Suryadevara, S. Kapoor, S. Dhatterwal, R. Naaz and A. Sharma, "Tongue as a Biometric Visualizes New Prospects of Cloud Computing Security," in: *International Conference on Information and Network Technology*, vol. 4, 2011.

[10] S.N.S. Raghava, "Iris Recognition on Handoop: a Biometrics System Implementation on Cloud Computing," in: *Proceedings of IEEE CCIS,* 2011.

[11] C. Senk and F. Dotzler, "Biometric Authentication as a Service for Enterprise Identity Management Deployment: A Data Protection Perspective," in: *International Conference on Availability, Reliability and Security*, pp. 43-50, 2011.

[12] E. Kohlwey, A. Sussman, J. Trost, and A. Maurer, "Leveraging the Cloud for Big Data Biometrics: Meeting the Performance Requirements of the Next Generation Biometric Systems," in: *IEEE World Congres on Services,* pp. 597-601, 2011.

[13] D.M. Dakhane and A.A. Arokar, "Data Secuity in Cloud Computing for Biometric Application," *International Journal of Scientific & Engineering Research*, vol. 3, no. 6, pp. 1-4, 2012.

[14] Cloud computing use case discussion group, "Cloud Computing Use Cases: White Paper" available from: http://cloudusecases.org/, last visited: 05.12.2012.

[15] Homepage of the Animetrics cloud-based face recognition solution, available from:

http://animetrics.com/cloud-face-recognition-services/, last visited: 03.10.2012.
[16] Homepage of the BioID cloud-based biometric recognition solution, available from: http://www.bioid.com/, last visited: 03.10.2012.
[17] Homepage of the Face.com cloud-based face recognition solution, available from: http://face.com/, last visited: 03.10.2012.
[18] Homepage of Ceelox ID Online, available from: http://www.ceelox.com/ceeloxidonline.html, last visited: 05.12.2012.
[19] Homepage of PasswordBank IDaaS, available from: http://www.passwordbank.com/passwordbank-private-cloud, last visited: 05.12.2012.
[20] Homepage of the Slovenian Information Commissioner, biometrics, available from: https://www.ip-rs.si/varstvo-osebnih-podatkov/informacijske-tehnologije-in-osebni-podatki/biometrija/, last visited: 03.10.2012.
[21] Information Commissioner, Cloud Security Alliance Slovenia Chapter, Slovenski institut za revizijo, Slovenski odsek ISACA, Zavod e-Oblak, Eurocloud Slovenia, "Varstvo osebnih podatkov & računalništvo v oblaku," pp. 31, 2012, available from: https://www.ip-rs.si/fileadmin/user_upload/Pdf/smernice/Smernice_rac_v_oblaku.pdf, last visited: 03.10.2012.
[22] Information technology, "Biometric data interchange formats – Part 5: Face image analysis," *Documents ISO/IEC 19794-5:2005*, 2004, available from: http://www.iso.org, last visited: 03.10.2012.
[23] Information technology, "Face recognition format for data interchange," *Document 385-2004 ANSI INCITS*, 2004, available from: http://www.iso.org, last visited: 03.10.2012.
[24] NIST standard, ANSI/NIST-ITL 1-2011, NIST Special Publication 500-290, Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information, available from: http://www.nist.gov/itl/iad/ig/ansi_standard.cfm, last visited: 03.10.2012.
[25] OASIS standard, "Biometric Identity Assurance Services (BIAS) SOAP Profile Version 1.0," pp. 210, May 2012, available from: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=bias, last visited: 03.10.2012.
[26] M. Tovšak, J. Bule, P. Peer, "Upgrading a system for verification based on fingerprints," in: *Electrotechnical and Computer Science Conference (ERK)*, vol. B, pp. 135-138, 2011.
[27] Moodle, open-source e-learning software platform, available from: http://moodle.org, last visited: 06.12.2012.

An Evaluation Engine for Dynamic Ranking of Cloud Providers

Pawel Czarnul
Narutowicza 11/12, 80-233 Gdansk, Poland
E-mail: pczarnul@eti.pg.gda.pl and http://pczarnul.eti.pg.gda.pl

Keywords: cloud computing, ranking cloud providers

Received: December 13, 2012

The paper focuses on creation of an effective dynamic ranking service for IaaS, PaaS and SaaS cloud providers. It considers building a quality model for this purpose along with definition of quality measurement procedures. The paper discusses several techniques known from already existing price comparison engines that could be modified and adopted for comparison of cloud providers. A technique for filtering measured data is proposed, in particular to avoid vendor lock-in issues. The paper presents a design and results from an engine for simulation of various ranking algorithms in response to streams of prices from various providers. Examples with various streams of provider prices and resulting rankings are presented that cope with the vendor lock-in issue as well as consider the impact of long or short-term price changes on the ranking.

Povzetek: Članek se osredotoča na izdelavo učinkovite storitve za dinamično rangiranje ponudnikov storitev tipa IaaS, PaaS in SaaS v oblaku.

1 Introduction

Cloud computing has become more and more widespread and popular in today's world with many offerings regarding infrastructure, ready-to-use platforms and services [1]. These can be categorized as follows:

- IaaS – Infrastructure as a Service - making an infrastructure (computing, storage, operating system) with a given configuration available to a client, examples: Google Compute Engine[1], Amazon Elastic Compute Cloud (EC22)[2], RackSpace Cloud Servers[3], Rack Space Cloud Files[4],
- PaaS – Platform as a Service - offering a complete platform with particular software required by users; examples include: Aneka [10], Google AppEngine[5], Windows Azure[6], RedHat Openshift[7], RackSpace Cloud Sites[8],
- SaaS – Software as a Service - particular software that is managed by its provider and accessed by users from any location. Examples include Google Apps[9] and Salesforce[10].

Following search engines and price comparison tools and engines for the traditional marketplaces, there have emerged tools for comparison of cloud offers as well. For instance, as of this writing a web search on "IaaS ranking" returns several surveys on IaaS: either static analyses [11] [12] [13] or rankings that depend on actual parameters of the offers (such as prices) that can change in time [14] [15]. Platforms such as Cloudorado[16] allow to preselect user requirements such as required processor computing capabilities or storage and return a ranking based on that. FindTheBest allows to select a cloud provider based on its type (IaaS, PaaS) but also the control interface, software license or subscription type.

It seems, however, that many of these rankings use unstructured quality comparison models, do not consider how qualities have been changing over time for providers and do not address issues such as vendor lock-in. It is a known fact that some Internet providers or shops used to offer very cheap prices to gain a market share (by being on top places in comparison rankings) only to deceive some customers later. The paper discusses a quality model for a dynamic ranking of cloud providers that addresses these issues. This work extends the concepts presented in [5] by proposing a design and implementation of a simulation engine for running various provider ranking algorithms and presentation of its results for various streams of input price offers from

[1] http://cloud.google.com/products/compute-engine.html
[2] http://aws.amazon.com/ec2/
[3] http://www.rackspace.com/cloud/cloud_hosting_products/servers/
[4] http://www.rackspace.com/cloud/cloud_hosting_products/files/
[5] https://developers.google.com/appengine/
[6] http://www.windowsazure.com
[7] https://openshift.redhat.com/app/
[8] http://www.rackspace.com/cloud/cloud_hosting_products/sites/
[9] http://www.google.com/Apps
[10] http://www.salesforce.com/eu/
[11] http://my-inner-voice.blogspot.com/2011/02/here-are-results.html
[12] http://insidehpc.com/2011/02/10/survey-results-on-cloud-iaas-providers/
[13] http://www.opsource.net/Info-Tech-Cloud-IaaS-Vendor-Landscape
[14] http://www.cloudreviews.com/top-ten/cloud-hosting-services.html
[15] http://cloud-computing.findthebest.com/
[16] http://www.cloudorado.com/

various providers.

The structure of the paper is as follows. Section 2 discusses the problem of quality assessment of services offered on the cloud. Next, Section 3 details the design and implementation of a simulator for ranking input streams of price offers from various providers. Experiments for various input streams are presented in Section 4 which is followed by a summary in Section 5.

2 Quality evaluation of cloud offers

Before educated selection of services can be performed, it is necessary to incorporate measurable quality assessment of the given service. This comprises several aspects that need to be addressed:

1. a quality model/ontology that defines metrics to be measured,
2. quality measurement procedures – e.g. how frequently the metrics should be measured – this may be different for various metrics; for instance availability may require more frequent monitoring than the price,
3. filters applied on top of the measured values – such may be used to address several issues such as:
   - preventing from short-term peaks in measured values to affect output; possibly only longer lasting changes should do that,
   - preventing from one or few providers to occupy top places all the time by offering too good to be true conditions,
   - considering or not sudden changes in the history of the provider which may affect user decisions who might be afraid of similar changes in the future – it may depend on the user whether he or she wants to consider this aspect.

For metrics, it is recommended to adopt and extend the already used techniques for marketplaces in the Internet. Namely, evaluation of the providers using a numerical scale such as [0,10] which is offered for almost any price comparison engine today along with physical location of a particular provider. In this case, a quality ontology is proposed for quality service evaluation of particular IaaS, PaaS, SaaS that will incorporate the following:

`accessibility` [11] – characterizes the network between the client in location and the service, several entries of this type could be inserted,

`availability` [12, 13, 11, 2] – characterizes the availability of the service itself. It can be measured by e.g. checking its availability vs availability of other services/servers in a similar geographical/provider location,

`reputation` [12] – reputation of the provider,

`security` [11] – offered by the provider,

`fidelity` [3] or `conformance` [11] – with standards,

`cost-effectiveness` – evaluated by clients,

`reconfiguration ability` – applicable to IaaS and PaaS,

`interface` – how easy it is to access the infrastructure and upload/download/execute applications.

As suggested in Section 3, various filters can be applied on top of measured values. For instance, a one time peak in measurements of a certain value might not change the overall score of the given metric. Only a longer lasting change would initiate this. A simple average would work as a low-pass filter. The regular average suffers from the historical effect i.e. results from the past affect the final average in the same way as the last input. It may depend on the client whether to rely more just on recent measurements. This could be further extended to a running score e.g. a running average of 10 or 100 values. Alternatively, the history of the provider might be important for the given client.

In order to avoid a situation when one provider wants to dominate the given segment of the market by e.g. using too good to be true prices it is possible to consider a certain number of best offers and rotation on the first ranking places, provided that results returned for the services are closer to each other than a predefined threshold. Even one company could then try to use different providers for parts of their businesses to avoid the lock-in problem.

3 Proposal of an evaluation engine and visualization for ranking algorithms

In a way, the proposed approach can be seen as a solution aiding sky computing [8] as the proposed engine tries to sort out available cloud options and offer best options at a higher level of cloud integration.

As mentioned above, the goal of the engine is to be able to:

1. monitor Quality of Service (QoS) dynamically which refers to periodic measurements of quality metrics applicable to cloud services,
2. avoid potential vendor lock-in problem.

3.1 Proposed simulation engine

Within this paper, the author has developed a simulator implemented in C along with visualization assisted by GNU Plot. The goal of the simulator is to model cloud provider

offers over time and simulate execution of a ranking algorithm that would output certain scores for particular offers at particular moments in time. From the cloud client's point of view that gives the preference in choosing "the best" offer by selecting the top offer. If some particular needs of the client are not considered in the ranking scheme, the next best offer can be selected as well. However, from the global point of view i.e. the population of clients, the ranking algorithm is supposed to provide a solution that copes well with the vendor lock-in issue. Namely, it does not to allow selection of just one best provider at all times even if its offer seems to be the best from the QoS perspective. This is to prevent from dumping practices or similar over a certain period of time just to gain market share.

Let us focus first on one quality metric such as price. The following notation will be used:

- $p_i(t)$ – the price offered by provider i at time t,
- $dp_i(t) = |p_i(t) - p_i(t-1)|$ – the price difference between successive discrete points in time,
- $dap_i^a(t) = \sum_{x=a}^{t} dp_i(x)$ – the accumulated sum of price differences offered by the particular provider; the goal of this metric is to assess an accumulated rate of price changes over period from a until t. The larger $t-a$ is the larger history has an impact on the current value of $dap_i^a(t)$.

The flow of the data through the simulation engine is shown in Figure 1. Several steps are performed including: computing the above values, then computing values $val_i(t) = f(p_i(t), dp_i(t), dap_i^a(t))$ against which sorting will be performed such that the lower the value of $val_i(t)$ the better place in the ranking provider i will be assigned.

Furthermore, this scheme is extensible i.e. it allows modeling of several behaviors of cloud providers as well as easily extend the ranking algorithm with:

new metrics. This can be done by extending the structure that currently contains $p_i(t), dp_i(t), dap_i(t)$. For instance, the metrics can include: reputation of the provider $r_i(t)$, availability $a_i(t)$ etc. This leads to consideration of $dr_i(r), dar_i(t), da_i(t), daa_i(t)$. The final value of $val_i(t)$ would be a function of all these metrics.

application of other digital filters in addition to dp and dap to process the data of a particular cloud provider over successive time steps and works for particular metrics. For instance, depending on the needs and particular metrics, either high or low pass filters can be used.

The whole system consists of the following programs that pass data using standard inputs and outputs as well as additional files:

1. datagenerator – generates input streams of data e.g. price offers,
2. simulator – implementing the aforementioned evaluation algorithm,
3. visualization tool – implemented using custom input scripts and the GNU Plot tool.

3.2 A wider perspective on QoS evaluation

From the client point of view, it would be desirable to have access to a comparison engine like Cloudorado with the aforementioned features. First of all, the engine can consider three categories of: IaaS, PaaS and SaaS. It can first match available offers in terms of functions and then evaluate based on the ranking discussed earlier. In order to make search better, two solutions are feasible:

1. categorization of features such as hardware and software parameters desired by the client:
   - memory size,
   - processor/core/GPU capabilities,
   - storage,
   - operating system,
   - particular software,
   - access interface.

   This is especially suitable for IaaS and PaaS offerings.
2. full text search as in [6]. This allows formulation of desired functions in the form of human readable text. Useful mainly for SaaS as it would allow searching and presentation of SaaS offers for a particular application.

The full text search mechanism could also be applied to any type of service when looking for comments of already existing clients.

This would also naturally lead to creation of runtime registries of particular IaaS, PaaS and SaaS offers [9]. SaaS options could then be categorized into various categories. One possibility is to adopt the well know technique from photo sharing sites i.e. augmenting descriptions with tags. Then selection of particular tags would narrow search results.

4 Experiments

In this section a series of experiments is provided along with graphs presenting:

1. input data from cloud providers i.e. prices offered over time,
2. output ranking from the simulation algorithm using various ranking algorithms.

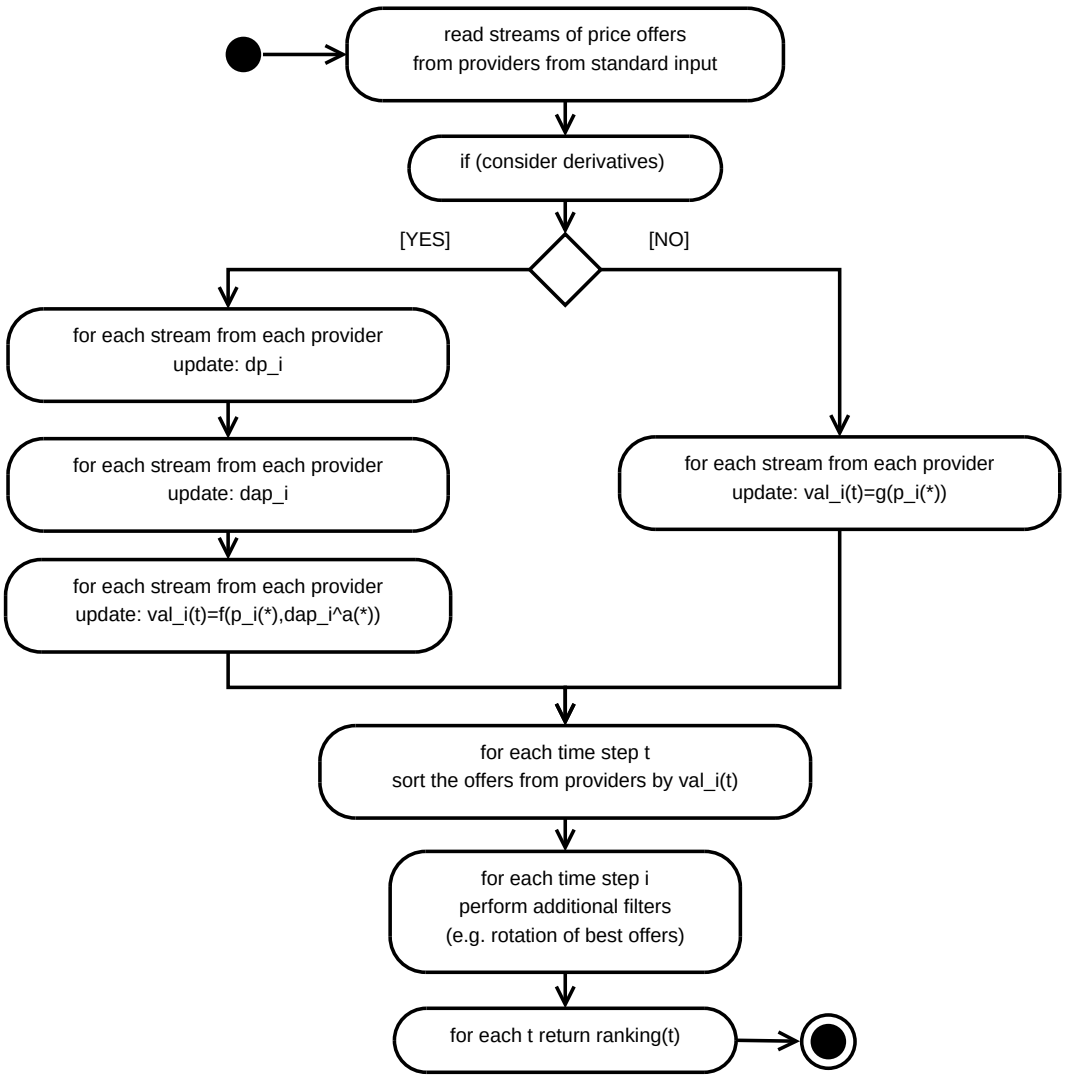


Figure 1: Steps for filtering input providers' offers

The basic assumptions for the following tests are as follows. There are 10 cloud providers that offer a service of a particular type (PaaS, IaaS or SaaS) and adjust their prices in successive time steps by introducing small variations to their base prices as shown in the following figures. For each of the input data streams outputs that denote ranking of particular providers are shown. For the end client, the provider that occupies the top spot at the particular moment should be selected. For each test case, several figures are shown: input streams of unmodified cloud offers, ranking by values that result from functions of the observed original prices and the latter modified by rotation of the best offers in the ranking.

4.1 Stable prices with reasonably small variations over time and elimination of vendor lock-in

For the input shown in Figure 2, the prices from various providers are close to each other which results in slight changes of the ranking by sorting just by $val_i(t) = p_i(t)$. The ranking that resulted from sorting by the current price only is shown in Figure 3. It can be seen that although there are changes in the ranking as the price ranges of some providers overlap, some offers result in the provider occupying one spot at all times. This may result in vendor lock-in if clients would choose the best offer at all times. Figure 4, on the other hand, shows ranking after additional mixing of the three best offers to get rid of this potential problem, as the prices of these providers do not differ by a large margin in absolute terms.

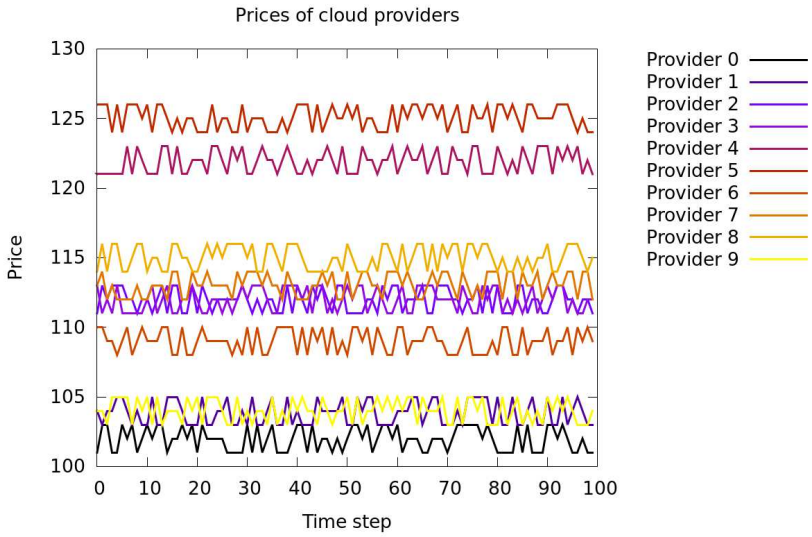


Figure 2: Offers from cloud providers in successive time steps

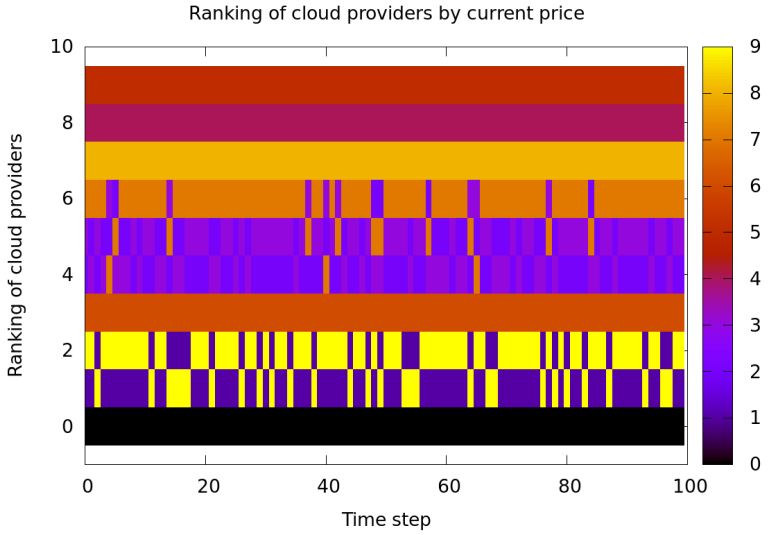


Figure 3: Ranking of cloud providers by $val_i(t) = p_i(t)$

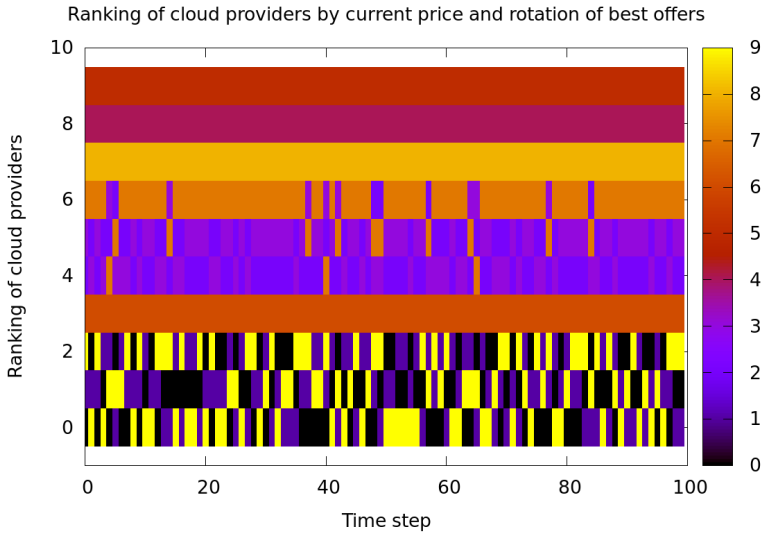


Figure 4: Ranking of cloud providers by $val_i(t) = p_i(t)$ and rotation of the best offers

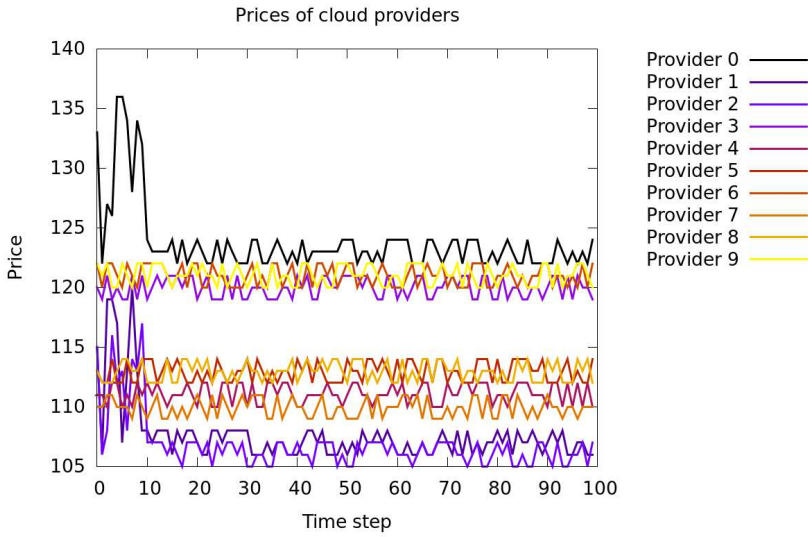


Figure 5: Offers from cloud providers in successive time steps

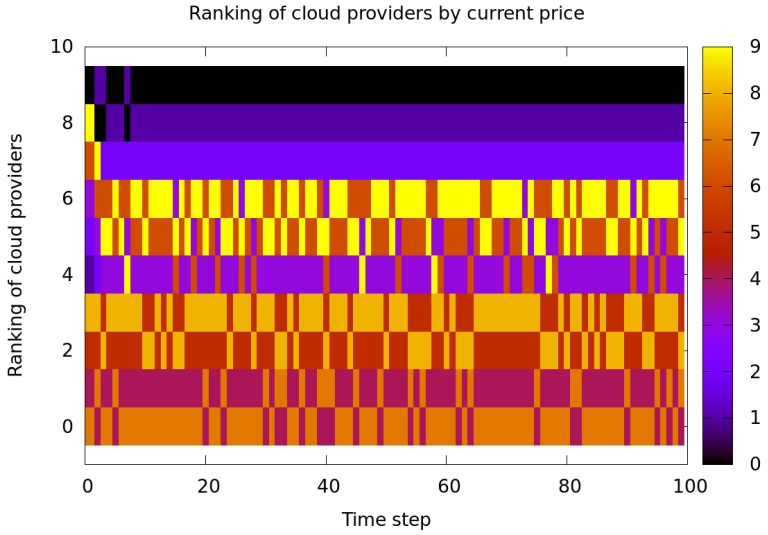


Figure 6: Ranking of cloud providers by $val_i(t) = p_i(t) + dap_i^0(t)$

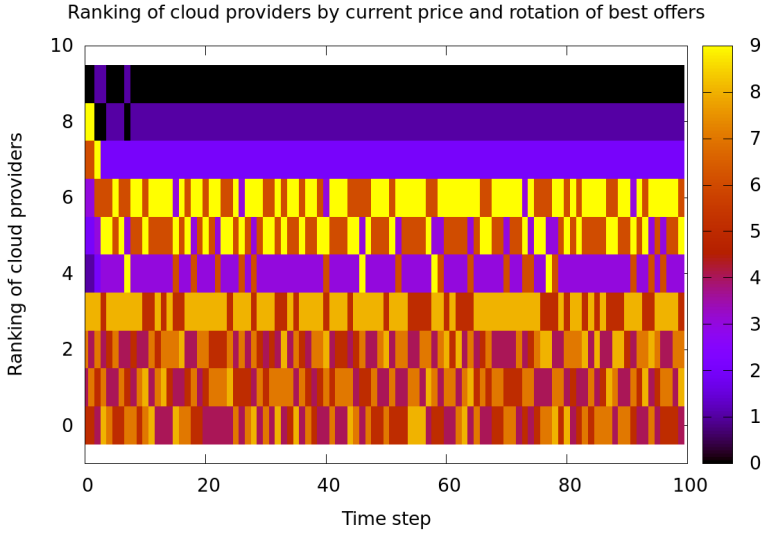


Figure 7: Ranking of cloud providers by $val_i(t) = p_i(t) + dap_i^0(t)$ and rotation of the best offers

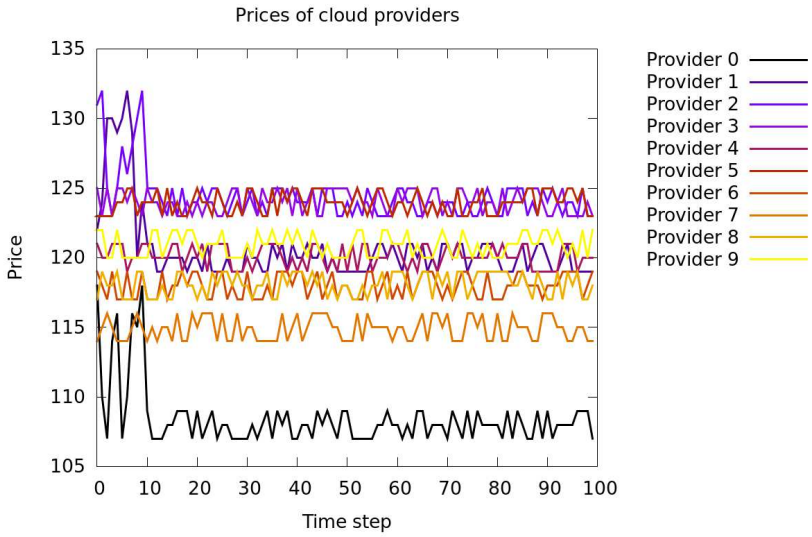


Figure 8: Offers from cloud providers in successive time steps

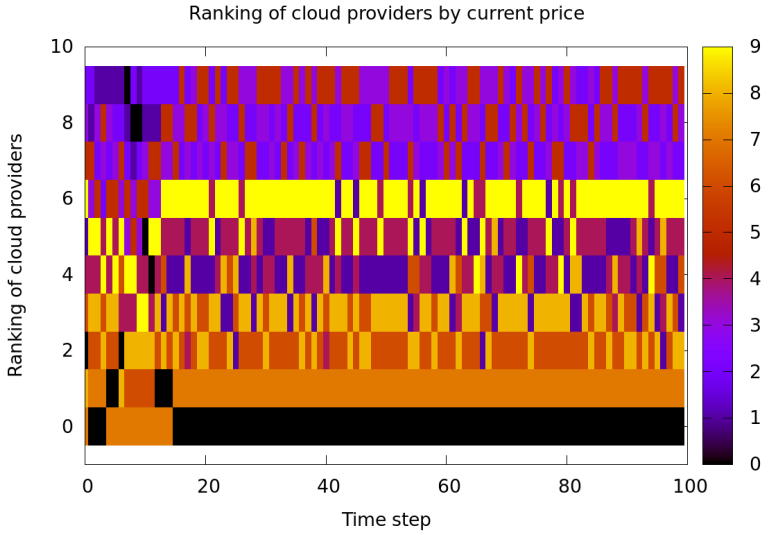


Figure 9: Ranking of cloud providers by $val_i(t) = p_i(t) + dap_i^{t-4}(t)$

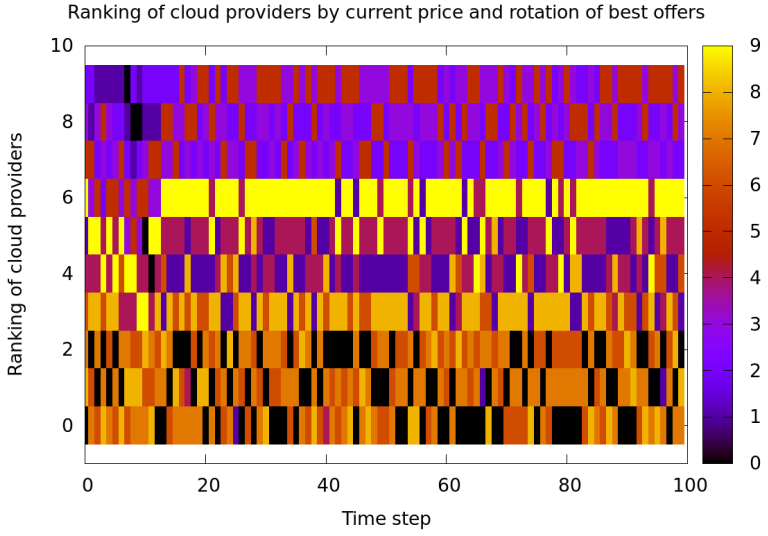


Figure 10: Ranking of cloud providers by $val_i(t) = p_i(t) + dap_i^{t-4}(t)$ and rotation of the best offers

4.2 Considering derivatives in ranking

In the following tests, considerable changes of prices of selected cloud providers were simulated in the first 10 time steps of the simulation. This is shown in both considered inputs in Figures 5 and 8.

Two different solutions were proposed here:

1. $dp_i(t)$s are computed for each time step i.e. absolute values of differences in prices between successive time steps. Then the accumulated sum of $dap_i^0(t)$ is computed. As shown in Figure 6, ranking by $val_i(t) = p_i(t) + dap_i^0(t)$ considers the whole past history of price changes of a particular provider. The larger the derivatives, the smaller chance the provider will occupy top spots of the ranking. It can be clearly seen that even though two providers offer the best current prices in later time steps as shown in Figure 5, the history of larger changes has put them back into further places in the ranking. Figure 7 shows additional mixing of the top three spots.

2. As shown in Figure 9, ranking by $val_i(t) = p_i(t) + dap_i^{t-4}(t)$ considers *only the recent* history of price changes of a particular provider. It can be seen very clearly that the provider offering the best current prices in the initial time steps falls down in the ranking but then recovers to the top spot. Figure 10 shows additional mixing of the top three spots.

Obviously, additional filters and combination of various QoS metrics can be obtained and programmed analogously just by adding additional processing functions to the flow proposed in Section 3. Depending on the client needs, a ranking is then created that allows to select the best offer at any time. For instance, it can also consider the providers that the client has already been using.

5 Summary and future work

The paper presented an idea, design and implementation of a simulator for ranking incoming streams of provider offers that may be applicable for real world cloud offers. The simulator allows to test various algorithms for ranking providers with easy changing to other algorithms or even filters within the algorithms. Practical applications include incorporation of the idea into Internet price comparison engines, cloud service search engines as well as integrated systems for workflow management where services need to be found for workflow subtasks.

Further work will focus on extension of the simulator with new filters and development of an integrated evaluation method for various QoS metrics. Additionally the engine will be deployed in the BeesyCluster middleware for assessment of its services and then used in discovering and incorporation of such services into workflow applications on grids [4]. Such workflows can also be run on clouds [7].

References

[1] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Gener. Comput. Syst.*, 25(6):599–616, June 2009.

[2] G. Canfora, M. Di Penta, R. Esposito, and M.L. Villani. A Lightweight Approach for QoS-Aware Service Composition, 2004. ICSOC forum paper, IBM Technical Report Draft.

[3] Jorge Cardoso, Amit Sheth, and John Miller. Workflow quality of service. Technical report, LSDIS Lab, Department of Computer Science, University of Georgia, Athens, GA 30602, USA, March 2002.

[4] Pawel Czarnul. Modeling, run-time optimization and execution of distributed workflow applications in the jee-based beesycluster environment. *The Journal of Supercomputing*, pages 1–26, 2010. 10.1007/s11227-010-0499-7, http://dx.doi.org/10.1007/s11227-010-0499-7.

[5] Pawel Czarnul. Dynamic ranking of cloud providers. In *Proceedings of the 4th International Workshop on Software Services – WoSS 2012*, pages 6–8. Univerza v Ljubljani, 2012. ISBN 978-961-6884-06-8, Eds.: Vlado Stankovski and Dana Petcu.

[6] Pawel Czarnul and Jakub Kurylowicz. Automatic conversion of legacy applications into services in beesycluster. In *Proceedings of 2nd International IEEE Conference on Information Technology ICIT'2010*, pages 21–24, Gdansk, Poland.

[7] G. Juve and E. Deelman. *Grids, Clouds and Virtualization*, chapter Scientific Workflows in the Cloud, pages 71–91. Springer, 2010.

[8] Katarzyna Keahey, Mauricio Tsugawa, Andrea Matsunaga, and Jose Fortes. Sky computing. *IEEE Internet Computing*, 13:43–51, 2009.

[9] S. Pandey, D. Karunamoorthy, and R. Buyya. *Cloud Computing: Principles and Paradigms*, chapter Workflow Engine for Clouds. Wiley Press, New York, USA, 2011. ISBN-13: 978-0470887998.

[10] Suraj Pandey, Dileban Karunamoorthy, and Rajkumar Buyya. *Cloud Computing: Principles and Paradigms*, chapter Workflow Engine for Clouds, pages 321–344. Wiley Press, New York, USA, February 2011. ISBN-13: 978-0470887998.

[11] Chintan Patel, Kaustubh Supekar, and Yugyung Lee. A QoS Oriented Framework for Adaptive Management of Web Service based Workflows. In *Proceedings of the 14th International Database and Expert Systems Applications Conference (DEXA 2003)*, LNCS, pages 826–835, Prague, Czech Republic, September 2003.

[12] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Sheng. Quality driven web services composition. In *Proceedings of WWW 2003*, Budapest, Hungary, May 2003.

[13] Liangzhao Zeng, Boualem Benatallah, Anne H.H. Ngu, Marlon Dumas, Jayant Kalagnanam, and Henry Chang. Qos-aware middleware for web services composition. *IEEE Trans. Softw. Eng.*, 30(5):311–327, 2004.

A Comparison of Hadoop Tools for Analyzing Tabular Data

Ivan Tomašić, Aleksandra Rashkovska, Matjaž Depolli and Roman Trobec
Jožef Stefan Institute, Slovenia
E-mail: ivan.tomasic@ijs.si, aleksandra.rashkovska@ijs.si, matjaz.depolli@ijs.si, roman.trobec@ijs.si

Keywords: Hadoop, MapReduce, Pig, Hive, BigData

Received: December 24, 2012

The paper describes the application of Hadoop modules: MapReduce, Pig and Hive, for processing and analyzing large amounts of tabular data acquired from a computer simulation of heat transfer in bio tissues. The Apache Hadoop is an open source environment for storing and analyzing BigData. It was installed on a cluster of six computing nodes, each with four cores. The implemented MapReduce job pipeline is described and the essential Java code segments are presented. The Java implementation employing MapReduce is compared to the Pig and Hive implementations regarding execution time and programming overhead. The experimental measurements of execution times of the employed parallel MapReduce tasks on 24 processor cores result in a speedup of 20, relative to the sequential execution, which indicates that a high level of parallelism is achieved. Furthermore, our test cases confirm that the direct employment of MapReduce in Java outperforms Pig and Hive by more than two times, while Hive being 20% faster than Pig. Still, Pig and Hive remain suitable and convenient alternatives for efficient operations on large data sets.

Povzetek: Prispevek opisuje uporabo Hadoop programskih modulov: MapReduce, Pig in Hive za procesiranje in analizo tabelaričnih podatkov o prenosu toplote v tkivih.

1 Introduction

Since 2004, when the famous publication "MapReduce: Simplified Data Processing on Large Clusters" [1] was published from the Google's team, the MapReduce paradigm has become one of the most popular tools for processing large datasets, mostly because it allows users to build complex distributed programs using a very simple model.

Apache Hadoop [2] is a highly popular set of open source modules for distributed computing, developed initially to support distribution for the Nutch search engine project. One of the key Hadoop components is the MapReduce on which the other, higher-level Hadoop-related components rely, e.g., Pig and Hive.

With the increasing popularity of the MapReduce and other non-relational data processing approaches, it became apparent that they can be used to construct efficient computing infrastructures. Furthermore, the Hadoop has proved its ability to store and analyze huge datasets often referred to as the BigData [3]. It is used by Yahoo and Facebook for their batch processing needs. Hadoop and Hive are among cornerstones of the storage and analytics infrastructure at Facebook [4]. Facebook Message, in particular, is the first ever user-facing application built on the Apache Hadoop platform [5].

The MapReduce can be seen as a complement to the parallel Relational Database Management System (RDBMS). It is a common opinion these days that the MapReduce is more suitable for batch processing analyzes of whole datasets and for applications where data is written once and read many times, whereas the RDBMS is better for databases that are continuously updated.

In this paper, we investigate the differences in approaches, performances, and usability, between MapReduce, Pig, and Hive Hadoop tools. Their performances were compared in the analysis of tabular simulation data of heat transfer in a biomedical application, in particular, cooling of a human knee after surgery [6]. Similar data sources can be found also in other scientific areas related to multi-parametric simulations [7], environmental data analysis [8], high energy physics [9], bioinformatics [10] etc., whereas some special problems may benefit from specific interfaces to the Hadoop [11].

2 Description of utilized Hadoop modules

Hadoop is composed of four modules:

- Common: support for other Hadoop modules,
- Hadoop Distributed file System (HDFS),
- YARN: a framework for job scheduling and cluster resource management, and
- MapReduce.

There is a number of Hadoop-related projects, but the ones most relevant to our data analyses are Pig and Hive.

2.1 Map/Reduce paradigm

MapReduce is a programming model and an associated implementation for processing and generating large data sets [1]. Some problems that can be simply solved by MapReduce are: distributed grep, count of URL access frequency, various representations of the graph structure of web documents, term-vector per host, inverted index, etc.

A MapReduce program execution consists of the four basic steps: (1) splitting the input, (2) iterating over each split and computing (key, value) pairs (parallel for each split), (3) grouping intermediate values by keys, (4) iterating over values associated with each unique key (in parallel for different keys), computing (usually reducing values for a given key) and outputting final (key, value) pairs.

The first step is done by the MapReduce framework, whereas for the second step a user provides a Map function, which is applied by the framework, commonly on each line of every split. Each Map function invocation outputs a list of (key, value) pairs. Note that each split is generally processed on different processor cores and machines in parallel.

As a simple example, let's consider the task of counting the number of occurrences for each word in a document. The Map function will count the number of occurrences of each word in a line and output a list of (key, value) pairs, for each line:

$$\{(\text{word\_1}, \text{num\_1}_i), (\text{word\_2}, \text{num\_2}_i), \ldots\},$$

where *i* is the line index.

The MapReduce framework groups together all intermediate values associated with the same intermediate key (step 3). The resulting (key, values) pairs are one by one sent to the user-specified Reduce function which aggregates or merges together the values to form a new, possibly smaller, set of values (step 4). In our example the Reduce function will accept each unique word, as a key, and the numbers of their occurrences in each line, as values, sum the numbers of occurrences and output one (key, value) pair per word:

$$(word\_N, \text{sum}\{num\_N_1, num\_N_2, \ldots, num\_N_i, \ldots\}).$$

The executions of the Map and Reduce functions are referred to as Map and Reduce tasks. A set of tasks executed for one application are referred to as a MapReduce job.

The main limitation of the MapReduce paradigm is that each Map and Reduce task must not depend on any data generated in other Map or Reduce tasks of the current job, as user cannot control the order in which the tasks execute. Consequently, the MapReduce is not directly applicable to recursive computations, and algorithms that depend on shared global state, like online learning and Monte Carlo simulations [12].

The MapReduce, as a paradigm, has different implementations. In the presented work, we have used MapReduce implemented in Apache Hadoop distributed in Cloudera [13]. A convenient comparison between MapReduce implementations is presented in [14].

2.2 Apache Hadoop MapReduce implementation

The splitting is introduced because it enables data processing scalability, which shortens the time needed to process the entire input data. The parallel processing can be better load-balanced if the splits are small. However, if the splits are too small, then the time needed to manage the splits and the time for the Map task creation may begin to dominate the total job execution time.

Hadoop splits are fixed-size, whereas a separate Map task is created for each split (Figure 1). The default Hadoop MapReduce split size is the same as the default size of an HDFS block, which is 64 MB. Hadoop performs data locality optimization by running the Map task on the node where the input data resides in the HDFS. With the default HDFS replication factor of three, files are concurrently stored on three nodes; hence, splits of the same file can be concurrently processed on three nodes without the need for being copied before.

In the Hadoop implementation, the Map tasks write their outputs to their local disks, not to the HDFS and are therefore not replicated. If an error happens on a node running a Map task before its output has been consumed by a Reduce task, then the Hadoop resolves the error by re-running the corrupted Map task on another node.

The Map tasks partition their outputs, creating one partition for each Reduce task (Figure 1 – each Map creates *r* output partitions). Each partition may contain

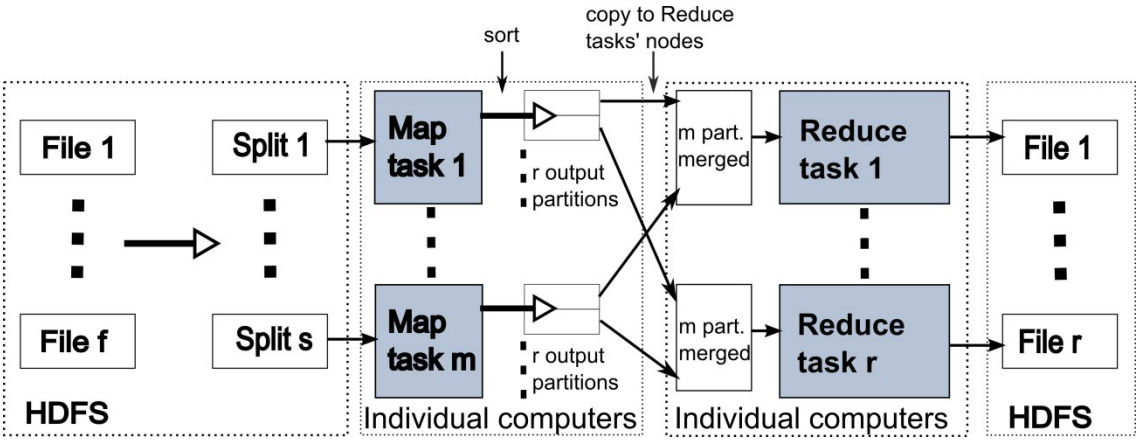


Figure 1: Schematic representation of Hadoop's MapReduce data flow (without using Combiners).

various keys and associated values. All records sharing the same key are processed by the same Reduce task. This is achieved by using the so-called Partitioner function. The default Hadoop MapReduce Partitioner employs a hash function on the keys from the Maps' outputs. Modulo function by the number of reducers is subsequently applied to the hash values resulting in the Reduce task indexes for each key.

The data flow between Map and Reduce tasks is colloquially known as "shuffle". The inputs for each Reduce task are pulled from the machines where the Map tasks ran. The input to a single Reduce task is generally formed from outputs of multiple Map tasks (Figure 1 – each reduce task receives *m* partitions, where *m* is the number of Map tasks); therefore Reduce tasks cannot convey on data locality. On nodes running Reduce tasks, the sorted map outputs are merged before being passed to a Reduce task. The number of Reduce tasks is specified independently for a given job. Each Reduce task outputs a single file, which is usually stored in the HDFS (Figure 1).

Hadoop allows a user to specify an additional so-called Combiner function, which can be executed on each node that runs Map tasks. It receives all the data emitted by the Map tasks on the same node as an input and forms the output that is further processed in the same way as the direct output from a Map task would be. The Combiner function may achieve data reduction on a node level, consequently minimizing data transfer over the network between the machines executing Map and Reduce tasks. The use of Combiner functions reduces the impact of the limited communication bandwidth on the performances of a MapReduce job. The Combiner function code is usually the same as the Reduce function.

2.3 Pig

The development cycle of a MapReduce program may be quite long. Furthermore, it requires an experienced programmer that knows how to describe a given data processing task as a set of MapReduce jobs.

Pig is a sequential language, called Pig Latin, which expresses operation on data, together with execution environment that runs Pig Latin programs [15]. A Pig Latin program comprises a series of high level data operations, translated to the MapReduce jobs that can be executed on a Hadoop cluster. Pig is designed to reduce programming time by providing a higher level procedural utilization of the MapReduce infrastructure. It allows a programmer to concentrate on the data rather than on the details of execution.

Pig runs as a client-side application and has an interactive shell named Grunt used for running Pig Latin programs.

2.4 Hive

A programmer familiar with SQL language may prefer to describe data operations with SQL language, even if the data is not stored in a RDBMS. Hive is Hadoop's data warehouse system that provides mechanism to project structure onto data stored in HDFS or a compatible file system [16]. It provides a SQL-like language called HiveQL. It does not support the full SQL-92 specification, but provides some extensions that are consequences of the MapReduce infrastructure supporting each Hive query. The primary way of interacting with Hive is the Hive shell used to insert and execute HiveQL instructions.

Like RDBMS, Hive stores data in tables. When the tables are loaded with data, Hive stores them in its warehouse directory [17]. Before execution, usually when the select statement is called, Hive, like Pig, transforms the instructions to a set of MapReduce jobs executed on a Hadoop cluster.

The most significant difference between Hive and Pig is that Pig Latin is a procedural programming language, whereas HiveQL is a declarative programming language. A Pig Latin program is a sequential list of operations on an input relation, in which each step is a single transformation. On the other hand, HiveQL is a language based on constraints that, when taken together, define a data operation.

3 Analyzing simulation data

3.1 Description of the Hadoop cluster

The Apache Hadoop open source Cloudera distribution was installed on a cluster built of six computing nodes. The nodes are connected with Gigabit Ethernet. Each node has a quad-core Intel Xeon 5520 processor, 6 GB of RAM and 500 GB hard disk. All nodes run 64-bit Ubuntu Server 12.04 operating system.

One of the nodes is designated as the namenode while others are the datanodes. The namenode also hosts the jobtracker. All machines in the cluster run an instance of a datanode and a tasktracker. For a description of the HDFS and MapReduce nodes please refer to [18, 19].

3.2 Input data

The computer simulation of two hours cooling of a human knee after surgery is performed for 10 different knee sizes, 10 different initial temperature states before cooling, and 10 different temperatures of the cooling pad. This results in 1000 simulation cases. The results of those simulation cases are gathered in 100 files, each for one knee size and one initial state, and for all cooling

CASE	*Parameters*
1	T1
2	T1-T5
3	T1,T6,T11,T16,T21
4	T1-T21
5	T1,T6,T11,T16,T21,T46,T51,T56,T61
6	T1-T21,T46-T61
7	T1,T6,T11,T16,T21,T26,T31,T36,T41,T46,T51, T56,T61,T66,T71,T76,T81
8	T1-T85

Table 1: List of test cases.

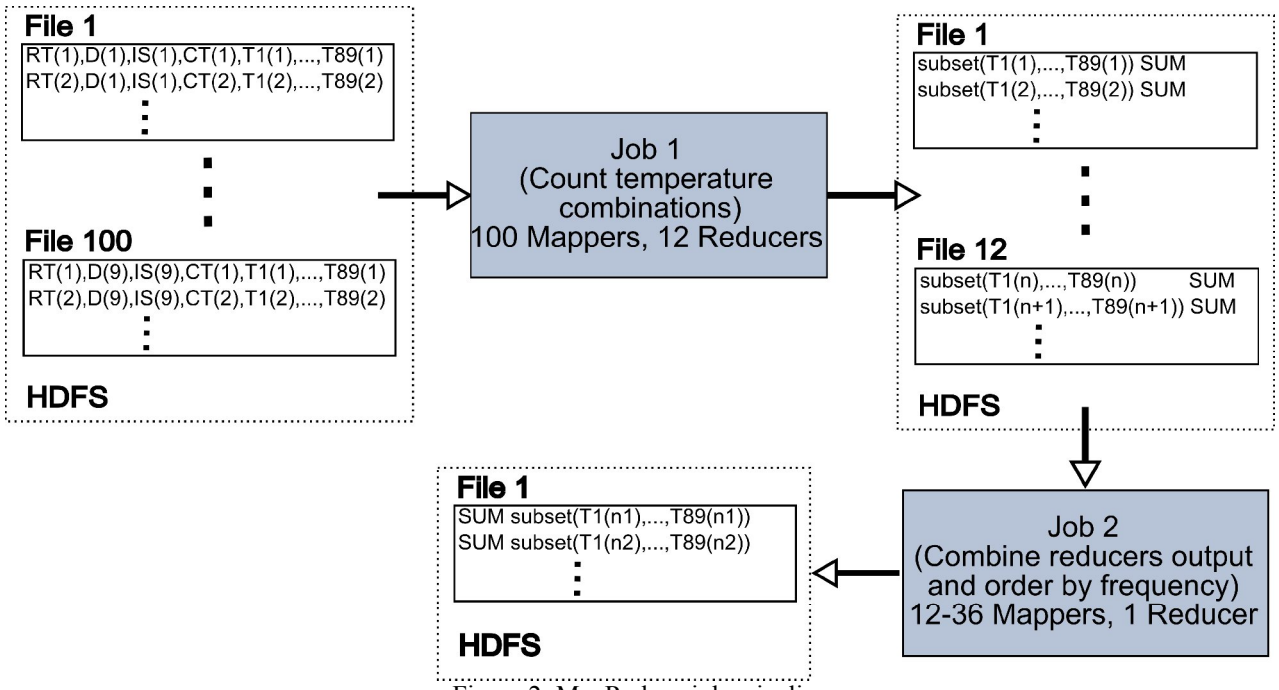


Figure 2: MapReduce jobs pipeline.

temperatures. Each file contains 71970 rows or approximately 44 MB of data. Each data row is composed of the following parameters, i.e., columns: RT, D, IS, CT, T1, T2, … , T85, where are: RT - relative time in a simulation case, D - knee size, IS –initial state, CT – cooling temperature, T1-T85 – inner and outer knee temperatures, i.e., temperatures at a particular location in the knee center, 8 locations on the knee skin and 8 respective locations under the cooling pad, all taken in the current and in previous time steps. In order to assess the periodicities in the knee simulation results, we demand from the MapReduce to count the occurrences of the same value arrays for a subset of knee temperatures T, more precisely, to count the occurrences of identical rows after having projected only columns of T that are of interest. For the SQL code of this operation please refer to the code in Figure 5. We will refer to, in the rest of the paper, the number of occurrences of identical rows as temperature frequencies.

We defined and examined 8 cases with different sets of T. The cases are given in Table 1. Cases with odd numbering take only the current values for the temperatures: Case 1 – the knee center; Case 3 – the knee center and 4 locations on the knee skin; Case 5 – the knee center, 4 locations on the knee skin, and 4 respective locations under the cooling pad; Case 7 – all current temperatures. The cases with even numbering incorporate denoted temperatures T and their value in 4 previous time steps, e.g., in Case 2, T1-T5 represents five temperature values at time steps t_i, t_{i-1}, t_{i-2}, t_{i-3}, t_{i-4}, for each of T from T1-T5, etc.

3.3 MapReduce

The MapReduce jobs pipeline, used for solving our test cases, is illustrated in Figure 2. The sizes of the input files are smaller than the HDFS block size (in our case: 64 MB). Hence, the number of input Map tasks in Job 1 is equal to the number of input files [20] (in our case: 100), i.e., each input file is processed by a different Map task and no additional splitting is performed. Because the number of Reduce tasks is not explicitly set for Job 1, it becomes, by default, equal to the number of task tracker nodes (in our case: 6), multiplied by the value of the *mapred.tasktracker.reduce.tasks.maximum* configuration property [20] (in our case: 2). The output of Job 1 consists therefore of 12 files. Each file contains a unique combination of temperatures and the number of their occurrences. Job 2 combines Reduce tasks' outputs from Job 1 into a single file (in Job 2, the number of Reduce tasks is explicitly set to 1). It also sorts the input columns in the output file by temperature frequencies. The number of Map tasks in Job 2 depends on the test case (Table 1) and varies between 12 for Case 1 and 36 for Case 8 as the amount of data emitted by Job 1 increases with the case number. The details of the jobs implementations are given in Figure 3 and the following text.

In the Map function of Job 1, from each input row, only the relevant columns (see Table 1) are extracted.

For example, in Case 2, only the columns belonging to T1-T5 will be extracted in the *SearchString* variable. Reduce functions sum, i.e., count the number of occurrences of each combination of temperatures (the key) and outputs it as the new value for the current key. Because all the values for the same key are processed by a single Reduce task, it is evident that the output from Job 1 consists of unique combinations of temperatures and the number of their occurrences.

In Job 2, the Map function inverts its (key, value) pairs, making temperature occurrences the keys, and emits them to the Reduce function that outputs the received pairs. The sorting by occurrence is done by the framework as explained in Section 2.2.

```
//Job 1
public void map(LongWritable key,Text value,
OutputCollector<Text,IntWritable> output, Reporter reporter)
throws IOException{
        String line = value.toString();
        String[] lineElements  = line.split(",");
        String SearchString = null
        //depending on a case (Table I) concatenate different
        lineElements in //SearchString
        …
        word.set(SearchString);
        output.collect(word, new IntWritable(1));
}
public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output, Reporter reporter)
throws IOException{
        int sum = 0;
        while (values.hasNext()){
                sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
}
```

```
//Job 2
public void map(LongWritable key, Text value,
OutputCollector<IntWritable,Text> output, Reporter reporter)
throws IOException{
        String line = value.toString();
        //\t is the default delimiter used by a reducer
        String[] lineElements  = line.split("\t");
        output.collect(new
        IntWritable(Integer.parseInt(lineElements[1])),
                                new Text(lineElements[0]));
}
public void reduce(IntWritable key, Iterator<Text> values,
OutputCollector<IntWritable, Text> output, Reporter reporter)
throws IOException{
        //there is only one value
        output.collect(key, values.next());
}
```

Figure 3: Java code segments of Map and Reduce tasks for Job 1 and Job 2.

3.4 Pig

The Pig program that has the same functionality as the MapReduce code described before must be tailored for each specific case. The Pig code for Case 2 is shown in Figure 4.

After having loaded the data files, we group the records by columns with ordinal numbers 4 to 8 corresponding to temperatures T1 to T5 (note that column indexes are zero based). For other cases, the ordinal numbers of columns are as defined in Table 1. Then we count the number of temperatures in each group and afterwards we order the grouped records by the temperature occurrence. At the end, the results are stored in an output file.

```
records = LOAD '/user/path_to_data_files/*'
                        USING PigStorage(',');
grouped_records = GROUP records BY ($4, $5, $6, $7, $8);
count_in_group = FOREACH grouped_records
                GENERATE group,
                COUNT(records) AS count_temp;
count_in_group_ordered = ORDER count_in_group
                        BY count_temp DESC
                        PARALLEL 1;
STORE count_in_group_ordered
                INTO 'path_to_destination folder';
```

Figure 4: The Pig program.

For the execution of the presented Pig program, we use the default settings with an exception: we use the keyword PARALLEL with the ORDER statements to specify that we want only one Reducer task to be executed for the ORDER statement. Hence, a single file is produced as a final result, as in the MapReduce approach. For the three given instructions: GROUP, FOREACH and ORDER, Pig generates three sequential MapReduce jobs named "GROUP BY", "SAMPLER" and "ORDER BY". We use the same names to refer to those generate jobs.

3.5 Hive

The Hive code that has the same functionality as the MapReduce and Pig programs described before is also tailored for each specific case. The Hive code for Case 2 is given in Figure 5.

First, we create the table Temp_Simul and load the simulation data in it. LOAD instruction is just a file system operation in which Hive copies the input files into Hive's warehouse directory. The resulting table *Results_Case_2* is generated for the results of the SELECT statement that evaluates temperature frequencies. The SELECT statement is customized for columns determined by Case 2. For other cases, the columns should be named as defined in Table 1.

When executing the SELECT statement, Hive generates and executes only two MapReduce jobs, in contrast to the Pig that executes three MapReduce jobs. Hive allows a specification of a maximum or a constant number of reducers. We have not specified them; therefore we gave Hive freedom in specifying the

```
CREATE TABLE `Temp_Simul` (`col_0` INT ,
                        `col_1` INT ,
                        `col_2` INT ,
                        `col_3` FLOAT ,
                        ...
                        col_88` FLOAT )
    COMMENT "Results from simulations"
    ROW FORMAT DELIMITED
    FIELDS TERMINATED BY ','
LOAD DATA INPATH 'path_to_data_files/*'
    INTO TABLE Temp_Simul;
CREATE TABLE Results_Case_2 AS
    SELECT col_4, col_5, col_6, col_7, col_8,
    COUNT(1) AS NumOfOccurences
        FROM Temp_Simul
        GROUP BY col_4, col_5, col_6, col_7, col_8
        ORDER BY NumOfOccurences DESC;
```

Figure 5: The Hive program.

Case:								
1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
11159	8933	391	387	298	294	298	294
11097	8860	323	319	228	224	228	224
10945	8778	298	294	227	217	215	211
10924	8351	271	267	221	216	199	181
10729	7807	264	232	220	211	194	168

Table 2: Top 5 temperature frequencies for each case.

number of reducers for each job. Still, the number of Reduce tasks for Job 2 was always equal to one.

4 Results and Discussion

As expected, the three presented approaches gave identical quantitative result. The five highest numbers of temperature frequencies, for each test case from Table 1, are given in Table 2. We have presented only temperature frequencies since the temperature values that are associated with these frequencies are specific to the knee simulation and are not in the scope of this paper. We see that the lowest numbers appear in Case 8, which was expected because in Case 8 the largest number of parameters (T) is projected from the source data.

	Job1									***Job2***								
Case:	1	2	3	4	5	6	7	8	**Total**	1	2	3	4	5	6	7	8	**Total**
No. of Map tasks	100	100	100	100	100	100	100	100		12	12	12	18	16	20	26	36	
No. of Reduce tasks	12	12	12	12	12	12	12	12		1	1	1	1	1	1	1	1	
Tot. time maps (s)	1122	1080	1119	1187	1121	1287	1162	1826	**9903**	32	31	51	78	59	184	64	443	**941**
Tot. time red. (s)	100	80	91	148	108	207	118	413	**1264**	4	4	10	16	12	31	12	50	**139**
CPU time spent (s)	588	618	667	790	686	933	719	1,494	6494	7	9	55	95	70	185	73	330	823
Total duration (s)	40	37	38	43	49	51	40	79	**377**	13	14	22	28	26	48	24	73	**248**

Table 3: MapReduce approach: MapReduce tasks execution times.

	Job1 (GROUP BY)								
Case:	1	2	3	4	5	6	7	8	**Total**
No. of Map tasks	34	34	34	34	34	34	34	34	
No. of Reduce tasks	5	5	5	5	5	5	5	5	
Total time spent by all maps in (s)	517	543	527	807	540	874	723	1077	**5608**
Total time spent by all reduces (s)	31	24	45	180	58	297	109	597	**1342**
CPU time spent (s)	371	394	446	695	502	1098	582	1734	5822
Total duration (s)	31	34	38	76	40	100	59	330	**708**

	Job2 (SAMPLER)									***Job3 (ORDER BY)***								
Case:	1	2	3	4	5	6	7	8	**Total**	1	2	3	4	5	6	7	8	**Total**
No. of Map tasks	1	1	1	10	3	18	5	35		1	1	1	10	3	18	5	35	
No. of Reduce tasks	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	
Tot. time maps (s)	7	6	9	52	17	105	26	315	**537**	7	7	18	108	35	215	56	536	**983**
Tot. time red. (s)	4	4	4	4	4	6	4	4	**32**	4	4	15	64	25	124	42	202	**481**
CPU time spent (s)	2	2	5	35	11	70	17	139	282	2	5	31	176	62	346	100	669	1391
Total duration (s)	13	15	17	17	18	19	18	23	**140**	13	15	36	82	43	145	62	226	**622**

Table 4: Pig approach: MapReduce tasks execution times.

	Job1									***Job2***								
Case:	1	2	3	4	5	6	7	8	**Total**	1	2	3	4	5	6	7	8	**Total**
No. of Map tasks	17	17	17	17	17	17	17	17		2	2	2	3	2	4	3	10	
No. of Reduce tasks	5	5	5	5	5	5	5	5		1	1	1	1	1	1	1	1	
Tot. time maps (s)	119	153	166	254	193	364	224	652	**2125**	8	9	24	66	32	103	44	186	**473**
Tot. time red. (s)	17	22	31	82	42	136	52	810	**1192**	3	3	15	77	26	146	41	265	**576**
CPU time spent (s)	113	138	202	389	237	580	291	991	2941	4	7	39	156	63	272	92	501	1134
Total duration (s)	19	23	27	39	31	58	35	215	**447**	12	14	32	112	49	182	68	293	**762**

Table 5: Hive approach: MapReduce tasks execution times.

Table 3 shows the MapReduce execution times for Job 1 and Job 2, for each test case. It also shows the total CPU time and associated total duration of the analysis for each case. Table 4 and Table 5 show corresponding execution times of the MapReduce tasks generated by Pig and Hive, respectively.

4.1 Execution time

Although the interpretation of the temperature values and their occurrence in a specified combination are not important for this paper, each execution case (Table 1) draws different amounts of data to the Map and Reduce functions in Job 1 and Job 2, which influences their execution times, as evident from Table 3.

We can calculate from Table 3 that the total time spent for Map and Reduce in Job 1 and Job 2, for all test cases on all executing nodes, is: $t_s = 9903 + 1264 + 941 + 139 = 12247$ s, while the total duration of the complete MapReduce analysis is: $t_m = 377 + 248 = 625$ s. The ratio t_s/t_m, which can assess the level of parallelism achieved, is 19.6. Consequently, we can conclude that the above analysis is about 20 times faster, if implemented by the MapReduce paradigm on 24 computing cores, relatively to the MapReduce execution time on a single core.

Table 4 and Table 5 show execution times of the MapReduce tasks generated by the Pig and Hive. The job durations, for all test cases and for all three approaches, are shown in Figure 6. The last triple presents the total execution times across all test cases. It is evident that the MapReduce tasks, written and executed directly, take, in average, approximately two times less time than those generated by Pig or Hive. Furthermore, Hive outperforms Pig for approximately 20%.

One can also notice that the Hive approach was faster than the direct MapReduce approach in the first three cases. By comparing Tables 3 and 5, it is evident that Hive gained the advantage in Job 1. This possibly happened because the number of Map and Reduce tasks, i.e., 17 and 5, applied by Hive, were more appropriate for the smaller amount of data. The superiority of the direct MapReduce approach is however more and more evident as the case number, therefore also the amount of data, increases.

5 Conclusion

In this paper, we have applied Hadoop tools for the analyses of tabular data coming from a complex computer simulation. Three approaches were applied; the first modeled the data operations directly with the MapReduce jobs, while the other two described the data operations using higher level languages Pig and Hive.

All three approaches gave the same quantitative result, but the execution times were different. From the presented time measurements it is evident that the directly programmed MapReduce tasks are in average two times faster than Pig or Hive. For our test cases, it is also evident that Hive outperforms Pig for 20%, probably because Hive generates one MapReduce job less than Pig. Hive outperformed the direct MapReduce approach in the cases with smaller amounts of data, probably because the number of Map and Reduced tasks employed by Hive was more optimal for smaller data sets.

As Pig and Hive use MapReduce in the background, it is expected that using the low-level approach will be faster. However, the high-level approaches could have advantages over the direct MapReduce approach if design efforts are also considered. Writing the mappers and reducers, compiling, debugging, packaging the code, submitting the jobs, and retrieving the results using the direct MapReduce approach takes developer's time. On

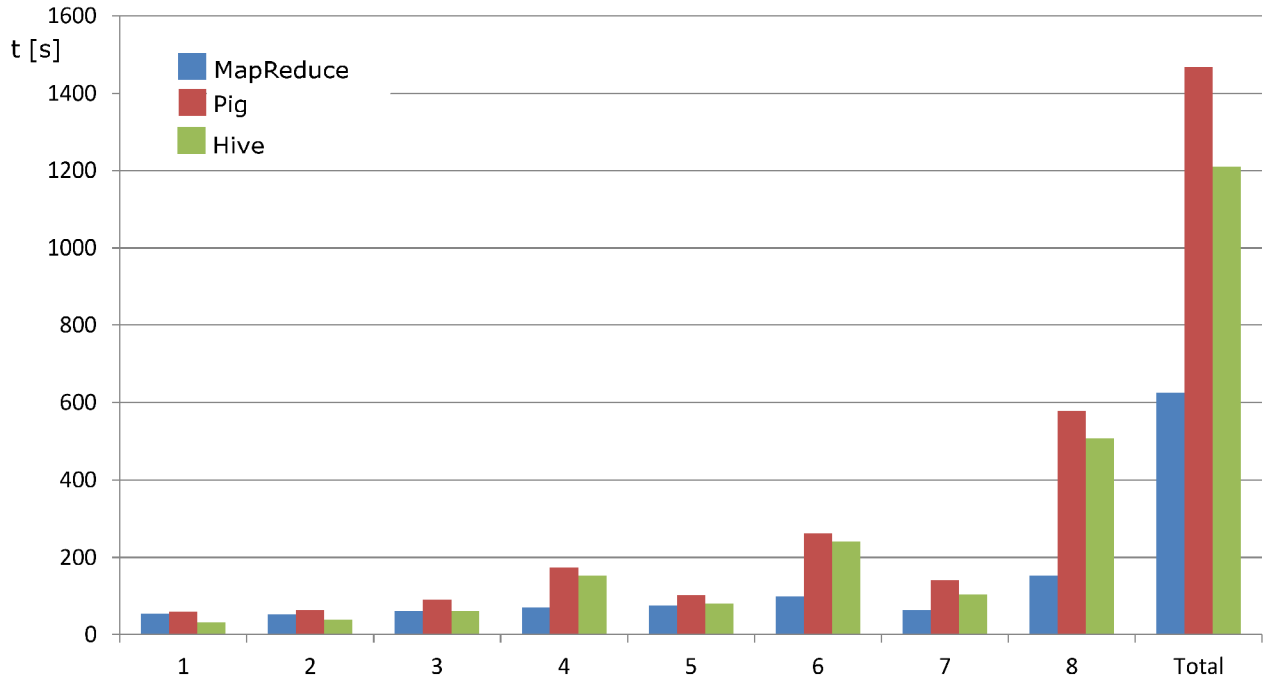


Figure 6: Sum of jobs execution times for each test case. The last series represents the total sum

the other hand, it is much easier to describe data operations with Pig or Hive for an user less familiar with the Java programing language. Users familiar with SQL language may prefer to use Hive, while users familiar with procedural languages would probably prefer to use Pig to describe the same data operations.

Future work is in implementing MapReduce paradigm with the MPI library [21] that could support more complex communication functions, which could result in more efficient execution of the computationally intensive services, on complex data sets in cloud environments [22].

Acknowledgement

The research was funded in part by the European Union, European Social Fund, Operational Programme for Human Resources, Development for the Period 2007-2013.

References

[1] J. Dean, and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in OSDI'04, 2004, pp. 137-149.

[2] "Welcome to Apache™ Hadoop®!," Oct., 2012; http://hadoop.apache.org/.

[3] B. Franks, "What is big data and why does it matter?," *Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics*, pp. 3-29, Hoboken, New Jersey: John Wiley & Sons, Inc., 2010.

[4] A. Thusoo, Z. Shao, S. Anthony *et al.*, “Data warehousing and analytics infrastructure at facebook,” in SIGMOD 2010, International conference on Management of data, pp. 1013-1020.

[5] D. Borthakur, J. Gray, J. S. Sarma *et al.*, “Apache hadoop goes realtime at Facebook,” in ACM SIGMOD International Conference on Management of Data, 2011, pp. 1071-1080.

[6] R. Trobec, M. Šterk, S. Almawed *et al.*, “Computer simulation of topical knee cooling,” *Comput. biol. med,* vol. 38, pp. 1076-1083, 2008.

[7] G. Kosec, Šarler, Božidar, “Solution of a low Prandtl number natural convection benchmark by a local meshless method.,” *International journal of numerical methods for heat & fluid flow,* vol. 23, no. 1, pp. 189-204, 2013.

[8] U. Stepišnik, and G. Kosec, “Modelling of slope processes on karst,” *Acta Carsologica,* vol. 40, no. 2, pp. 267-273, 2011.

[9] L. Wang, J. Tao, R. Ranjan *et al.*, “G-Hadoop: MapReduce across distributed data centers for data-intensive computing,” *Future Generation Computer Systems,* vol. 29, no. 3, pp. 739-750, 2013.

[10] R. C. Taylor, “An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics,” *BMC Bioinformatics,* vol. 11, no. SUPPL. 12, 2010.

[11] M. Niemenmaa, A. Kallio, A. Schumacher *et al.*, “Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud,” *Bioinformatics,* vol. 28, no. 6, pp. 876-877, 2012.

[12] J. Lin, and C. Dyer, "Limitations of MapReduce," *Data-Intensive Text Processing with MapReduce*, Synthesis Lectures on Human Language Technologies, pp. 143-145: Morgan & Claypool Publishers, 2010.

[13] I. Cloudera. "CDH Proven, enterprise-ready Hadoop distribution – 100% open source," Oct, 2012; http://www.cloudera.com/hadoop/.

[14] Z. Fadika, E. Dede, M. Govindaraju *et al.*, "Benchmarking MapReduce Implementations for Application Usage Scenarios." 12th IEEE/ACM International Conference on Grid Computing (GRID). pp. 90-97, 2011.

[15] "Welcome to Apache Pig!," December, 2012; http://pig.apache.org/.

[16] "Welcome to Hive!," December, 2012; http://hive.apache.org/.

[17] T. White, "Hive," *Hadoop: The Definitive Guide*, pp. 365-409, Gravenstein Highway North, Sebastopol: O’Reilly Media, Inc., 2010.

[18] T. White, "The Hadoop Distributed Filesystem," *Hadoop: The Definitive Guide*, pp. 41-73, Gravenstein Highway North, Sebastopol: O’Reilly Media, Inc., 2010.

[19] T. White, "MapReduce," *Hadoop: The Definitive Guide*, pp. 15-40, Gravenstein Highway North, Sebastopol: O’Reilly Media, Inc., 2010.

[20] T. White, "MapReduce Types and Formats," *Hadoop: The Definitive Guide*, pp. 189-224, Gravenstein Highway North, Sebastopol: O’Reilly Media, Inc., 2010.

[21] T. Hoefler, A. Lumsdaine, and J. Dongarra, "Towards efficient mapreduce using MPI," 16th European Parallel Virtual Machine and Message Passing Interface Users' Group Meeting, EuroPVM/MPI, 2009, pp. 240-249.

[22] H. Mohamed, and S. Marchand-Maillet, "Distributed media indexing based on MPI and MapReduce," *2012 10th International Workshop on Content-Based Multimedia Indexing, CBMI 2012*. pp. 236-241.

QoS Prediction for Web Services Based on Similarity-Aware Slope One Collaborative Filtering

Chengying Mao and Jifu Chen
School of Software and Communication Engineering,
Jiangxi University of Finance and Economics, 330013 Nanchang, China
E-mail: maochy@yeah.net

Keywords: Web services, QoS prediction, Slope One, similarity, collaborative filtering

Received: December 15, 2012

Web services have become the primary source for constructing software system over Internet. The quality of whole system greatly dependents on the QoS of single Web service, so QoS information is an important indicator for service selection. In reality, QoSs of some Web services may be unavailable for users. How to predicate the missing QoS value of Web service through fully using the existing information is a difficult problem. This paper attempts to settle this difficulty through combining Pearson similarity and Slope One method together for QoS prediction. In the paper, we adopt the Pearson similarity between two services as the weight of their deviation. Meanwhile, some strategies like weight adjustment and SPC-based smoothing are also utilized for reducing prediction error. In order to evaluate the validity of our algorithm (i.e., similarity-aware Slope One algorithm, SASO), comparative experiments are performed on the real-world data set. The results show that SASO algorithm exhibits better prediction precision than both basic Slope One and the well-known WsRec algorithm in most cases. Meanwhile, our approach has the strong ability of reducing the impact of noise data.

Povzetek: Članek poskuša razrešiti problem ocenjevanja kakovosti storitve s kombiniranjem Parsonove podobnosti in metode Slope One.

1 Introduction

In recent years, the pattern of service-oriented computing (SOC) has been widely accepted to build large-scale system over Internet [1]. In this new style of software development paradigm, software is no longer built via the traditional process, but in the way of service unit reuse. Accordingly, some new problems such as service discovery, selection and composition are emerging, and play a great impact on the quality of service-based system.

In general, service unit is self-describing component to complete a specific task. Quality-of-Service (QoS) is an important way to describe non-functional characteristics of Web services. When several functionally-equivalent Web services exist in the network, QoS is viewed as a critical issue for picking out the appropriate service from equivalent service set. Web service QoS usually includes a number of properties, such as response time, throughput, failure probability, availability, price, popularity, and so on [2]. Due to different network environments, service users will have different QoS metrics for the same Web service. Therefore, each service user has to understand QoSs of all services to be invoked at his/her end.

In order to construct the software meeting the actual requirements, it needs to make the existing service units work together in accordance with the pre-defined business logic, that is the so-called Web service composition (WSC). During service selection, the quality of each service unit should be carefully considered so as to ensure the trustworthiness of WSC. However, service invoker may be lack of adequate historical information for some specific Web services. He/She has to estimate the QoS value of a given Web service before determining to introduce it into WSC, i.e., QoS prediction for Web services. Since the service user has not even invoked the service in past, the estimation for such service's QoS has to get help from other similar users or self's invocation records on other Web services.

The similar work firstly emerged in the field of E-commerce, vendors used consumer's historical purchase records and the similarity between costumers to recommend products [3]. In contrast, the prediction of Web service's QoS is much harder than product recommendation. Web service is merely an encapsulated and distributed Web API over network. Therefore, for service users, the information related with service execution are hardly collected. In order to improve the prediction precision, the limited available Web services invocation records should be fully utilized. As far as we known, study in [4] is the first work of predicting Web service's QoS through collaborative filtering (CF). Shao *et al.*'s work mainly considered the similarity among user's experiences on Web services, and proposed a service users' similarity-based prediction method, in which the similarity is measured by Pearson correlation coefficient. Subsequently, Zheng *et al.* [5] presented a

more comprehensive method for QoS prediction, in which they combined the traditional user-based and item-based collaborative filtering methods together through confidence weights. Recently, some improved methods based on personalized context [6, 7] or hierarchical and side information [8] are also proposed.

It is important to note that, most above mentioned methods are in accordance with Pearson-based similarity. Although this kind of similarity can provide good prediction effect, it not only cost much computation time but also lose performance for the very sparse data set. Besides the similarity-based collaborative filtering, Slope One [9] has been validated as an effective prediction method due to its simpleness and high performance. In the paper, we presented a hybrid QoS prediction method through introducing Pearson-based similarity into Slope One method. The experimental results revealed that our hybrid method (named *similarity-aware Slope One*, SASO) could outperform the basic Slope One and Pearson-based collaborative filtering methods in term of prediction precision.

The main contributions of this paper can be addressed as follows.

(1) A prediction algorithm of Slope One co-operated with Pearson similarity measurement has been proposed for providing QoS information for Web service user.

(2) Some strategies like weight adjustment and SPC-based smoothing are presented for improving the prediction precision.

(3) The detailed performance analysis on real-world data set is performed to verify the effectiveness of our method. Moreover, the two-stage filling strategy is also validated through experimental analysis.

The structure of the paper is as follows. In the next section, we state the QoS prediction problem for Web services, and introduce two typical collaborative filtering algorithms. In section 3, the overall QoS prediction framework is firstly addressed, and then the similarity-aware Slope One algorithm is described in details. The performance comparison and analysis are discussed in section 4. Section 5 gives some existing researches that are closely related with our prediction approach. Finally, section 6 concludes the paper.

2 Background

2.1 QoS prediction for Web services

When Web service users prepare to adopt some service units to construct an enterprise-level application, in general, they have to replace each abstract service in service orchestration plan with a concrete service. For each abstract service, perhaps quite a few service implementations will meet the requirement of its function. Therefore, the rational way is to pick out a service with high QoS from the candidate set. However, for a specific service user, the QoS values of some Web services may be not available. As a consequence, it is necessary to estimate the QoSs of such services according to the limited existing information, that is so-called QoS prediction problem.

Motivating Example. Here, we provide a simple illustration to address the QoS prediction for Web services. As shown in Table 1, there are response time (i.e. RT) records of three Web services w.r.t five users. The element $r_{i,j}$ means the RT value of user i for service j, and "NA" represents the corresponding value not available at present. Assume user $u3$ has some interests on the third service, since there is no ready record in the table, he has to predicate the issue $r_{3,3}$ according to his own and others' service invocation records.

User	Response time (second)		
	*service*1	*service*2	*service*3
*u*1	0.4	1.6	NA
*u*2	0.9	NA	1.9
*u*3	2.8	3.5	**??**
*u*4	NA	3.0	4.0
*u*5	0.8	NA	0.9

Table 1: An motivated example for illustrating QoS prediction problem.

How to estimate the missing value? Besides $u3$'s existing records on other two services (i.e. $r_{3,1}$ and $r_{3,2}$), the available service invocation records of other four users also should be taken into consideration. With regard to prediction techniques, experiences tell us that *collaborative filtering* (CF) techniques can be viewed as a good choice.

2.2 Review on collaborative filtering

In general, collaborative filtering is a technique of suggesting particularly interesting items or patterns based on past evaluations of a large group of users. The fundamental assumption of CF is that if users have similar tastes on some items, and hence they will rate or act on other items similarly. At present, CF techniques can be classified into three categories [10, 11]: (1) memory-based methods, (2) model-based methods, and (3) hybrid methods. Memory-based CF utilizes the user rating data to calculate the similarity or weight between users or items, and then make predictions according to those similarity values. This type of CF is the earlier mechanism and used in many commercial systems such as Amazon, Barnes and Noble. According to the background and feature of QoS prediction problem, memory-based CF is treated as the main research issue in the paper. Especially, two well-known methods, i.e., Pearson correlation CF and Slope One approach, are taken into consideration.

2.2.1 Pearson correlation-based method

In a typical CF scenario, there is a list of m users $\{u_1, u_2, \cdots, u_m\}$ and a list of n items $\{i_1, i_2, \cdots, i_n\}$, and

each user u_i has a list of items (i.e., Iu_i), which the user has rated, or about which their preferences have been inferred through their behaviors [10]. Generally speaking, the basic procedure of CF-based recommendation or prediction can be summarized as the following two steps:

(1) Look for users sharing the similar interests or rating patterns with a given user (called active user).

(2) Use the information from those like-minded users found in step (1) to calculate a prediction for the active user.

Here, we mainly address the case from the perspective of users, but the above process is also suitable for item-oriented analysis. It is not hard to find that, how to find the similar users (or items) for a specific user (or item) is a critical task in the whole process of CF. In practice, the common interests or patterns are expressed via the correlation between users (or items).

At present, *Pearson correlation coefficient* has been introduced for computing similarity between users or items according to the user-item data like in Table 1, which is usually called *user-item matrix*. For two given users a and u, their similarity can be computed as follows.

$$Sim(a,u) = \frac{\sum_{i\in I}(r_{a,i}-\bar{r}_a)(r_{u,i}-\bar{r}_u)}{\sqrt{\sum_{i\in I}(r_{a,i}-\bar{r}_a)^2}\sqrt{\sum_{i\in I}(r_{u,i}-\bar{r}_u)^2}} \quad (1)$$

where $I = I_a \bigcap I_u$ is the subset of items which both user a and u have invoked previously, $r_{a,i}$ is a vector of item i observed (or rated) by user a, and $\bar{r}_a$ and $\bar{r}_u$ represent average values of different items observed (or rated) by user a and u, respectively.

The prediction method based on two users' similarity is referred as *user-based CF*. Similarly, CF can also be conducted through the similarity computation between two items, that is, *item-based CF*. According to the studies from other researchers, item-based CF can outperform user-based CF in most conditions, and has been treated as a preferred choice for prediction or recommendation problems.

As mentioned earlier, Shao *et al*. firstly adopted Pearson correlation-based CF for Web services' QoS prediction [4]. Recently, Zheng *et al*. improved prediction precision problem through combining item-based and user-based CF together [5]. Their WsRec algorithm exhibits better performance than other basic prediction methods, and has caused much attention in these two years.

2.2.2 Slope One method

Although previous studies have revealed that Pearson scheme CF can gain good prediction precision, its performance is not so satisfactory for the case of extremely sparse data. Meanwhile, Pearson-based method will cost a lot of computational overhead to measure the similarity between users or items. Fortunately, another well-known method called Slope One [9] can make up such deficiencies. On the one hand, Slope One can show good prediction effect for sparse data. On the other hand, this method can perform prediction activity with less computing cost.

As stated by Lemire *et al*., Slope One algorithm works on the intuitive principle of a "popularity differential"between items for users. In this algorithm, how much better one item is liked than another is determined in a pairwise fashion. Firstly, the difference between the averages of two items can be calculated via subtract operation. Then, once one item's value is available, the other's value can be predicted according to such difference. The process can be illustrated in Figure 1. For two users (a and b) and two items (i and j) in user-item matrix, the values of these two items for user a are known and the differential from i to j is 1.5-1=0.5. Thus, the item j's value for user b can be predicted via this mapping relationship, that is, 2+(1.5-1)=2.5. Of course, many such differentials exist in a training set for each unknown rating, the average of these differentials will be taken for predication.

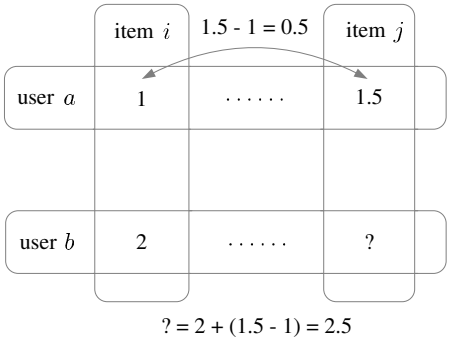


Figure 1: Illustration for Slope One prediction algorithm.

Formally speaking, for a given user-item matrix, the set of the users who contain rating records both on item i and item j can be computed and denoted as $U_{i,j}$ here. Obviously, $U_{i,j} = U_{j,i}$. Then, the average deviation of item i with respect to item j can be denoted as:

$$\text{dev}_{j,i} = \sum_{u\in U_{j,i}} \frac{r_{u,j}-r_{u,i}}{card(U_{j,i})} \quad (2)$$

where $card(U_{j,i})$ returns the element number of set $U_{j,i}$.

Based on the deviations of items, the rating of user u for item j, i.e. $r_{u,j}$, can be predicated via the following way.

$$P(r_{u,j}) = \frac{1}{card(R_j)} \sum_{i\in R_j} (\text{dev}_{j,i} + r_{u,i}) \quad (3)$$

where R_j=$\{i|r_{u,i} \neq \text{NA}, i \neq j \text{ and } card(U_{j,i}) > 0\}$ is the set of items which have co-occurrence relationship with item j.

The above discussion belongs to user-oriented prediction. Obviously, Slope One method can also be used in the other style, i.e., item-oriented prediction. In addition, several kinds of extensions are proposed. For instance, single or bivariate regression is used for finding the best mapping

relation [12, 13], bi-polar strategy is used for users' two different attitudes [9]. However, variant algorithms can't lead to obvious improvements over the basic form in all cases.

3 Similarity-Aware Slope One for QoS prediction

With regard to the usage scenario of Web services, services' QoS data from different users can form a sparse matrix of service invocation records. In order to help service user make a rational decision about service selection, the prediction for a specific service's QoS w.r.t. of the current user is very necessary. In this paper, we provide a hybrid prediction method through comprehensively adopt the merits both from Pearson correlation-based algorithm and Slope One algorithm.

3.1 The overall prediction framework

For an active service user u, the number of services which have been invoked by u is named *given number* (i.e. GN). For all n service items, GN is usually a little part. In order to provide precise QoS estimations for the remaining service items w.r.t user u, we should take full use of other users' invocation records for these services. Here, we assume the historical QoS data about m users for n service items is matrix $\mathcal{M}$. Similarly, each service user only has partial QoS information in that matrix. The proportion of existing QoS data in matrix is denoted as *density* (d for short).

In our investigations on collaborative filtering techniques, we have found a fact as follows: Slope One method is suitable for the very sparse data set (i.e. very low density data), whereas Pearson-based CF can achieve desired prediction results for the case of high density data. Therefore, in our method, we mainly adopt Slope One method for prediction and compute Pearson correlation between services to adjust the reference weight. The closer relation between a service and the subject service for user u, the higher weight should be assigned to the QoS deviation between these two services.

The whole procedure of Web service QoS prediction is shown in Figure 2. At the initial stage, the historical QoS records of n Web services for m users can be collected. Here, we call it training data $\mathcal{M}$. In general, a service user could not have QoS records for all n services, and usually has only very limited ones of them. As a result, training data is a sparse matrix in real-world scenarios. The matrix $\mathcal{M}$ should be filled as full as possible so that it can provide more useful information for QoS prediction. In the second step, we present a similarity-aware Slope One algorithm (SASO for short) as a way to fill the 'NA' (a.k.a. *null*) records in the training data set. For the perspective of Web service execution, there maybe exist some abnormal QoS records in the above training data, especially for the QoS attribute with wide scale values. In order to handle this problem, in the third step, we adopt *statistical process control* (SPC) strategy to adjust such exception data.

Based on the above treatments, the training data set has been enhanced and its data density has a great promotion. According to the renewed training matrix, SASO algorithm is also utilized for predicting Web service's QoS for active user. Finally, prediction quality is measured via error analysis.

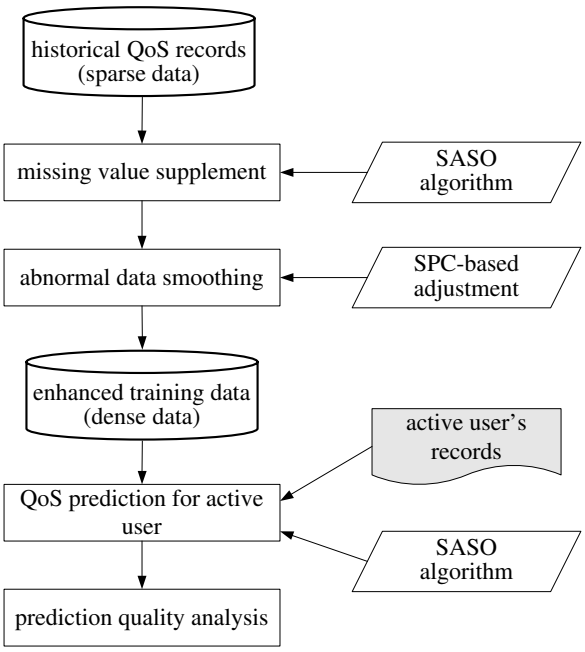


Figure 2: The overall framework of Web service's QoS prediction based on similarity-aware Slope One (SASO).

3.2 Prediction method

With regard to QoS prediction framework, it is not hard to find that SASO algorithm and SPC-based adjustment strategy play important roles for improving the precision. The details of these two key algorithms are addressed as follows.

3.2.1 SASO algorithm

As mentioned before, Slope One-based CF exhibits its advantage for sparse data. Since each active user has only GN (usually $GN << n$) QoS records for n Web services, we adopt item-oriented Slope One method to predict QoS value for active user. However, the similarity between items is not taken into consideration in the basic Slope One prediction method. In our work, we introduce the similarity between two items into Slope One method to form a new QoS prediction algorithm for Web services. The basic idea is that, the service with the higher similarity should give the higher priority when considering the deviation in Slope One method.

Here, we adopt item-based Pearson correlation to measure the similarity between two Web services. For service

i and j, their's similarity can be calculated as follows.

$$Sim(i,j) = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}} \quad (4)$$

where $U = U_i \bigcap U_j$ is the subset of users who have QoS records both on service i and service j previously (i.e., identical to $U_{i,j}$ in equation (2)), and $\bar{r}_i$ represents the average QoS value of service i observed by different users.

To predict a missing value $r_{u,j}$ in the user-item matrix, we have to measure the similarities between j and other services invoked by user u, that is, $I_u - \{j\}$. After removing the services with negative similarity to service j from them, the remaining items are called the related services of j w.r.t. user u, denoted as $R(j|u)$. It can be formally expressed as follows.

$$R(j|u) = \{i | i \in I_u,\ Sim(i,j) > 0,\ i \neq j\} \quad (5)$$

Then, we give the prediction formula based on similarity-aware Slope One algorithm as below.

$$P(r_{u,j}) = \frac{1}{card(R(j|u))} \sum_{i \in R(j|u)} (w_{i,j} \cdot \text{dev}_{j,i} + r_{u,i}) \quad (6)$$

where $w_{i,j}$ is an adjustment weight in accordance with the similarity between j and another service i ($i \in R(j|u)$).

As for a further comment, $w_{i,j}$ can be computed by the following formula:

$$w_{i,j} = \frac{Sim^{\lambda}(i,j)}{\sum_{k \in R(j|u)} Sim^{\lambda}(k,j)} \quad (7)$$

where λ (=1, 2 or 3) is a factor of adjustment strength, higher value means stronger adjustment. Meanwhile, $Sim^{\lambda}(k,j)$ is λ power of $Sim(k,j)$, i.e. $[Sim(k,j)]^{\lambda}$.

It should be noted that, both two steps of missing value supplement and the final QoS prediction for active user adopt SASO algorithm (cf. equation (6)) to provide prediction value.

3.2.2 SPC-based smoothing strategy

Here, we denote the original service's QoS record matrix as $\mathcal{M}$, and call the intermediate matrix after filling the missing values as $\mathcal{M}'$. On the one hand, some exceptional QoS records for Web services perhaps exist in $\mathcal{M}$. The so-called exceptional (or abnormal) record, means the QoS value of a specific user is far away from the records of neighbor users. On the other hand, it is also very sparse in the original state. As a result, the filled matrix $\mathcal{M}'$ maybe contain some QoS items which are far from the common situation. Obviously, these abnormal records will cause bad influence on the next stage of prediction. Thus, we should identify them out from matrix $\mathcal{M}'$ firstly, and then smooth them via a heuristic strategy.

In the paper, we borrow the idea from *statistical process control* (SPC) [14] to tackle the abnormal QoS data in $\mathcal{M}'$. SPC is a realtime monitoring technique for the process of industrial production in the way of statistical analysis. It can scientifically distinguish the exceptional fluctuation from the normal random fluctuation, so it is used for providing early warning for production process to manager. We mainly utilize this technique to pick out the abnormal QoS values so as to achieve better prediction performance.

At first, for matrix $\mathcal{M}'$, we judge whether item $r_{u,i}$ (i.e., the QoS of service i for user u) is an exception or not according to the following rule.

$$isAbn(r_{u,i}) = \begin{cases} true, & \mu_i - \theta \cdot \sigma_i < r_{u,i} \\ & < \mu_i + \theta \cdot \sigma_i \\ false, & \text{otherwise} \end{cases} \quad (8)$$

where μ_i is the average QoS value of service i ($1 \leq i \leq n$), and σ_i is the standard deviation of service i's QoS records from different users. θ is a positive integer used for regulating the normal range of QoS value. It is usually set to 3 in most applications of SPC.

When a suspected record of abnormal QoS is detected through the above approach, this isolated item should be smoothed before the prediction step. Here, we introduce a strategy called "small amplitude shift"for smoothing treatment. Suppose $r_{u,i}$ is an abnormal issue according to judgement of equation (8), the smoothing action can be performed via the following formula. The value after adjustment is denoted as $\tilde{r}_{u,i}$.

$$\tilde{r}_{u,i} = \begin{cases} \mu_i - \theta \cdot \sigma_i, & r_{u,i} < \mu_i - \theta \cdot \sigma_i \\ \mu_i + \theta \cdot \sigma_i, & r_{u,i} > \mu_i + \theta \cdot \sigma_i \\ r_{u,i}, & \text{otherwise} \end{cases} \quad (9)$$

That is to say, we use the upper (or lower) limit to replace the unusually high (or low) QoS record, respectively.

3.3 Computational complexity analysis

As shown in Figure 2, our algorithm mainly includes three linear steps as below.

(1) *Complexity of missing value supplement*. Obviously, the computational complexity for computing the similarity $Sim(i,j)$ between two services (i.e. i and j) is $O(m)$. Then, the complexity of computing similarities of all service pairs is $O(mn^2)$. At the same time, the computational complexity for calculating the deviation of each service pair is also $O(mn^2)$. Based on the above interim results, the complexity of providing the supplement value for each missing item is $O(dn)$ (here, d stands for *density*). Accordingly, $O(d(1-d)mn^2)$ is the complexity in respect to fill all missing items. Thus, the complexity of this step is $O(mn^2)$.

(2) *Complexity of abnormal data smoothing*. The complexity of computing the mean value of QoS is $O(m)$ for each Web service, so $O(mn)$ is for all n services. Meanwhile, the complexity of smoothing action for all items in matrix $\mathcal{M}$ is also $O(mn)$. Therefore, the complexity for smoothing the exceptional data is $O(mn)$.

(3) *Complexity of QoS prediction*. In the third step, each active user has $n - GN$ missing values to be predicated.

The complexity of computing the similarity and deviation between current Web service with all known GN services is $O(m \cdot GN)$. Therefore, the complexity of predicting all $n - GN$ missing values is $O(m \cdot GN(n - GN))$.

Altogether, the computational complexity of our approach for an active user is $O(mn^2)$. In literature [5], the complexities of five steps have been discussed in detail. Based on the comprehensive analysis on the above complexities, the overall computational complexity of WSRec algorithm is $O(m^2n + mn^2)$. As a result, our method and WSRec haven't obvious distinction from the perspective of computation time.

4 Implementation and Experiments

4.1 Experimental setup

In order to validate the effectiveness of our proposed algorithm for QoS prediction, some experiments are employed on a public published data set[1], which is collected by Zheng *et al.* [5] and has been widely adopted in the current researches [7, 15]. The original data set contains 5825 service invocation records from 339 users, and QoS attributes include *response time* (RT) and *throughput* (TP).

In our experiments, we select partial QoS records related with 100 Web services and 150 users from the original data. Then, this data is randomly divided into two parts: training data and test data. Here, 100 users are selected as training users, that is, the data about them is treated as training data. The remaining 50 users are viewed as test users (i.e., the *active user* in the above section). In the real-world situation, the known QoS records for a user only occupy a very small part of all 100 services. For satisfying the actual condition, some records are removed from training data matrix to construct three kinds of sparse data sets, whose data densities are set as 5%, 10% and 15%, respectively.

Similarly, for active users, we only retain GN (=5, 10 or 20) QoS records for each one of them. The records which are kicked out from test data set are treated as real data for prediction quality evaluation. The main parameters in our experiments are listed in Table 2.

4.2 Comparative analysis

In general, recommendation system uses mean absolute error (MAE) to evaluate prediction effect. It is the average of difference values between the predicted QoS and the real record.

$$\text{MAE} = \frac{\sum_{U,S} |P(r_{u,s}) - r_{u,s}|}{N} \tag{10}$$

where $P(r_{u,s})$ is the predicted QoS value of service s observed by user u, $r_{u,s}$ is the real QoS value of service s w.r.t. user u, and N is the total number of predictions.

Since the range of service's QoS value may be different from each other, MAE is not objective enough to reflect the accuracy of prediction algorithm. Here, we adopt the normalized MAE (NMAE) as a metric to compare the prediction quality of three algorithms.

$$\text{NMAE} = \frac{\text{MAE}}{\sum_{U,S} \frac{r_{u,s}}{N}} \tag{11}$$

It is not hard to find that, the smaller NMAE value means the more accurate prediction algorithm.

For the purpose of comparison, WSRec algorithm and basic Slope One algorithm are also implemented in our experiments. All three algorithms run on the same data set described in the above subsection. Other particular settings of WSRec algorithm are in accordance with reference [5]. For QoS attribute response time (RT) and throughput (TP), the comparisons on three algorithms are performed respectively. In the experiments, we repeated 100 times for each case of $density$ and GN value, and reported the average NMAE metrics.

The experimental results (i.e. NMAEs) on QoS attribute response time (RT) are shown in Table 3. It is clear that our SASO algorithm can outperform WsRec and basic Slope One algorithm for most cases. When $density$=5%, the basic Slope One can get the best result for the case of GN=5, but algorithm SASO (λ=1) overcomes other two algorithms for the remaining cases about GN. For the rest values (i.e. 10% and 15%) of $density$, algorithm SASO (λ=3) can achieve the lowest NMSE for nearly all cases except of $density$=15% and GN=20. On the whole, SASO's performance is better than those of WsRec and basic Slope One for almost all situations, especially λ=2 or 3.

The NMAE values of three algorithms on QoS attribute throughput (TP) are shown in Table 4. It is not hard to find our algorithm SASO (λ=3) has obvious improvement both for WsRec and basic Slope One, except of the case of $density$=15% and GN=20. With regard to SASO algorithm itself, the predication error of SASO reduces with the increase of λ value. When λ reaches to 2, algorithm SASO outperforms other two algorithms in most conditions.

According to the above experimental analysis, we can reasonably draw a conclusion that our SASO algorithm is a better choice than WsRec and basic Slope One for service's QoS prediction, especially when the data density of user-service record matrix is low.

4.3 Filling pattern analysis

As we stated earlier, the advantage of Slope One-based method is mainly for the sparse training data. However, Pearson correlation-based method will exhibit its merit with the increase of data density. In the above experiments, we only used one way (i.e. our SASO algorithm) to fill the missing data in training matrix. How about the effect of missing value supplement with two kinds of approaches? Suppose the density of training matrix during the procedure of missing value supplement is d' (obviously, $d' > d$), and the boundary point for switching filling approach is denoted as ρ. Here, we attempt to answer the above problem

[1] WS-DREAM data set, http://www.wsdream.net:8080 /wsdream/

No.	Parameter	Value	Description
1	m	150	The number of service users, 100 users for training and the rest for test.
2	n	100	Service number.
3	*density* (d)	5%, 10% or 15%	Data density of the training matrix.
4	Given Number (GN)	5, 10 or 20	The number of known QoS records for each active user.
5	λ	1, 2 or 3	The factor of adjustment strength.

Table 2: Parameter settings for service QoS prediction algorithm.

Algorithm		d=5%			d=10%			d=15%		
		GN=5	GN=10	GN=20	GN=5	GN=10	GN=20	GN=5	GN=10	GN=20
Slope One		**0.6306**	0.6142	0.6015	0.6050	0.5878	0.5718	0.5951	0.5819	0.5665
WsRec		0.6463	0.6240	0.6110	0.6001	0.5762	0.5578	0.5755	0.5596	**0.5221**
SASO	λ=1	0.6330	**0.6107**	**0.5957**	0.5923	0.5719	0.5570	0.5821	0.5645	0.5492
	λ=2	0.6350	0.6123	0.5960	0.5856	0.5654	0.5514	0.5721	0.5537	0.5401
	λ=3	0.6375	0.6155	0.5987	**0.5825**	**0.5627**	**0.5491**	**0.5652**	**0.5472**	0.5348

Table 3: Experimental results (NMAEs) for the algorithm Slope One, WsRec and SASO for the QoS attribute RT.

by using a two-stage filling strategy: SASO algorithm can be used for filling data when the matrix is relative sparse (i.e. $d' \leq \rho$). Once training matrix reaches to a certain degree of density (i.e. $d' > \rho$), we used Pearson correlation-based method to supply the missing values.

In order to validate the effect of the above two-stage filling pattern, the training matrixes with different ρ are prepared for SASO prediction algorithm. The experimental results are shown in Figure 3-6. On the whole, the two-stage filling strategy has certain improvement for some situations, but it is not so obvious. Specifically speaking, for the case of *density*=5%, the optimal boundary point is $\rho = 0.4$ for QoS attribute response time (TR). The NMAE value gradually declines when ρ is lower than 0.4. Instead, when ρ exceeds the best point 0.4, prediction error will slowly climb with the increase of ρ's value. On the second attribute throughput (TP), the trend of NMAE's change is relatively simple, that is, the error descends with the increase of value of boundary point (ρ).

For the second case *density*=10%, the change trends of prediction error for two QoS attributes are highly consistent. From the overall point of view, NMAE basically decreases along with the growth of ρ's value. However, NMAE has a little drop at the point of ρ=0.7. As a consequence, the best boundary point for this case is 0.7.

While considering the last case (i.e. *density*=15%), the change trend of prediction error for attribute RT is very similar to the second case of this attribute, just the current fluctuation is too small. There is an exceptional case for attribute TP when ρ=0.2, the corresponding NMAE value is suddenly low. Meanwhile, the prediction error value has a relatively high value at point ρ=0.3. Subsequently, it has a small reduction at first and then gradually takes off. The optimal boundary point in this case can be considered as 0.5 for most situations.

On the whole, we can argue that the two-stage filling strategy has a small improvement w.r.t prediction error. Considering the selection of boundary point, the value in the domain from 0.5 to 0.7 is worth considering in practice.

5 Related work

From the perspective of service users, how to select a suitable service is a critical step to build a reliable software system. In general, service selection is mainly in accordance with the property of QoS. Accordingly, QoS prediction for Web services has caused widespread attention in the field of service computing.

As we mentioned earlier, Pearson correlation-based algorithms are the main-stream strategies to treat such problem at current stage. Shao *et al.* [4] firstly attempted to use Pearson similarity-based collaborative filtering to provide the QoS value of a specific Web service. But their experiments are performed on a data set in small scale, and the error analysis is not so sufficient. Subsequently, Zheng *et al.* [5] firstly collected plenty of QoS records from different service users via a monitoring platform Planet-lab[2]. Then, they combined user-based and item-based CF together to form a comprehensive algorithm (i.e. WsRec) for service's QoS prediction. Their WsRec exhibits better performance than the single user-based or item-based prediction algorithm.

Recently, some improvements on Pearson correlation-based algorithm have been proposed. Liu's research group presented a personalized hybrid collaborative filtering (PHCF) algorithm by considering the personal information about service user [7]. However, it is not so easy to obtain such personal information, so the application of their method is limited. Reference [15] adopted an improved similarity measure for Web service similarity computation, and the corresponding normal recovery collaborative filtering (NRCF) was proposed for personalized Web service recommendation. In essence, it is only a minor modify

[2]http://www.planet-lab.org

Algorithm		d=5%			d=10%			d=15%		
		GN=5	GN=10	GN=20	GN=5	GN=10	GN=20	GN=5	GN=10	GN=20
Slope One		0.5115	0.5083	0.5054	0.4901	0.4873	0.4815	0.4793	0.4795	0.4707
WsRec		0.5326	0.5245	0.5195	0.4798	0.4724	0.4670	0.4579	0.4520	**0.4404**
SASO	λ=1	0.5050	0.5007	0.4968	0.4793	0.4736	0.4657	0.4687	0.4657	0.4571
	λ=2	0.5027	0.4967	0.4918	0.4735	0.4663	0.4582	0.4616	0.4566	0.4488
	λ=3	**0.5015**	**0.4946**	**0.4888**	**0.4701**	**0.4620**	**0.4533**	**0.4567**	**0.4502**	0.4426

Table 4: Experimental results (NMAEs) for the algorithm Slope One, WsRec and SASO for the QoS attribute TP.

on the similarity measure for the WsRec prediction framework. In addition, Shi *et al*. [16] presented a linear regression prediction algorithm for Web service's QoS based on clustering user in respect to location and network condition. It is not hard to find that the distance between users plays a significant role for prediction precision, however, which is not easily measured in practice.

Of course, there are also some Slope One-based methods for service's QoS prediction. Reference [6] presented a personalized context-aware QoS prediction method based on the Slope One approach. In this work, the basic Slope One algorithm is used for prediction, but it has been validated to be not very precise in our experiments. Then, Li *et al*. [17] utilized an enhanced Slope One method called Bi-Polar Slope One to predict the ratings of Web services. On the one hand, their approach mainly aims at the rating prediction problem. On the other hand, Bi-Polar phenomenon maybe exists in the data set in rating style, but not obvious in QoS data (i.e. the continuous data type).

With regard to the combination of Slope One and Pearson similarity, the preliminary researches in [18] and [19] have contributed an incipient idea for blending them together. However, the above works merely provide a primitive form of similarity-aware Slope One prediction algorithm, that is, the case of λ=1 in our work. As shown in our experimental results, this basic form without weight adjustment is not very effective for QoS prediction problem. At the same time, the experimental analysis and discussion are very limited in their work. Besides the weight adjustment strategy illustrated in formula (7), here, a more important strategy named SPC-based smoothing is also proposed to reduce prediction error.

6 Conclusion

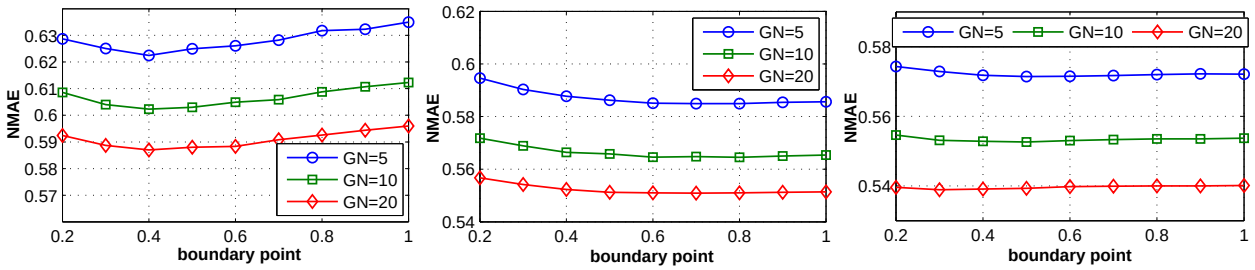
With the widespread application of service computing, Web services have been viewed as a prevalent form of components for building software on the Web. In order to ensure the reliability and trustworthy of the composite software system, users generally are very concerned about the quality of service. Unfortunately, the QoS metrics of some services can not be provided due to the actual situation. Therefore, how to predicate QoS of Web service becomes a valuable task in the field of service engineering.

In the paper, we introduce the Pearson similarity between Web services into Slope One collaborative filtering for solving QoS prediction problem. Instead of assigning the identical weight to each service, we adjust Pearson similarity as a weight for differentiating the deviation between services. In order to improve the prediction accuracy, a SPC-based smoothing is presented for correcting the exceptional data. In the empirical aspects, besides our approach, the basic Slope One and the well-known WsRec algorithm are also implemented. Meanwhile, the comparative analysis is also performed on the public published data set. The experimental results indicate that our hybrid algorithm (SASO) outperforms other two methods in the term of prediction precision. The SPC-based smoothing strategy can effectively handle the noise data so as to reduce prediction error. Furthermore, an additional strategy called two-stage filling is studied, and the appropriate boundary point for transforming filling methods is also suggested here.

The practice of SASO algorithm is obvious, it can guide users to pick out desired services from cloud platform. At the same time, this algorithm can also be used in the field of E-commerce to help consumers choose goods. Of course, although our approach achieves some promising results at present, there are still quite a few complicated issues should be further investigated. For instance, the QoS prediction for Web services from the dynamic perspective [20], as well as the service quality prediction in the environment of mobile computing. In addition, to find more effective data filling algorithm for training data is an interesting research direction.

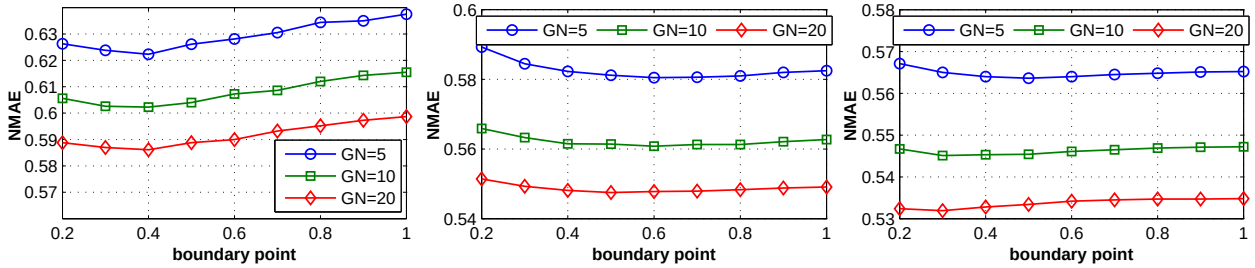
Acknowledgement

The authors would like to express their appreciation to the reviewers for their thoughtful and constructive comments. The authors would also like to thank Xiaomei Xie for her helpful feedback on the earlier version of this paper. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 60803046 and 61063013, the Natural Science Foundation of Jiangxi Province under Grant No. 2010GZS0044, the Science Foundation of Jiangxi Educational Committee under Grant No. GJJ10433, the Open Foundation of State Key Laboratory of Software Engineering under Grant No. SKLSE2010-08-23, and the Program for Outstanding Young Academic Talent in Jiangxi University of Finance and Economics.



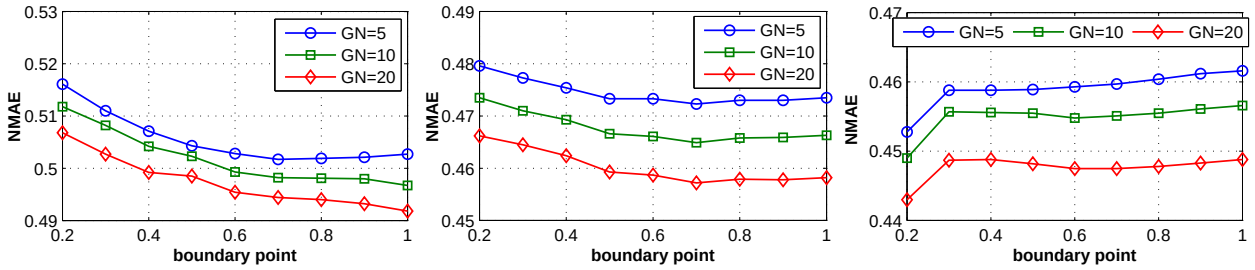
(a) *density*=5% (b) *density*=10% (c) *density*=15%

Figure 3: The NMAEs on attribute RT for different boundary points (λ=2).



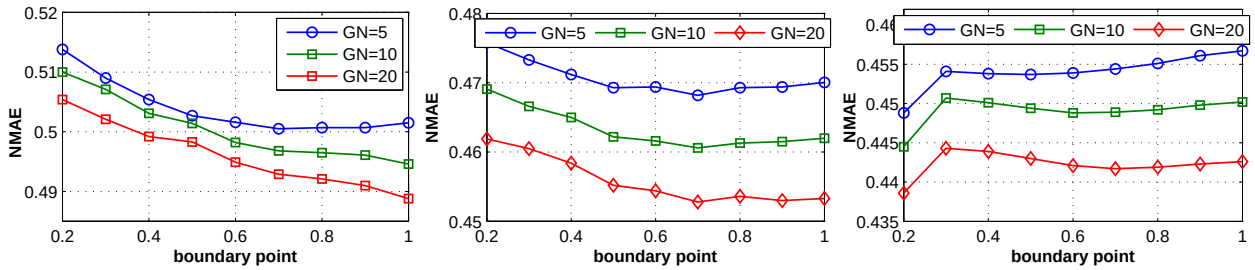
(a) *density*=5% (b) *density*=10% (c) *density*=15%

Figure 4: The NMAEs on attribute RT for different boundary points (λ=3).



(a) *density*=5% (b) *density*=10% (c) *density*=15%

Figure 5: The NMAEs on attribute TP for different boundary points (λ=2).



(a) *density*=5% (b) *density*=10% (c) *density*=15%

Figure 6: The NMAEs on attribute TP for different boundary points (λ=3).

References

[1] M. P. Papazoglou and D. Georgakopoulos, (2003). Serive-Oriented Computing, *Communcications of the ACM*, Vol. 46, No. 10, ACM Press, pp. 25–65.

[2] D. A. Menasce, (2002). QoS issues in Web services, *IEEE Internet Computing*, Vol. 6, No. 6, IEEE CS Press, pp. 72–75.

[3] J. B. Schafer, J. Konstan, and J. Riedi, (1999). Recommender Systems in E-Commerce, *Proc. of the 1st ACM Conference on Electronic Commerce (EC'09)*, ACM Press, Denver, CO, USA, pp. 158–166.

[4] L. Shao, J. Zhang, Yong Wei, and *et al*., (2007). Personalized QoS Prediction forWeb Services via Collaborative Filtering, *Proc. of the IEEE International Conference on Web Services (ICWS'07)*, IEEE CS Press, Salt Lake City, Utah, USA, pp. 439–446.

[5] Z. Zheng, H. Ma, M. R. Lyu, and I. King, (2011). QoS-Aware Web Service Recommendation by Collaborative Filtering, *IEEE Trans. on Services Computing*, Vol.4, No. 2, IEEE CS Press, pp. 140–152.

[6] Q. Xie, K. Wu, J. Xu, and *et al*., (2010). Personalized Context-Aware QoS Prediction for Web Services Based on Collaborative Filtering, *Proc. of the 6th International Conference on Advanced Data Mining and Applications (ADMA'10)*, Part II, Springer-Verlag Berlin, Chongqing, China, pp. 368–375.

[7] Y. Jiang, J. Liu, M. Tang, and X. Liu, (2011). An Effective Web Service Recommendation Method based on Personalized Collaborative Filtering, *Proc. of IEEE International Conference on Web Services (ICWS'11)*, IEEE CS Press, Washington, DC, USA, pp. 211–218.

[8] A. K. Menon, K. P. Chitrapura, S. Garg, and *et al*., (2011). Response Prediction using Collaborative Filtering with Hierarchies and Side-information, *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*, ACM Press, San Diego, CA, USA, pp. 141–149.

[9] D. Lemire and A. Maclachlan, (2005). Slope One Predictors for Online Rating-Based Collaborative Filtering, *Proc. of the 2005 SIAM International Data Mining Conference (SDM'05)*, Newport Beach, California, USA, pp. 1–5.

[10] X. Su and T. M. Khoshgoftaar, (2009). A Survey of Collaborative Filtering Techniques, *Advances in Artificial Intelligence*, Hindawi Publishing Corporation, pp. 1–19.

[11] Linyuan Lü, Matéš Medo, Chi Ho Yeung, and *et al*., (2012). Recommender Systems, *Physics Reports*, Vol. 519, No. 1, Elsevier B. V., pp. 1–49.

[12] S. Vucetic and Z. Obradovic, (2000). A Regression-based Approach for Scaling-up Personalized Recommender Systems, *Proc. of the ACM WebKDD Workshop (WebKDD'00)*, Boston, MA, USA, pp. 1–9.

[13] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, (2001). Item-based Collaborative Filtering Recommender Algorithms, *Proc. of the 10th International Conference on World Wide Web (WWW'01)*, ACM Press, Hong Kong, China, pp. 285–295.

[14] J. Oakland, (2003). *Statistical Process Control*, the 5th Revised Edition, Butterworth-Heinemann Ltd.

[15] H. Sun, Z. Zheng, J. Chen, and M. R. Lyu, (2011). NRCF: A Novel Collaborative Filtering Method for Service Recommendation, *Proc. of IEEE International Conference on Web Services (ICWS'11)*, IEEE CS Press, Washington, DC, USA, pp. 702–703.

[16] Y. Shi, K. Zhang, B. Liu, and L. Cui, (2011). A New QoS Prediction Approach Based on User Clustering and Regression Algorithms, *Proc. of IEEE International Conference on Web Services (ICWS'11)*, IEEE CS Press, Washington, DC, USA, pp. 726–727.

[17] J. Li, L. Sun, and J. Wang, (2012). A Slope One Collaborative Filtering Recommendation Algorithm Using Uncertain Neighbors Optimizing, *Proc. of WAIM 2011 International Workshops*, Springer-Verlag Berlin, Wuhan, China, pp. 160–166.

[18] P. Wang and H. W. Ye, (2009). A Personalized Recommendation Algorithm Combining Slope One Scheme and User Based Collaborative Filtering, *Proc. of International Conference on Industrial and Information Systems (IIS'09)*, IEEE CS Press, Haikou, China, pp. 152–154.

[19] D. J. Zhang, (2009). An Item-based Collaborative Filtering Recommendation Algorithm Using Slope One Scheme Smoothing, *Proc. of the Second International Symposium on Electronic Commerce and Security (ISECS'09)*, Vol. 2, IEEE CS Press, Nanchang, China, pp. 215–217.

[20] X. Li, F. Zhou, and X. Yang, (2011). Research on Trust Prediction Model for Selecting Web Services based on Multiple Decision Factors, *International Journal of Software Engineering and Knowledge Engineering*, Vol. 21, No. 8, World Scientific Publishing Company, pp. 1075–1096.

Enhanced Time-Bound Ticket-Based Mutual Authentication Scheme for Cloud Computing

Ravi Singh Pippal
Radharaman Institute of Research and Technology, Bhopal, India
E-mail: ravesingh@gmail.com

Jaidhar C. D.
Defence Institute of Advanced Technology, Girinagar, Pune, India
E-mail: jaidharcd@diat.ac.in

Shashikala Tapaswi
ABV-Indian Institute of Information Technology and Management, Gwalior, India
E-mail: stapaswi@iiitm.ac.in

Keywords: authentication, cloud computing, cryptanalysis, impersonation attack, smart card

Received: December 3, 2012

Cloud computing is a recently developed technology for complex systems with services sharing among various registered users. Therefore, proper mutual authentication is needed between users and cloud server prior to avail the services provided by cloud servers. Recently, Hao et al. [26] proposed time-bound ticket-based mutual authentication scheme for cloud computing. However, this paper shows that their scheme is vulnerable to Denial-of-Service attack and insecure password change phase. Besides, enhanced scheme is proposed to overcome these security pitfalls. Moreover, performance comparison of both the schemes proves that the enhanced scheme is more efficient in comparison with Hao et al.'s scheme.

Povzetek: V tem članku je predlagana okrepljena shema medsebojne avtentifikacije aplikacij v oblaku, ki odpravi nekatere varnostne slabosti.

1 Introduction

Cloud computing is a new computing paradigm and got wide popularity from both industries as well as academia since 2007. It is employed because of its powerful computing and storage capabilities necessary in a distributed environment [1]. Its attractive characteristics include on-demand self-service, measured service, location independent resource pooling, ubiquitous network access and rapid elasticity. Three types of service offered by cloud computing are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Several firms like Google, Amazon, Microsoft, IBM and Yahoo are the ancestors that offer services for Internet users. Some more firms like Facebook, Salesforce, Myspace, Youtube, etc. are also started offering cloud computing services.

Users who are acquainted to use Internet can avail the computing resources, storage space and software services as per their demands to solve their problems. Further, users can also store their data in cloud servers and the same can be accessed from anywhere over the Internet as on-demand. This offers great flexibility for remote users.

Although, it provides a number of advantages such as cost reduction, dynamic resource provisioning, increased flexibility, low capital expenditures and time saving for new service deployment. However, still it is not matured enough to preserve data confidentiality as well as integrity. Many security issues, like data security either in store form or transmission form, application security, monitoring and metering need to be addressed and so on. Number of security issues have been discussed [2, 3, 4, 5, 6] and few research works address the security issues [7, 8, 9, 10].

One of the primary security needs is user authentication. Several authentication schemes have been proposed in the literature but most widely used one is password based authentication scheme [11, 12, 13, 14]. However, single factor password based authentication is not secure enough in the present scenario. Two factor authentication is a better option using password as one and smart card as other factor. Smart card is a tamper resistant integrated circuit card with memory to store personal information and a processor capable of performing computations [15].

In this context, many password based smart card authentication schemes have been proposed in order to avoid the use of the verification tables [16, 17, 18, 19]. Subsequently, authentication based on smart card has been employed continuously in several applications like healthcare [20], key exchange in IPTV broadcasting [21, 22], wireless networks

[23], authentication in multi-server environment [24], wireless sensor networks [25] and many more.

1.1 Contribution of this Paper

Cloud servers authenticate the remote users prior to offer any services to them. Recently, Hao *et al.* [26] proposed time-bound ticket-based mutual authentication scheme for cloud computing. It is claimed that the scheme resists lost smart card attacks, offline password guessing attack, lost ticket attack, masquerade attack and replay attack. In addition, it provides mutual authentication and secure session key generation. This paper shows vulnerabilities of Hao *et al.*'s scheme, i.e. vulnerable to Denial-of-Service attack and insecure password change phase. To resist these weaknesses, this paper proposes an enhancement to Hao *et al.*'s scheme.

The rest of this paper is organized as follows. Section 2 gives review of Hao *et al.*'s scheme. Security pitfalls of Hao *et al.*'s scheme is shown in section 3. Section 4 describes the proposed enhanced mutual authentication scheme. An in-depth security analysis and performance comparison is discussed in section 5. Finally, section 6 concludes the paper.

2 Review of Hao *et al.*'s Scheme

This section describes Hao *et al.*'s time-bound ticket-based mutual authentication scheme for cloud computing [26] (see Figure 1). The scheme consists of four phases: Registration phase, Verification request phase, Mutual authentication phase and Password change phase. The notations used throughout this paper are summarized in Table 1.

Table 1: Notations used in this paper

Symbols	Their meaning
U_i	Remote user
ID_i	Identity of U_i
PW_i	Password chosen by U_i
S	Cloud server
U_a	Attacker
PW_a	Password chosen by U_a
t	Number of digital tickets needed by U_i
$T_i^{(j)}$	j^{th} ticket of U_i
$TID_i^{(j)}$	j^{th} ticket ID
$VP_i^{(j)}$	Valid period of $T_i^{(j)}$
k_1,k_2	Two long term secret keys of S
$H(\cdot)$	Cryptographic hash function
$H_k(\cdot)$	Keyed hash function
$\|$	Concatenation
$\oplus$	Bitwise XOR operation
r_u	Random nonce generated by U_i
r_s	Random nonce generated by S
r_a	Random nonce generated by U_a
K_c/K_s	Shared session key between U_i and S

2.1 Registration Phase

This phase is invoked when a new user registers with the cloud server. The cloud server issues 't' tickets, in which each ticket can be used only once. In this phase, U_i selects ID_i, PW_i and a random number b, computes $IPB_i = H(ID_i \parallel H(PW_i \oplus b))$ and submits $\{ID_i, IPB_i, t\}$ to S over a secure channel, where 't' is the number of digital tickets needed by U_i.

Upon receiving the registration request and ticket fee from U_i, S generates t tickets for U_i. j^{th} ticket of U_i and its validity is represented as $\{(TID_i^{(j)}, VP_i^{(j)}), j = 1, 2, ..t\}$. S computes

$$W_i = IPB_i \oplus H(ID_i, K_1)$$

$$\alpha_i^{(j)} = H_{K_2}(ID_i \parallel TID_i^{(j)} \parallel VP_i^{(j)})$$

$$\beta_i^{(j)} = \alpha_i^{(j)} \oplus IPB_i$$

$T_i^{(j)}$ has two parts,

$$T_i^{(j)} = (T_i^{(j)1}, T_i^{(j)2})$$

in which

$$T_i^{(j)1} = (TID_i^{(j)}, VP_i^{(j)})$$

$$T_i^{(j)2} = \beta_i^{(j)}$$

S also computes $Z_i = H_{K_2}(ID_i) \oplus IPB_i$ and issues a smart card to U_i by storing $\{ID_i, t, W_i, Z_i, T_i^{(j)}\}$ into smart card memory over secure channel. After receiving, U_i stores b into smart card memory.

2.2 Verification Request Phase

As U_i receives t tickets, these tickets can be used to perform data verification at most t times. Suppose, for k^{th} verification request, U_i inserts the smart card to the card reader and keys in ID_i and PW_i. The smart card generates a nonce r_u and computes

$$IPB_i = H(ID_i \parallel H(PW_i \oplus b))$$

$$H_i = W_i \oplus IPB_i$$

$$C_1 = r_u \oplus H_i$$

$$C_2 = H(r_u) \oplus T_i^{(k)2} \oplus IPB_i$$

U_i sends the verification request $\{ID_i, T_i^{(k)1}, C_1, C_2\}$ to S in order to pass the mutual authentication phase.

2.3 Mutual Authentication Phase

Once the verification request has been received, S first checks the validity of ID_i to accept/reject the verification request. S rejects the request when it finds invalidity otherwise checks whether $TID_i^{(k)}$ is on the bulletin board or not. If it's on the bulletin board, S rejects U_i's request and terminates the process. S checks whether the current date is within the range of $VP_i^{(k)}$ or not. If not, S rejects U_i's request and terminates the process.

If all these conditions hold, S computes

$$D_0 = H(ID_i, K_1)$$

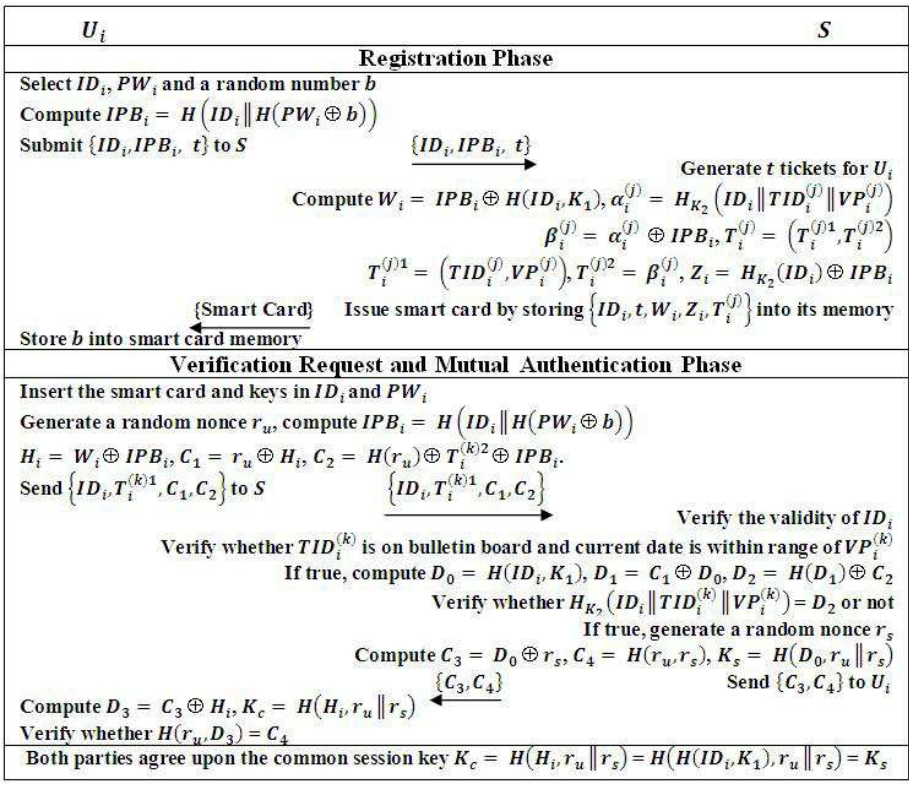


Figure 1: Hao *et al.*'s Scheme

$$D_1 = C_1 \oplus D_0$$
$$D_2 = H(D_1) \oplus C_2$$

S computes $H_{K_2}(ID_i \parallel TID_i^{(k)} \parallel VP_i^{(k)})$ and checks whether it is equal to D_2 or not. If true, S generates a random nonce r_s, computes $C_3 = D_0 \oplus r_s$, $C_4 = H(r_u, r_s)$ and sends the message $\{C_3, C_4\}$ to U_i. S also computes $K_s = H(D_0, r_u \parallel r_s)$ as the session key.

After getting the message $\{C_3, C_4\}$ from S, U_i computes $D_3 = C_3 \oplus H_i$ and compares $H(r_u, D_3)$ with C_4. If true, U_i authenticates S successfully otherwise terminates the session. Subsequently, U_i computes $K_c = H(H_i, r_u \parallel r_s)$. Both parties agree upon the common session key $K_c = H(H_i, r_u \parallel r_s) = H(H(ID_i, K_1), r_u \parallel r_s) = K_s$.

2.4 Password Change Phase

This phase is invoked when U_i wants to change the password. U_i inserts the smart card to the card reader and keys the credentials such as ID_i and PW_i. The smart card generates a nonce r_u and computes

$$IPB_i = H(ID_i \parallel H(PW_i \oplus b))$$
$$C_1 = r_u \oplus W_i \oplus IPB_i$$
$$C_2 = H(r_u) \oplus Z_i \oplus IPB_i$$

The smart card sends $\{update, ID_i, C_1, C_2\}$ to S, in which, $update$ denotes that it's a password change request. After receiving, S checks the validity of ID_i to accept/reject the request. If it is invalid, then S rejects the request otherwise computes

$$D_1 = C_1 \oplus H(ID_i, K_1)$$
$$D_2 = H(D_1) \oplus C_2$$

S computes $H_{K_2}(ID_i)$ and checks whether it is equal to D_2 or not. If true, S generates a random nonce r_s, computes $C_3 = H(ID_i, K_1) \oplus r_s$, $C_4 = H(r_u, r_s)$ and sends the message $\{C_3, C_4\}$ to U_i. Upon receiving the message $\{C_3, C_4\}$, smart card computes $D_3 = C_3 \oplus W_i \oplus IPB_i$ and compares $H(r_u, D_3)$ with C_4. If true, U_i authenticates S successfully otherwise terminates the session. Subsequently, smart card prompts U_i to enter a new password PW_i^{new}. Then, smart card computes

$$IPB_i^{new} = H(ID_i \parallel H(PW_i^{new} \oplus b))$$
$$W_i^{new} = W_i \oplus IPB_i \oplus IPB_i^{new} = H(ID_i, K_1) \oplus IPB_i^{new}$$
$$Z_i^{new} = Z_i \oplus IPB_i \oplus IPB_i^{new} = H_{K_2}(ID_i) \oplus IPB_i^{new}$$

The smart card updates $T_i^{(j)2}$ to $T_i^{(j)2} \oplus IPB_i \oplus IPB_i^{new}$ for all remaining tickets which yields $\alpha_i^{(j)} \oplus IPB_i^{new}$.

3 Weakness in Hao *et al.*'s Scheme

This section provides security flaws in Hao *et al.*'s scheme. They are (a) exposed to Denial-of-Service attack due to lack of early wrong password detection prior to verification request creation and (b) inefficient password change phase. It is assumed that the attacker U_a is able to intercept all the messages exchanged between U_i and S.

3.1 Denial-of-Service Attack

To check whether or not the requested user is a legitimate bearer of smart card, entered password must be verified at the smart card level before login request creation [27]. In this scheme, if U_a gets U_i's smart card by any means, he or she can create invalid login request by entering wrong password which is verified only at the cloud server side not at the user side.

Assume, U_a gets/steals U_i's smart card, inserts the smart card into the card reader and enters the wrong password PW_a as well as ID_a. Smart card creates an invalid login request without verifying the correctness of entered password or identifier. The smart card generates a nonce r_a and computes

$$IPB_a = H(ID_a \parallel H(PW_a \oplus b))$$
$$H_a = W_i \oplus IPB_a = IPB_i \oplus H(ID_i, K_1) \oplus IPB_a$$
$$C_{1a} = r_a \oplus H_a = r_a \oplus IPB_i \oplus H(ID_i, K_1) \oplus IPB_a$$
$$C_{2a} = H(r_a) \oplus T_i^{(k)2} \oplus IPB_a$$

U_a sends the verification request $\{ID_i, T_i^{(k)1}, C_{1a}, C_{2a}\}$ to S. This request fails to pass the authentication phase at the cloud server side. As a result, load on S increases which leads to Denial-of-Service attack. To overcome this attack, both password and identifier must be verified at the user side prior to compute verification request.

3.2 Insecure Password Change Phase

Communication is needed between S and U_i during the password change phase. Password change at the user side without interacting with S strengthen the security and reduces the load on S. Further, password change phase leads to Denial-of-Service attack because of non existence of earlier password as well as identifier verification before the update request creation [27].

4 Proposed Enhanced Mutual Authentication Scheme

This section describes proposed enhanced mutual authentication scheme over Hao *et al.*'s scheme (see Figure 2). The scheme consists of four phases: Registration phase, Verification request phase, Mutual authentication phase and Password change phase. The details of these phases are as follows:

4.1 Registration Phase

In this phase, U_i selects ID_i, PW_i and a random number b, computes $H(PW_i \oplus b)$ and submits $\{ID_i, H(PW_i \oplus b), t\}$ to S over a secure channel, where 't' is the number of digital tickets needed by U_i. Upon receiving the registration request and ticket fee from U_i, S generates t tickets for U_i. j^{th} ticket of U_i and its validity is represented as $\{(TID_i^{(j)}, VP_i^{(j)}), j = 1, 2, ..t\}$. S computes

$$W_i = H(ID_i \parallel H(PW_i \oplus b))$$

$$X_i^{(j)} = H_x(ID_i \parallel TID_i^{(j)} \parallel VP_i^{(j)}) \oplus H(ID_i, x)$$

where 'x' is long term secret key of S. $T_i^{(j)}$ has two parts,

$$T_i^{(j)} = (T_i^{(j)1}, T_i^{(j)2})$$

in which

$$T_i^{(j)1} = (TID_i^{(j)}, VP_i^{(j)})$$
$$T_i^{(j)2} = X_i^{(j)}$$

S issues a smart card over secure channel to U_i by storing $\{ID_i, t, W_i, T_i^{(j)}\}$ into smart card memory. After receiving, U_i stores b into smart card memory.

4.2 Verification Request Phase

As U_i receives t tickets, these tickets can be used to perform data verification at most t times. Assume for k^{th} verification request, U_i inserts the smart card to the card reader and keys the credentials, ID_i' and PW_i'. The smart card computes $W_i' = H(ID_i' \parallel H(PW_i' \oplus b))$ and compares it with the stored W_i. If true, U_i is the valid owner of smart card.

The smart card generates a nonce r_u and computes $Y_i = H_{T_i^{(k)2}}(T_i^{(k)2} \parallel r_u)$. U_i sends the verification request $\{ID_i, T_i^{(k)1}, Y_i, r_u\}$ to S.

4.3 Mutual Authentication Phase

Upon receiving the verification request $\{ID_i, T_i^{(k)1}, Y_i, r_u\}$; S first checks the validity of ID_i to accept/reject the verification request. S rejects the request when it finds invalidity otherwise checks whether $TID_i^{(k)}$ is on the bulletin board or not. If it's on the bulletin board, S rejects U_i's request and terminates the process. S checks whether the current date is within the range of $VP_i^{(k)}$ or not. If not, S rejects U_i's request and terminates the process.

If all these conditions hold, S computes $X_i^{(k)} = H_x(ID_i \parallel TID_i^{(k)} \parallel VP_i^{(k)}) \oplus H(ID_i, x)$. S computes $Y_i' = H_{X_i^{(k)}}(X_i^{(k)} \parallel r_u)$ and checks whether it is equal to received Y_i or not. If true, S authenticates U_i otherwise rejects the request. S generates a random nonce r_s, computes $Z_i = H_{X_i^{(k)}}(r_u \parallel r_s \parallel X_i^{(k)})$ and sends the message $\{ID_i, Z_i, r_s\}$ to U_i. S also computes $K_s = H(ID_i \parallel r_u \parallel r_s \parallel X_i^{(k)})$ as the session key.

After getting the message $\{ID_i, Z_i, r_s\}$ from S, U_i computes $Z_i' = H_{T_i^{(k)2}}(r_u \parallel r_s \parallel T_i^{(k)2})$ and compares it with the received Z_i. If true, U_i authenticates S successfully otherwise terminates the session. Subsequently, U_i computes $K_c = H(ID_i \parallel r_u \parallel r_s \parallel T_i^{(k)2})$. Both parties agree upon the common session key $K_c = H(ID_i \parallel r_u \parallel r_s \parallel T_i^{(k)2}) = H(ID_i \parallel r_u \parallel r_s \parallel X_i^{(k)}) = K_s$.

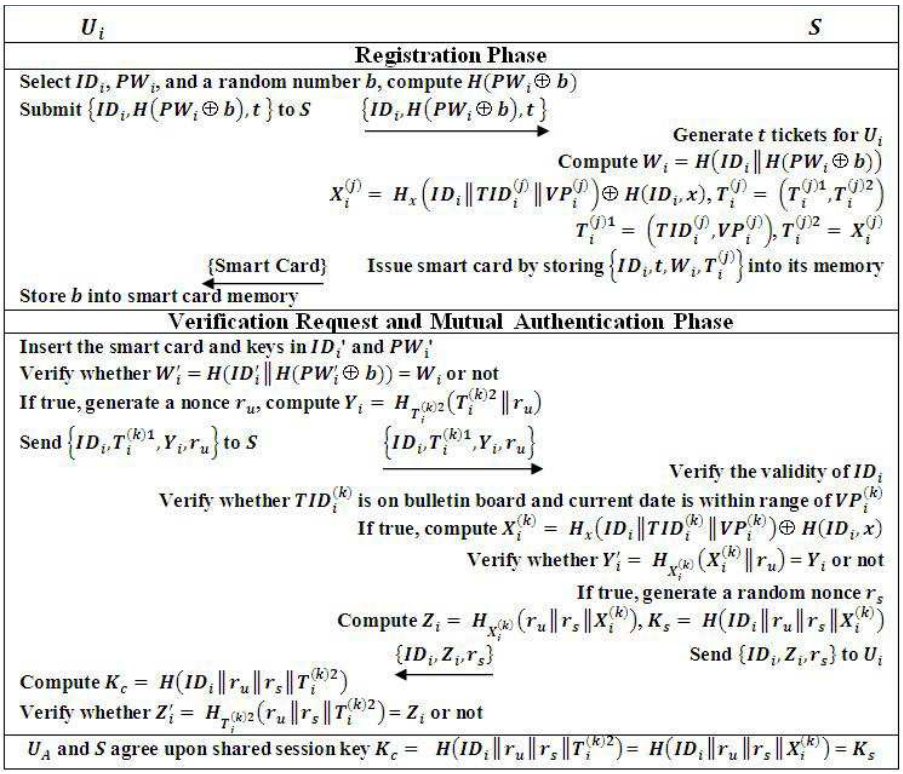


Figure 2: Proposed Enhanced Mutual Authentication Scheme

4.4 Password Change Phase

This phase is invoked when U_i wants to change the password. U_i inserts the smart card to the card reader and keys the credentials such as ID_i' and PW_i'. The smart card computes $W_i' = H(ID_i' \parallel H(PW_i' \oplus b))$ and compares it with the stored W_i. If true, U_i is the legitimate bearer of smart card.

Subsequently, smart card prompts U_i to enter a new password PW_i^{new}. Then, smart card computes $W_i^{new} = H(ID_i \parallel H(PW_i^{new} \oplus b))$. The smart card updates W_i to W_i^{new} in the smart card memory.

5 Security Analysis and Performance Comparison

This section discusses security analysis of the proposed enhanced mutual authentication scheme and provides performance analysis in comparison with Hao *et al.*'s scheme.

5.1 Impersonation Attack

Suppose, U_a has complete hold on the insecure communication channel and can intercept all the communicating messages transmitted between U_i and S. U_a is unable to create a forged verification request as the value of $T_i^{(k)2}$ is needed to compute fake Y_i. Further, it is not possible to get $T_i^{(k)2}$ from intercepted $T_i^{(k)1}$ without knowing 'x', long term secret key of S. Moreover, without the information about $T_i^{(k)2}$, U_a cannot masquerade as a legitimate S. Hence, U_a is unable to forge the verification request to impersonate a valid U_i or forge the response message to impersonate a legitimate S.

5.2 Password Guessing Attack

One of the most important features provided by any authentication scheme is the security of passwords of users. The scheme must be structured in such a way that no one can guess the password. In the proposed scheme, password is used only in the card holder verification. It is not used in the calculation of any of the verification request parameters. Hence, there is no chance of offline password guessing attack. To resist online password guessing attack, the number of attempts made by user can be limited to some fixed value.

5.3 Replay Attack

An adversary may try to act as an authentic user by resending previously intercepted messages. This scheme uses unique ticket ID TID_i and random nonces r_u and r_s which are different from session to session. As a consequence, U_a cannot enter the system by resending previously transmitted messages to impersonate legal U_i.

Assume that the intercepted verification request

$\{ID_i, T_i^{(k)1}, Y_i, r_u\}$ is replayed to pass the mutual authentication phase. Upon receiving the verification request, S first checks the validity of ID_i and then checks whether $TID_i^{(k)}$ is on the bulletin board or not. Obviously, S will find that $TID_i^{(k)}$ is on the bulletin board. S rejects the service request and terminates the process.

5.4 Reflection and Parallel Session Attack

To resist reflection and parallel session attacks, the given scheme employs asymmetric structure of communicating messages, i.e., $\{ID_i, T_i^{(k)1}, Y_i, r_u\}$ and $\{ID_i, Z_i, r_s\}$. There is no symmetry in the values of $Y_i = H_{T_i^{(k)2}}(T_i^{(k)2} \parallel r_u)$ and $Z_i = H_{X_i^{(k)}}(r_u \parallel r_s \parallel X_i^{(k)})$. Hence, U_a is unable to launch parallel session attack by replaying cloud server response message as the user verification request or reflection attack by resending user verification request as the cloud server response message.

5.5 Privileged Insider Attack

For remembrance, many users employ same password to access different servers. Nevertheless, a privileged insider of server can get this password and then try to utilize it for personal benefit. In the given scheme, U_i sends $H(PW_i \oplus b)$ to S instead of PW_i to resist privileged insider attack. Hence, this scheme provides security against privileged insider attack.

5.6 Valid Period Extending Attack

In the proposed scheme, no one can use the ticket after the expiration date. It helps to control the database growth maintained by S. Let us suppose, U_i wants to reuse the k^{th} ticket $T_i^{(k)}$. U_i changes $VP_i^{(k)}$ to $VP_i^{(k')}$ (by including the current date) and sends $\{ID_i, T_i^{(k')1}, Y_i, r_u\}$ to S.

Once received, S computes $X_i^{(k')} = H_x(ID_i \parallel TID_i^{(k)} \parallel VP_i^{(k')}) \oplus H(ID_i, x)$. Obviously, S finds Y_i' = $H_{X_i^{(k')}}(X_i^{(k')} \parallel r_u) \neq Y_i$ and rejects the request. Hence, the enhanced scheme is able to prevent the user from extending the expiration date of any ticket.

5.7 Early Wrong Password Detection

To provide security against Denial-of-Service attack, identity of users must be verified at the user side prior to creation of verification request. The enhanced scheme verifies the entered password and identifier by comparing W_i' with the stored W_i during the verification request phase. If U_i enters either password or identifier incorrect, the smart card prompt U_i to re-enter correct password as well as correct identifier. In addition, it is infeasible to guess correct identifier and password simultaneously by using stolen smart card. Hence, there is no chance for Denial-of-Service attack.

5.8 Efficient Password Change Phase

In the proposed scheme, U_i can choose and change the password without any support from S. The smart card compares the computed W_i' with the stored W_i to verify the legitimacy of U_i before the update of new password. If it holds, smart card asks U_i to enter a new password PW_i^{new}, computes W_i^{new} and updates W_i to W_i^{new} in the smart card memory. It eliminates the role of S during password change phase which diminishes burden on S.

5.9 Performance Comparison

In order to measure the security in terms of possible attacks, proposed scheme is compared with Hao *et al.*'s scheme. From Table 2, it can be clearly seen that the proposed scheme is more secure in comparison with Hao *et al.*'s scheme. It includes early wrong password and wrong identifier detection which resists Denial-of-Service attack either during verification request phase or password change phase.

Table 3 shows comparative results for Hao *et al.*'s scheme and the proposed enhanced scheme in terms of computational complexity. In this table, t denotes the number of tickets issued to user U_i and r denotes the number of tickets remaining. From both the tables, it is clear that the proposed scheme is more efficient in comparison with Hao *et al.*'s scheme.

6 Conclusion

Nowadays, cloud has become one of the most popular business transaction platform. However, the growing security threat emerging due to the present security attacks obfuscates this powerful network. Weak authentication of responses and requests allows the attackers to compromise the cloud infrastructure. Hence, authentication of both the users and the cloud servers is a vital issue. To address this aforementioned issue, Hao *et al.* [26] proposed time-bound ticket-based mutual authentication scheme for cloud computing.

This paper pointed out that Hao *et al.*'s scheme is inadequate to provide security against Denial-of-Service attack. Further, password change phase is also insecure. To overcome these security flaws, this paper proposes an enhanced scheme over Hao *et al.*'s scheme. The enhanced scheme inherits all the merits of Hao *et al.*'s scheme and resists the identified security attacks. In addition, user can choose and change the password securely without any assistance from the cloud server.

Acknowledgement

The authors would like to thank ABV-Indian Institute of Information Technology and Management, Gwalior, India for providing the academic support.

Table 2: Comparison between proposed scheme and Hao *et al.*'s scheme in terms of security properties

Security Properties	**Hao *et al.*'s Scheme**	**Proposed Scheme**
User is allowed to choose and change the password	Yes	Yes
Provides mutual authentication	Yes	Yes
Provides secure session key generation	Yes	Yes
Resists replay attack	Yes	Yes
Resists guessing attack	Yes	Yes
Resists parallel session attack	Yes	Yes
Resists reflection attack	Yes	Yes
Resists privileged insider attack	Yes	Yes
Resists valid period extending attack	Yes	Yes
Resists impersonation attack	Yes	Yes
Resists Denial-of-Service attack	No	Yes
Free from cloud server involvement during password change	No	Yes
Provides early wrong password detection	No	Yes
Provides early wrong identifier detection	No	Yes

Table 3: Comparison between proposed scheme and Hao *et al.*'s scheme in terms of computational complexity

Authentication Schemes	**Name of Phases**	**No. of Hash Functions (H)**	**No. of Exclusive-or Operations (XOR)**	**Total No. of Operations**
Hao *et al.*'s Scheme	Registration Phase	$(4+t)$	$(3+t)$	$(24+t)$ H
	Verification Request Phase	(3)	(5)	
	Mutual Authentication Phase	(7)	(4)	$(27+t+2r)$ XOR
	Password Change Phase	(10)	$(15+2r)$	
Proposed Scheme	Registration Phase	$(3+t)$	$(1+t)$	$(17+t)$ H
	Verification Request Phase	(3)	(1)	
	Mutual Authentication Phase	(7)	(1)	$(5+t)$ XOR
	Password Change Phase	(4)	(2)	

References

[1] Li, Z., Chen, C. and Wang, K. (2011). Cloud computing for agent-based urban transportation systems. *IEEE Intelligent Systems*, 26(1), pp. 73–79.

[2] Zhou, M., Zhang, R., Xie, W., Qian, W. and Zhou, A. (2010). Security and privacy in cloud computing: A survey. *In Proceedings of* 6^{th} *International Conference on Semantics, Knowledge and Grid*, Shanghai, China, pp. 105–112.

[3] Subashini, S. and Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, 34(1), pp. 1–11.

[4] Pearson, S. and Benameur, A. (2010). Privacy, security and trust issues arising from cloud computing. *In Proceedings of* 2^{nd} *IEEE International Conference on Cloud Computing Technology and Science*, Bristol, U.K., pp. 693–702.

[5] Jensen, M., Schwenk, J., Gruschka, N. and Iacono, L. (2009). On technical security issues in cloud computing. *In Proceedings of IEEE International Conference on Cloud Computing*, Bangalore, India, pp. 109–116.

[6] Kandukuri, B.R., Ramakrishna, P.V. and Rakshit, A. (2009). Cloud security issues. *In Proceedings of IEEE International Conference on Services Computing*, Bangalore, India, pp. 517–520.

[7] Takabi, H., Joshi, J.B.D. and Ahn, G.J. (2010). SecureCloud: Towards a comprehensive security framework for cloud computing environments. *In Proceedings of* 34^{th} *Annual IEEE Computer Software and Applications Conference Workshops*, P.A., U.S.A., pp. 393–398.

[8] Wang, C. and Yan, H. (2010). Study of cloud computing security based on private face recognition. *In Proceedings of International Conference on Computational Intelligence and Software Engineering*, Beijing, China, pp. 1–5.

[9] Shen, Z. and Tong, Q. (2010). The security of cloud computing system enabled by trusted computing technology. *In Proceedings of* 2^{nd} *International Conference on Signal Processing Systems*, Wuhan, China, pp. 11–14.

[10] Zech, P. (2011). Risk-based security testing in cloud computing environments. *In Proceedings of* 4^{th} *IEEE International Conference on Software Testing, Verification and Validation*, Innsbruck, Austria, pp. 411–414.

[11] Hwang, M.S., Lee, C.C. and Tang, Y.L. (2001). An improvement of SPLICE/AS in WIDE against guessing attack. *Informatica*, 12(2), pp. 297–302.

[12] Yang, C.C., Chang, T.Y. and Hwang, M.S. (2003). Security of improvement on methods for protecting password transmission. *Informatica*, 14(4), pp. 551–558.

[13] Yoon, E.J., Ryu, E.K. and Yoo, K.Y. (2005). Attacks and solutions of Yang *et al.*'s protected password changing scheme. *Informatica*, 16(2), pp. 285–294.

[14] Ku, W.C. and Tsai, H.C. (2005). Weaknesses and improvements of Yang-Chang-Hwang's password authentication scheme. *Informatica*, 16(2), pp. 203–212.

[15] http://en.wikipedia.org/wiki/Smart_card.

[16] Chang, C.C. and Wu, T.C. (1991). Remote password authentication with smart cards. *IEE Proceedings E: Computers and Digital Techniques*, 138, pp. 165–168.

[17] Chen, T.H., Horng, G. and Wu, K.C. (2007). A secure YS-like user authentication scheme. *Informatica*, 18(1), pp. 27–36.

[18] Liao, C.H., Chen, H.C. and Wang, C.T. (2009). An exquisite mutual authentication scheme with key agreement using smart card. *Informatica*, 33(2), pp. 125–132.

[19] Pippal, R.S., Jaidhar, C.D. and Tapaswi, S. (2012). Highly secured remote user authentication scheme using smart cards. *In Proceedings of* 7^{th} *IEEE International Conference on Industrial Electronics and Applications*, Singapore, pp. 988–992.

[20] Hu, J., Chen, H.H. and Hou, T.W. (2010). A hybrid public key infrastructure solution (HPKI) for HIPAA privacy/security regulations. *Computer Standards and Interfaces*, 32(5-6), pp. 274–280.

[21] Yoon, E.J. and Yoo, K.Y. (2009). Robust key exchange protocol between set-top box and smart card in DTV broadcasting. *Informatica*, 20(1), pp. 139–150.

[22] Pippal, R.S., Tapaswi, S. and Jaidhar, C.D. (2012). Secure key exchange scheme for IPTV broadcasting. *Informatica*, 36(1), pp. 47–52.

[23] He, D., Ma, M., Zhang, Y., Chen, C. and Bu, J. (2011). A strong user authentication scheme with smart cards for wireless communications. *Computer Communications*, 34(3), pp. 367–374.

[24] Pippal, R.S., Jaidhar, C.D. and Tapaswi, S. (2013). Robust Smart Card Authentication Scheme for Multiserver Architecture. *Wireless Personal Communications*. DOI: 10.1007/s11277-013-1039-6.

[25] Fan, R., He, D., Pan, X. and Ping, L. (2011). An efficient and DoS-resistant user authentication scheme for two-tiered wireless sensor networks. *Journal of Zhejiang University-SCIENCE C (Computers and Electronics)*, 12(7), pp. 550–560.

[26] Hao, Z., Zhong, S. and Yu, N. (2011). A time-bound ticket-based mutual authentication scheme for cloud computing. *International Journal of Computers, Communications and Control*, 6(2), pp. 227–235.

[27] Yoon, E.J., Ryu, E.K. and Yoo, K.Y. (2005). An improvement of Hwang-Lee-Tang's simple remote user authentication scheme. *Computers and Security*, 24(1), pp. 50–56.

A Hybrid Metaheuristic Algorithm for Job Scheduling on Computational Grids

Zahra Pooranian
Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran
E-mail: Zahra.Pooranian@gmail.com

Mohammad Shojafar
Department of Information Engineering, Electronics and Telecommunication (DIET), "Sapienza" University of Rome, Via Eudossiana 18, 00184, Rome, Italy
E-mail: Shojafar@diet.uniroma1.it

Reza Tavoli
Department of Mathematics, Islamic Azad University, Chalous Branch (IAUC)17Shahrivar Ave., P.O. Box 46615-397, Chalous, Iran
E-mail: r.tavoli@gmail.com

Mukesh Singhal
Computer Science & Engineering, University of California, Merced, CA S&E 296,USA
E-mail: msinghal@ucmerced.edu

Ajith Abraham
Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, P.O. Box 2259, Auburn, WA 98071-2259, USA
E-mail: Ajith.Abraham@ieee.org

Keywords: **g**rid computing, genetic algorithm, gravitational emulation local search (GELS), independent task scheduling.

Received: April 25, 2013

The dynamic nature of grid resources and the demands of users produce complexity in the grid scheduling problem that cannot be addressed by deterministic algorithms with polynomial complexity. One of the best methods for grid scheduling is the genetic algorithm (GA); the simple and parallel features of this algorithm make it applicable to several optimization problems. A GA searches the problem space globally and is unable to search locally. Therefore, scholars have investigated combining GAs with other meta-heuristic methods to resolve the local search problem. This is the focus of the present contribution, where we have developed a new hybrid scheduling algorithm GGA that combines GA and the gravitational emulation local search (GELS) algorithm. The noteworthy feature of the proposed optimal scheduler is that it decreases runtime and the number of submitted tasks whose deadlines are missed. A comparison of the performance of our proposed joint optimal scheduler to similar methods shows that it produces more optimal computation time.

Povzetek: Predlagana je metoda genetskih algoritmov za razvrščanje poslov v grid sistemih.

1 Introduction

Grid computing has emerged as a new approach for solving large-scale problems in scientific, engineering, and commercial fields [1].

A deciding factor in grid computing design is the purpose for which it will be used. Design goals can be divided into three major groups: increasing the efficiency of an application, improving data access, and increasing and improving services. Grid systems can be classified according to these objectives as, respectively, grid computing systems, data grids, and service grids. Further, grid-computing systems can be classified into two main categories: distributed supercomputing and high-throughput grids [2].

Data grids provide a platform for assembling new databases from distributed data sources such as digital libraries or providers' data warehouses. Although grid computing also needs to provide data services, the major difference between a data grid and grid computing is that the former provides a special platform to manage data storage and access for applications, while in grid

computing, the applications themselves must implement the storage management schema. An example use for data grids is data mining that gathers information from various sources. Two organizations that are working on developing large-scale data collections are the European data grid and Globus [2].

Systems in a service grid provide services that cannot be provided with a single machine [3]. Most research in grid computing falls under one of these classifications (data, computing, and service grids).

Among existing uses of grids, grid computing is the most prevalent. By utilizing the processing power of CPUs during their idle periods, grid computing can be several times faster than what a single computer can achieve today. Therefore, acceptable task scheduling for resources plays a crucial role in grids, especially scheduling computing resources for tasks. The main goal of most schedulers is to find a balance between execution cost and the runtime for tasks. This means that given a deadline for completing execution, the running costs are kept low, or given a fixed cost for execution; the necessary time to perform the tasks will be minimized.

Generally, there are three methods for scheduling:

1. Manual scheduling. The user divides the tasks between different resources.

2. Application-mode scheduling. Applications perform the scheduling, with each application defining the resources, such as MPI programs, required for its execution. A list of machines that have MPI programs is given to the user at runtime.

3.Scheduling that is independent of applications, such as scheduling by a grid broker. This method is much more appropriate for grid scheduling. For task processing and task analysis, applications deliver their requirements to the broker, based on the quality of service required for their tasks.

We should note that the resources for grid task scheduling are distributed in various locations. One or more resources are selected for running a task, which is then sent to those resources. The grid scheduler has no ownership or control over resources. Rather, tasks are delivered to local resource managers (LRMs) for execution. After that, the LRMs control the running status and execution of the tasks they have received.

The first phase of grid task scheduling is resource discovery, which generates a list of potential resources. The second phase includes gathering information about these resources and choosing the best set of resources matching the application's requirements. In the third phase, the task is executed, which involves file staging and cleanup [4].

Grid systems consist of heterogeneous resources, managerial systems, policies, and applications with different requirements. Since these resources are heterogeneous and distributed and are used in common, grid efficiency is highly dependent on an effective and efficient design for its scheduler. Grid scheduling is considered to be an NP-hard problem. Deterministic algorithms do not have the necessary efficiency for solving this problem. Therefore, much research has been directed toward heuristic methods. Most of these methods attempt to minimize makespan.

Many heuristic algorithms have recently been suggested for task scheduling in grid computing, including hierarchical stochastic Petri net schedulers (HSPNs) [5-8], genetic algorithms (GAs) [9], the group leaders' optimization algorithm (GLOA) [10], simulated annealing (SA) [11], the queen bee method [12], and the tabu search (TS) [13] and others [29-36]. Among these, GAs provide the best heuristic method because they are inherently parallel and can search several aspects of a problem space simultaneously. Since the convergence of a GA is slow for global optimization and has been proved to be unstable in different implementations, the efficiency of GAs can be improved by combining them with other algorithms such as GPSO [14] that it combines GELS method with PSO.

This research combines a GA and the gravitational emulation local search (GELS) algorithm. GA's are weak for local searches and strong for global searches. Conversely, GELS is a local search algorithm that imitates gravitational attraction and is therefore strong for local searches and weak for global searches. Combining the benefits of these two algorithms can solve the grid-scheduling problem. This paper presents a static scheduling algorithm for scheduling independent tasks in a grid system. "Static scheduling" means that all necessary data about tasks, resources, and the number of resources should be specified before execution. The advantage of static scheduling is that no overhead is exerted on the system. In addition to decreasing makespan, our proposed algorithm considers quality of service (QOS) to minimize the number of tasks that miss their deadlines.

The remainder of this paper is organized as follows. Section 2 briefly describes previous related work and the intelligent GA and the GELS algorithm, respectively. Section 3 describes the task-scheduling problem. Section 4 presents our proposed algorithm in detail. Section 5 compares our proposed algorithm with several similar algorithms, and Section 6 presents our conclusions and future research directions.

2 Related Work

In the following Section, we provide an overview of Genetic algorithm and GELS algorithm and explain various methods, which describe different hybrid, and joint method that applied for scheduling in grid computing.

2.1 Genetic algorithms

GAs were first proposed in 1975 by John Holland et al. [15] at Michigan University. In optimization methods, a GA or optimization inspired by nature is considered to be the most natural evaluation method. A GA selects the most suitable strings from organized stochastic information that is searched and gathered by humans. In each generation, a new set of strings is produced based on artificial strings with the help of the most suitable bits

and elements among the old elements. The new set is tested stochastically, and its strength or fitness level is evaluated. The general form of a GA is as follows:

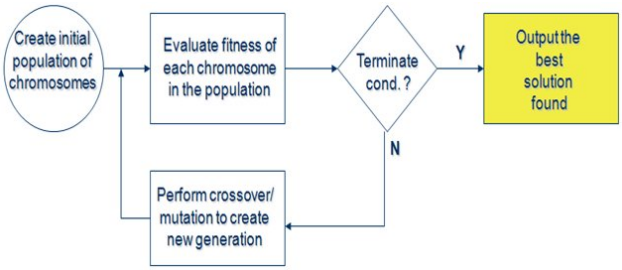


Figure 1: Genetic algorithm.

2.2 Gravitational Emulation Local Search

In 1995, Voudouris and his colleagues [16] proposed the Guided Local Search (GLS) algorithm for searching in a search space with an NP-hard solution. In 2004, Vebster [17] presented GLS as a strong algorithm, the GELS algorithm. GELS mimics gravitational attraction for searching within a search space. Each response has different neighbors that can be grouped based on problem-dependent criteria. The neighbors obtained in each neighbor group are called a *dimension*. A primary velocity is defined for each dimension, and a dimension with a greater primary velocity is more responsive for the problem. The GELS algorithm accounts for gravitational force in the responses in a search space through two methods. In the first method, a response is selected from the local neighbor space of the current response and the gravitational force between these two responses is calculated. In the second method, the gravitational force is calculated using all of the neighbor responses in a neighbor space of the current response rather than being limited to one response. GELS also implements movement into the search space with two methods. The first method allows movement from the current response toward the response to the current response in local neighbor spaces. The second method allows movement toward responses outside of the current response local neighbor spaces in addition to the neighbor responses of the current response. Each of these movement methods can be used in combination with each gravitation force calculation method, so that there are four models for the GELS algorithm.

In 2007, Blachandar [18] used the GELS algorithm to solve the Travelling Salesman Problem and compared it with other algorithms such as hill climbing and SA. The results showed that whenever the size of a problem is small, all algorithms perform roughly the same, but whenever the size of the problem is large, the GELS algorithm obtains better results than the other algorithms.

The algorithm begins with a primary response and a primary velocity vector that consists of a primary velocity specified by the user or randomly generated. After the primary velocity vector has been examined, the responsive dimension with the greatest primary velocity among the neighbor dimensions is selected for movement (select and obtain neighbor response).

The algorithm uses a pointer object that can move within the search space. This object always refers to the response with the most weight. In the first iteration of the algorithm using the first method, a dimension is selected for obtaining a neighbor response from the current response and a candidate response is selected from the local neighbor space of the current response in terms of this dimension. The gravitational force between the current and candidate responses is calculated and then added to the primary velocity of the dimension from which the candidate response was obtained. This is called the *updated primary velocity*. In the next iteration, the primary velocity vector is examined and a new movement direction is selected for continuing the response search. Each iteration of the algorithm using the second method is generally similar to the first method except that instead of calculating gravitational force and updating the primary velocity vector for just one candidate response in the current dimension, gravitational force is calculated and the primary velocity updated for each candidate response in the current dimension. In this algorithm, the gravitational force between two entities is calculated using Equation (1):

$$f = \frac{G(CU - CA)}{R^2} \tag{1}$$

where CA and CU are the candidate response and current response, respectively; G is the constant 6.672; and R is the neighbor radius of two parameters in the search space. R may be constant or can change intelligently in each iteration. The algorithm terminates when one of the following happens: either the primary velocity for all equal response dimensions (all elements of the primary velocity vector) are equal to zero or the maximum number of iterations of the algorithm has been reached [19].

Another parameter used in this algorithm is the maximum primary velocity. This parameter is the maximum value that can be used in the primary velocity vector. The primary velocity vector is used to select the movement direction for obtaining a neighbor, and this parameter prevents the move from increasing the primary velocity vector elements beyond a certain limit.

In [20], the authors tried to optimize the convergence speed of a GA with two changing points in the standard GA. After executing the crossover action, if the fitness value of the produced population is less than the average fitness or the best individual of the population, secondary preferential hybridization or mutation is also used after the primary mutation action.

Cruz-Chávez[21] proposed a hybrid genetic/annealing evolutionary algorithm for the independent task scheduling problem. The main purpose of this algorithm was to find the solution that minimizes the total runtime. GAs are weak for local searches, while SA is powerful for local searches. The authors combined these two methods to use both their abilities to search the problem space. The GA includes a stochastic population

generator, an elitism selection operator, and mutations and crossovers with the help of SA. Based on the fitness function, the selection operator selects the best half of the chromosomes in the population, the crossover is performed, and new children are produced for the next generation. Using a crossover leads to complete searches of the problem space. The iteration operation used as a mutation produces an optimized population, and a better population is found during the SA searching iteration. This process is repeated for each generation. It should be noted that thermal simulation techniques are performed on populations of individuals who have been run off.

In [22], some modifications of GAs are proposed to improve scheduling efficiency. These changes consist of the combination of the greedy algorithms, modified critical path (MCP) [23] and duplication scheduling heuristic (DSH) [24], with a GA to minimize the start time for tasks until, in the end, makespan is minimized. The algorithm also uses idle processor time. The algorithm has two fitness functions. The first function searches for chromosomes with the shortest makespan and the second function are designed to find the most appropriate chromosomes with respect to load balance.

In [25], aGA is presented in which chaotic variables are used instead of random variables for chromosome production. This leads to a distribution of solutions over the entire search space and avoids local minima, so that the best solutions and productions are obtained in a shorter time.

In [26], aGA is combined with the hill-climbing algorithm to repair chromosomes. This work modifies invalid individuals in each generation until they become valid individuals in a new population.

In [27], the GELS algorithm is used for resource reservation and independent task scheduling, so that in the objective function, if one resource can't execute a task within its specified deadline, the task is allocated to another resource for execution. Simulation results show that this algorithm decreases makespan compared to GAs. In previous methods, a decrease of the entire execution time was considered, while the number of tasks missing their deadlines and the load balance problem were not also considered. Our proposed algorithm tries to consider these three parameters simultaneously. Also, because a GA is weak in local searches, our proposed algorithm combines it with a local search algorithm to address this weakness. A combination of a GA and GELS is used because GELS searches the problem space well and finds better solutions compared to other local search algorithms such as hill-climbing and SA.

3 Scheduling Problem Description

The scheduling problem for independent tasks is an NP-hard problem that consists of N tasks and M machines. Each task should be considered to be processed by each of the M machines, so that the makespan is minimized. However, this only considers one of the QOS parameters, the time constraint, and ignores the cost. Therefore, we have introduced a deadline for every task such that each task should complete its execution before its deadline. Each task can be executed on only one resource and is not stopped before its execution is complete.

We use the expected time to compute the ETC matrix model described in [28]. Since our proposed scheduling algorithm is static, we assume that the expected execution time for each task i on each resource j has already been determined and has been set in the ETC matrix at ETC[i,j]. Also, the ready time (Ready [j]) for each machine j indicates when j has finished its previous task. The makespan is equal to the maximum complete time Completion_Time [i,j](Equation 2):

$$\text{makespan}=\text{Max}(\text{Completion\_Time}[i,j]) \quad \{1\le i\le N, 1\le j\le M\} \tag{2}$$

Completion_Time [i,j] is the time at which task i ends on resource j and is calculated according to Equation(3):

$$\text{Completion\_Time}[i,j]=\text{Ready}[j]+\text{ETC}[i,j] \tag{3}$$

The purpose of scheduling is to assign tasks to resources so that the final makespan and the number of tasks that miss their deadlines are minimized.

4 The Proposed Method (GGA)

The efficiency of genetic algorithms is highly dependent on how the chromosomes are represented. Here we use a simple method for representing chromosomes, in order to simplify the work of the crossover and mutation operators. Natural numbers are used for encoding the chromosomes. The numbers inside the genes are random numbers between 1 and M. The chromosome lengths are assumed to be task numbers. Figure 2 shows an example of the chromosome representation. For example, in the figure, task4—orT4—executes on Resource2.

T1	T2	T4	T3
1	3	2	4

Figure 2: Chromosome representation.

Initial Population: The initial population is created randomly. A source is selected randomly until the task being considered is executed on it. Each of the chromosomes produced is assumed to be a dimension of the problem (in fact, the problem's dimensions are just the neighbouring solutions that are obtained by changing the current solution). An initial random velocity is given to each of the problem's dimensions, ranging between one and the maximum velocity.

First Fitness Function: The basic purpose of task scheduling is to minimize makespan. This is the total time required until all of the input tasks complete their execution. It should be noted that this time should always be less than or equal to the maximum deadline among all the tasks. In our proposed method for task scheduling, a solution is more appropriate if in addition to decreasing makespan, it minimizes the number of tasks that miss their deadlines. Equation (4) calculates the first fitness function for each chromosome:

$$\mathrm{Fit}_1(\mathrm{ch}_i) = \frac{1}{\mathrm{makespan}(\mathrm{ch}_i)} + \frac{1}{\mathrm{miss\_task} * \mathrm{MD}} \quad (4)$$

Where miss_task is the number of tasks that have missed their deadlines in chromosome ch_i andMD is the maximum deadline for all tasks. As the equation shows, when the makespan and the number of tasks missing their deadlines are smaller, the fitness function value is greater, indicating the more promising chromosomes.

Second Fitness function: With respect to the basic purpose of task scheduling, minimizing makespan, several chromosomes may be found that have similar makespans but don't all balance the load among their resources. Hence, the second fitness function considers this factor after obtaining solutions with similar makespans, to find the most appropriate solution with respect to load balance.

If the execution time for resource R_j is E_time [R_j], the average execution time (avg) for all resources is as shown in Equation (5):

$$\mathrm{avg} = \sum_{j=1}^{\mathrm{num\_resources}} \left(\mathrm{E\_time}[R_j] \Big/ \mathrm{num\_resources} \right) \quad (5)$$

where num_resources is the number of resources. The load balance for resource i, $\mathrm{Cpu\_LB}_i$, can then be calculated with Equation (6):

$$\mathrm{Cpu\_LB}_i = \mathrm{makespan}/\mathrm{avg} \quad (6)$$

Equation (7)shows the second fitness function that considers the load balance:

$$\mathrm{Fit}_2(\mathrm{ch}_i) = \frac{1}{\mathrm{Cpu\_LB}_i} \quad (7)$$

Select an Operation: Before the mutation and crossover operators apply, the selection phase is first executed. In our proposed algorithm, we use the GELS algorithm instead of traditional genetic operators such as tournament, elitism, etc. These operators provide the possibility of creating the best solutions in each generation, but the GELS algorithm is used to select solutions because one chromosome may not initially have a good fitness value but turn out to be better after the mutation and crossover operations. Using the GELS algorithm, the two chromosomes that have a greater primary velocity are selected.

Crossover Operator: Our proposed algorithm uses a two- point crossover operator. Two points are selected randomly from chromosomes from the previous phase. Then all of the genes within these two points of the first and second chromosomes are removed.

Mutation Operator: A point on each chromosome from the previous phase is randomly selected and then changed to a random number between 1 and M.

Force Calculation: After applying the crossover and mutation operations, the gravitational force between the primary chromosome and the produced chromosome are calculated as in Equation (8). Then the gravitational force is added to the velocity of that dimension. This leads to no copying in the candidate population, if the produced chromosomes have worse fitness values than the primary chromosomes.

$$\mathrm{Force} = 6.672 * \left(\frac{\mathrm{Fit}_1(\mathrm{Candidate\_ch}_i)}{R^2} - \frac{\mathrm{Fit}_1(\mathrm{Current\_ch}_i)}{R^2} \right) \quad (8)$$

Terminating Conditions:The algorithm terminates when the primary velocity is equal to zero for all dimensions or the maximum number of algorithm iterations has been reached.

Algorithm1 shows the GGA pseudocode.

Algorithm 1 GGA Algorithm (pseudocode)

Input: Tasks Populations;
Output: Scheduled tasks based on Fit1 and Fit2;

```
1:  Generate K chromosomes to initialize the population
2:  Velocity_Vector[1..K]=Initial velocity for each
    Dimension();
3:  While (i<=max_iteration and Velocity_Vector[...]≠0)
    {
4:  /* select current_ch1 and current_ch2 such that the velocity
    is larger and generate two offspring, candidate_ch1 and
    candidate_ch2, by crossover and mutation*/
5:  If (Fit1(candidate_ch1) > Fit1(current_ch1) )
        current_ch1= candidate_ch1;
6:  If (Fit1(candidate_ch2)> Fit1(current_ch2))
        current_ch2= candidate_ch2;
7:  Calculate gravitational force between candidate_ch1 and
    candidate_ch2 using Equation (8)
8:  Update Velocity_Vector for each dimension by
    gravitational force of chromosome;
9:  }// end while
10: If many chromosomes with same Fit1exist
        Select Best chromosome using Fit2;
```

5 Performance Evaluation

Here, we explain the experimental descriptions.

5.1 Experimental Results

The GGA algorithm was implemented using Java software running under the Win XP operating system on a 2.66GHZ CPU with 4GB RAM. In our proposed algorithm, we assumed that the crossover rate CR =0.98 and the mutation rate MR =0.05.

The contents of the primary velocity vector for the chromosomes were randomly assigned. The results of

simulations comparing GGA with the GELS, GA, and GSA algorithms are shown in Figures 4, 5, and 6.

5.2 Experimental Results

Here, we have tested our work on various tasks; Generations and different fitness function orderly.

The diagram in Figure 3 shows a number of scheduled tasks ranging between 20 and 60 allocated to 20 resources using the comparison algorithms. As the figure shows, when the number of tasks increases, the makespan increases as well. The diagram shows that our proposed algorithm produces a smaller makespan than the other algorithms.

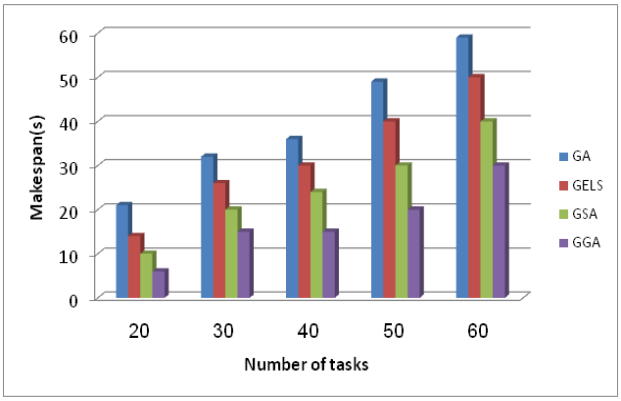


Figure 3: Comparison of make spans.

Figure 4 compares the algorithms for various numbers of iterations. It is clear that GGA, which is a combination of a globally searching GA and the local GELS algorithm, pays more attention to convergence velocity and optimization than the other algorithms, since unlike SA, the GELS algorithm doesn't have an absolute probability state.

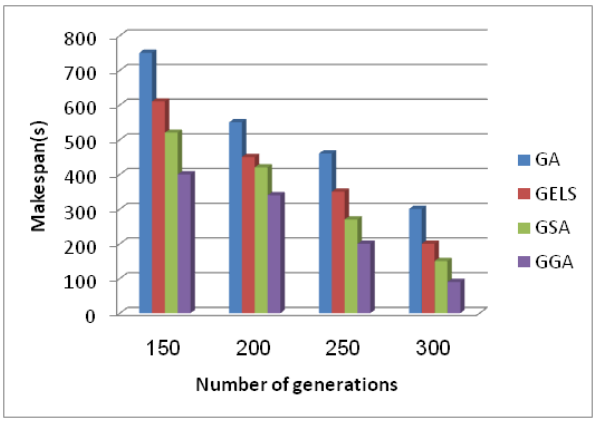


Figure 4: Comparison of evolutionary process for the different algorithms.

Figure 5 compares the algorithms with respect to the percentage of tasks that miss their deadlines. In this diagram, the fitness value is plotted against the rate of tasks missing deadlines. As the diagram shows, whenever the fitness value increases, the rate of tasks missing deadlines decreases. This means that the number of tasks missing deadlines decreases as a result of the completion of their makespan. The figure shows that fewer tasks miss their deadlines in the GGA algorithm than in the other algorithms.

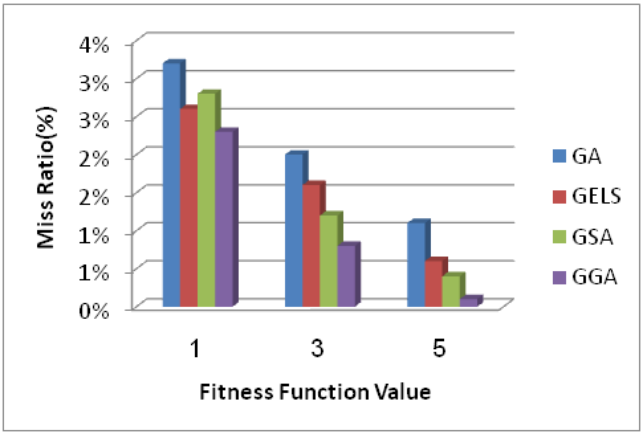


Figure 5: Comparison of the average missed deadline ratios for different fitness function values and algorithms.

6 Conclusions

This paper presented an algorithm for solving the grid task scheduling problem through a combination of a GA, which is a global search algorithm, and the GELS algorithm, which searches locally. The algorithm aims at minimizing makespan as well as the number of tasks that miss their deadlines. Local search algorithms such as hill climbing and SA always move to the solutions that have a better fitness function value, and they search the problem space randomly. Although the GELS algorithm shares the special behaviour of greedy algorithms, it doesn't always move directly to a solution with a better fitness function value but rather works by examining existing solutions. Although the GELS algorithm uses some random elements, it doesn't always move among them in the same way, which is why it doesn't stop with locally optimal solutions. By combining the advantages of the GELS algorithm and GAs, both the convergence velocity and the GA's identification of an optimal response are improved. We compared our proposed algorithm to other algorithms, and our simulation results showed that GGA produces smaller makespans than the other algorithms and also minimizes the number of tasks that miss their deadlines.

References

[1] J. Kołodzie and F. Xhafa, "Meeting security and user behavior requirements in Grid scheduling,"*Simulation Modeling Practice and Theory*vol.19, no. 1, pp. 213–226, 2011.

[2] W.T. Sullivan, D. Werthimer, S. Bowyer, J. Cobb, D. Gedye and D. Anderson, "A new major SETI project based on Project SERENDIP data and 100000 personal computers,"in *Proc. of the Fifth*

International Conference on Bioastronomy, no. 61, 1997.

[3] I. Foster, C. Kesselman, J. Nick and S. Tuecke, "the Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration,"*Computer*, 35 (6),pp. 1-4, 2002.

[4] B.Yan-ping, Zh.Wei and Y.Jin-shou, "An Improved PSO Algorithm and Its Application to Grid Scheduling Problem,"in *International Symposium on Computer Science and Computational Technology ISCSCT '08*, 2008, pp. 352-355.

[5] M. Shojafar, S. Barzegar and M.R. Meybodi, "A new Method on Resource Scheduling in grid systems based on Hierarchical Stochastic Petri net," in *Proc. of third International Conference on Computer and Electrical Engineering (ICCEE 2010)*, 2010, pp. 175-180.

[6] M. Shojafar, Z. Pooranian, J.H. Abawajy and M.R. Meybodi, "An Efficient Scheduling Method for Grid Systems Based on a Hierarchical Stochastic Petri Net," *Journal of Computing Science and Engineering (JCSE)*, 7(1), pp. 44-52, 2013.

[7] M. Shojafar, S. Barzegar and M.R. Meybodi, "Msc.Thesis: Time Optimizing in Economical Grid Using Adaptive Stochastic Petri Net Based on Learning Automata,"M.s.c. Thesis, Islamic Azad University of Qazvin, Qazvin, Iran,September 2010.

[8] M. Shojafar, S. Barzegar, and M. R. Maybodi, "Time optimizing in Economical Grid Using Adaptive Stochastic Petri Net Based on Learning Automata", in *Proc. of International Conference on Grid Computing & Applications(GCA)*, WORLDCOMP, 2011, pp. 67-73.

[9] G. Falzon and M. Li, "Enhancing genetic algorithms for dependent job scheduling in grid computing environments,"*The Journal of Supercomputing*, Springer,62(1), pp. 290–314, 2012.

[10] Z. Pooranian, M. Shojafar, J.H. Abawajy and M. Singhal, "GLOA: A new Job Scheduling Algorithm for Grid Computing," *International Journal of Artificial Intelligence and Interactive Multimedia (IJIMAI)*, 2(1), pp. 59-64, 2013.

[11] W. Abdulal and S. Ramachandram, "Reliability-Aware Scheduling Based on a Novel Simulated Annealing in Grid," in *Proc. in Fourth International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 665-670. 2012.

[12] Z. Pooranian, M. Shojafar and B. Javadi, "Independent Task Scheduling in Grid Computing Based on Queen Bee Algorithm," *IAES International Journal of Artificial Intelligence (IJ-AI)*, 1(4), pp. 171-181, 2012.

[13] F. Xhafaa and A. Abraham, "Computational models and heuristic methods for Grid scheduling problems," *Future Generation Computer Systems*,26(4),pp. 608–621, 2010.

[14] Z. Pooranian, A. Harounabadi, M. Shojafar, J. Mirabedini, "Hybrid PSO for Independent Task scheduling in Grid Computing to Decrease Makespan," in *Proc. of International Conference on Future Information Technology,IPCSIT'11*, vol. 13, 2011, pp. 435-439.

[15] J. Holland, "Adaptation in Natural and Artificial Systems," University of Michigan Press, Ann Arbor, ISBN: 0-262-58111-6, 1975.

[16] C. Voudouris and T. Edward,"Guided Local Search. Technical Report CSM-247," Department of Computer Science, University of Essex, UK, August 1995.

[17] L. W. Barry, "Solving Combinatorial Optimization Problems Using a New Algorithm Based on Gravitational Attraction," Ph.D. Thesis, Florida Institute of Technology Melbourne, FL, USA, May 2004.

[18] S. R. Balachandar and K. Kannan, "Randomized gravitational emulation search algorithm for symmetric traveling salesman problem," *Applied Mathematics and Computation*, 192(2), pp. 413–421, 2007.

[19] J. Li, Y. Lou and Y. Shi, "An Optimization Algorithm Based on Binary Difference and Gravitational Evolution, "International *Journal of Computational Intelligence Systems*,5(3), pp. 483-493,2012.

[20] X. ZhangandW. Zeng, "Grid Workflow Scheduling Based on Improved Genetic Algorithm,"in *Proc. of International Conference on Computer Design and Applications (ICCDA 2010)*, 2010, pp. 270-273.

[21] M.Cruz-Chavez, A.Rodriguez-Leon, E.Avila-Melgar, F.Juarez-Perez, M.Cruz-Rosales and R. Rivera-Lopez,"Genetic-Annealing Algorithm in Grid Environment for Scheduling Problems,"*Security-Enriched Urban Computing and Smart GridCommunications in Computer and Information Science, Springer*, vol. 78, 2010,pp. 1-9.

[22] F.A. Omaraa and M.M. Arafa, "Genetic algorithms for task scheduling problem,"*Journal Parallel Distributed Computing*, Elsevier, 70(1), pp. 13-22, 2010.

[23] M. Wu and D.D. Gajski, "Hyper tool: A programming aid for message-passing systems,"*IEEE Transactions on Parallel and Distributed Systems*, 1(3), pp. 330-343,1990.

[24] A.P. Engelbrech, "Fundamentals of computational swarm intelligence,"John Wiley & Sons Inc.,2005.

[25] G. Gharoonifard,F. Moeindarbari, H. Deldari and A. Morvaridi, "Scheduling of scientific workflows using a chaos- genetic algorithm,"in *Proc. of International Conference on Computational Science ICCS2010*, 2010, pp. 1439-1448.

[26] A. Lifeng and T. Maolin, "QoS-Based Web Service Composition Accommodating Inter-Service Dependencies Using Minimal-Conflict Hill-Climbing Repair Genetic Algorithm," in *Proc. of Fourth IEEE International Conference on Science*, 2008, pp. 119-126.

[27] B. Barzegar, A.M. Rahmani and K. Zamanifar, "Gravitational Emulation Local Search Algorithm

for Advanced Reservation and Scheduling in Grid Systems," in *Proc. of First Asian Himalayas International Conference on (2009)*, 2009,pp. 1-5.

[28] T.D. Braun., H.J. Siegel, N. Beck, L.L. Boloni, M. Maheswaran, A.L. Reuther, J.P. Robertson, M.D. Theys, B. Yao, D. Hensgen and R.F. Freund, "A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems,"*Journal of Parallel and distributed Computing*, 61(6), pp. 680- 1983, 2001.

[29] S.S. Kim, J.H. Byeon, H. Liu, A. Abraham and Sean McLoone, "Optimal job scheduling in grid computing using efficient binary artificial bee colony optimization, *Soft Computing*," 17(5), pp. 867-882, 2013.

[30] H. Liu, A. Abraham and A. Hassanien, "Scheduling Jobs on Computational Grids Using Fuzzy Particle Swarm Algorithm," *Future Generation Computing Systems, Elsevier Science*, 26 (8), pp. 1336-1343, 2010.

[31] H. Izakian, B.T. Ladani, A. Abraham and V. Snasel, "A Discrete Particle Swarm Optimization Approach for Grid Job Scheduling," International Journal of Innovative Computing, Information and Control, 6(9), pp. 4219-4233, 2010.

[32] H. Izakian, A. Abraham and V. Snasel, "Performance Comparison of Six Efficient Pure Heuristics for Scheduling Meta-Tasks on Heterogeneous Distributed Environments," *Neural Network World*, 19(6), pp. 695-710, 2009.

[33] H. Liu, A. Abraham and Z. Wang, "A Multi-swarm approaches to Multi-objective Flexible Job-shop Scheduling Problems," *Fundamental Informaticae Journal, IOS Press*, Netherlands, 95(4), pp.465-489, 2009.

[34] H. Izakian, A. Abraham and V. Snasel, "Metaheuristic Based Scheduling Meta-Tasks in Distributed Heterogeneous Computing Systems," *Sensors, Molecular Diversity Preservation International Switzerland*, 9(7), pp. 5339-5350, 2009.

[35] A. Abraham, H. Liu and M. Zhao, "Particle Swarm Scheduling for Work-Flow Applications in Distributed Computing Environments, Metaheuristics for Scheduling: Industrial and Manufacturing Applications," *Studies in Computational Intelligence, Springer Verlag, Germany*, ISBN 978-3-540-78984-0, pp. 327-342, 2008.

[36] A. Abraham, H. Liu, C. Grosan and F. Xhafa, "Nature Inspired Metaheuristics for Grid Scheduling: Single and Multi objective Optimization Approaches, Metaheuristics for Scheduling: Distributed Computing Environments," *Studies in Computational Intelligence, Springer Verlag, Germany*, ISBN: 978-3-540-69260-7, pp. 247-272, 2008.

CroNER: Recognizing Named Entities in Croatian Using Conditional Random Fields

Mladen Karan, Goran Glavaš, Frane Šarić, Jan Šnajder, Jure Mijić, Artur Šilić and Bojana Dalbelo Bašić
University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
E-mail: takelab@fer.hr and http://takelab.fer.hr

Keywords: named entity recognition, conditional random fields, natural language processing, information extraction, Croatian language

Received: February 27, 2013

In this paper we present CroNER, a named entity recognition and classification system for Croatian language based on supervised sequence labeling with conditional random fields (CRF). We use a rich set of lexical and gazetteer-based features and different methods for enforcing document-level label consistency. Extensive evaluation shows that our method achieves state-of-the-art results (MUC F1 90.73%, Exact F1 87.42%) when compared to existing NERC systems for Croatian and other Slavic languages.

*Povzetek: V pričujočem prispevku je predstavljen CroNER, sistem za prepoznavanje in klasifikacijo imenskih entitet za hrvaščino, ki temelji na nadzorovanemu označevanju s pomočjo pogojnih naključnih polj (*conditional random fields – *CRF).*

1 Introduction

Named Entity Recognition and Classification (NERC) is a well-known natural language processing (NLP) and Information Extraction (IE) task. NERC aims to extract and classify all names (*enamexes*), temporal expressions (*timexes*), and numerical expressions (*numexes*) appearing in natural language texts. The classes of named entities typically extracted by NERC systems are names of people, organizations, and locations as well as dates, temporal expressions, monetary expressions, and percentages.

In this paper we present CroNER, a supervised NERC for the Croatian language. We use sequence labeling with conditional random fields (CRF) [13] to extract and classify named entities from newspaper text. We use a rich set of features, including lexical and gazetteer-based features, with many of them incorporating morphological and lexical peculiarities of the Croatian language. We implemented two different methods for document-level consistency of NE labels: postprocessing rules (hard consistency constraint) and a two-stage CRF (soft consistency constraint). Postprocessing rules are hand-crafted patterns designed to extract or re-label named entities omitted or misclassified by the CRF model. Two-stage CRF [12] aims to consolidate NE label predictions on document and corpus level by employing a second CRF model that uses features computed from the output of the first CRF model. We evaluate the performance of the system using standard MUC and Exact NERC evaluation schemes [19].

The rest of the paper is structured as follows. In Section 2 we present related work on named entity extraction for Croatian and other Slavic languages. Section 3 discusses the details of corpus annotation. In Section 4 we thoroughly describe the feature set and the extensions used (rule-based postprocessing and two-stage CRF). Section 5 presents experimental setup and evaluation results. In Section 6 we conclude and outline future work.

2 Related work

Identifying references to named entities in text was recognized as one of the important subtasks of IE, and it has been a target of intense research for the last twenty years. The task was formalized at the Sixth Message Understanding Conference in 1995 [10]. There is a large body of NERC work for English [18, 17, 6, 12] and other major languages [7, 26, 4, 22]. Substantially less research has targeted Slavic (especially South Slavic) languages; NERC systems have been reported for Russian [23], Polish [21, 16], Czech [11], and Bulgarian [5, 9]. In [9] it was shown that CRF-based NERC with a rich set of features outperforms all other methods for Bulgarian, as well as other Slavic languages.

The rule-based system from [2], which uses a cascade of finite-state transducers, is the only reported work on NERC for Croatian language. In [15] a method for generating a morphological lexicon of organizational names was proposed, a valuable resource for morphologically rich languages. We used a similar approach to expand morphological lexica with inflectional forms of Croatian proper names, but we include first names, surnames, and toponyms in ad-

dition to organization names.

To the best of our knowledge, we are the first to use supervised machine learning for named entity recognition and classification in Croatian language. Using a machine learning method, we avoid the need for specialized linguistic knowledge required to design a rule-based system. This way we also avoid the explicit modelling of complex dependencies between rules and their application order. We instead focus on designing a rich set of features and let the CRF algorithm uncover the dependencies between them.

3 Corpus annotation

The training and testing corpus consists of 591 news articles (about 310,000 tokens) from the Croatian newspaper *Vjesnik*, spanning years 1999 to 2009. The preprocessing of the corpus involved sentence splitting and tokenization. For annotation we used seven standard MUC-7 types: *Organization*, *Person*, *Location*, *Date*, *Time*, *Money*, and *Percent*. We also introduced five additional types: *Ethnic* (names of ethnic groups), *PersonPossessive* (possesive adjectives derived from person names), *Product* (names of branded products), *OrganizationAsLocation* (organization names used as metonyms for locations, as in *"The entrance of the PBZ bank building"*), and *LocationAsOrganization* (location names used as metonyms for organizations, as in *"Zagreb has sent a demarche to Rome"*). The additional types were introduced for experimental reasons; in this work only the *Ethnic* tag was retained, while other additional tags were not used (i.e., the *Product* tag was discarded, while the remaining three subtype tags were mapped to the corresponding basic tags). Thus, in the end we trained our models using eight types of named entities.

The annotation guidelines we used are similar to MUC-7 guidelines, with some adjustments specifically for the Croatian language. The corpus was independently annotated by six annotators. To ensure high annotation quality, the annotators were first asked to independently annotate a calibration set of about 10,000 tokens. On this set, all the disagreements have been resolved by consensus, the borderlines were discussed, and the guidelines revised accordingly. Afterwards, each of the remaining documents was annotated by two independent annotators, while a third annotator resolved the disagreements. For annotating we used an in-house developed annotation tool.

The inter-annotator agreement (calculated in terms of MUC F1 and Exact F1 score and averaged over all pairs of annotators) is shown in Table 1. The inter-annotated was measured on a subset of about 10,000 tokens that was annotated by all six annotators. Notice that the overall quality of the annotations improved after resolving the disagreements, but – because each subset was resolved by a single annotator – we cannot objectively measure the resulting improvement in annotation quality.

Table 1: Inter-annotator agreement

Tag	F1 Exact	F1 MUC
Person	98.05	98.55
Ethnic	97.19	97.19
Percent	92.00	96.77
Location	93.95	94.93
Money	91.95	94.15
Organization	89.35	93.58
Date	71.47	85.79
Time	67.55	71.04

4 CroNER

CroNER is based on sequence labeling with CRF. We use the CRFsuite [20] implementation of CRF. At the token level, named entities are annotated according to the Begins-Inside-Outside (B-I-O) scheme, often used for sequence labeling tasks. Following is a description of the features used for sentence-level label prediction and the techniques for imposing document-level label consistency.

4.1 Sentence-level features

Most of the features can be characterized as lexical, gazetteer-based, or numerical. Some of the features were *templated* on a window of size two, both to the left and to the right of the current word. This means that that the feature vector for the current word consists of features for this word, two previous words, and two following words.

Lexical features. The following is the list of the lexical features used (templated features are indicated as such).

1. Word, lemma, stem, and POS tag (*templated*) – For lemmatization we use the morphological lexicon described in [25]. For stemming, we simply remove the word's suffix after the last vowel (or the penultimate vowel, if the last letter is a vowel). Words shorter than 5 letters are not stemmed. For POS tagging, we use a statistical tagger with five basic tags.
2. Full and short shape of the word – describe the ordering of uppercased and lowercased letters in the word. For example, *"Zagreb"* has the shape *"ULLLLL"* and short shape *"UL"*, while *"iPhone"* has the shape *"LULLLL"* and short shape *"LUL"*.
3. Sentence start – indicates whether the token is the first token of the sentence.
4. Word ending – the suffix of the word taken from the last vowel till the end of the word (or the penultimate vowel, if the last letter of the word is a vowel).
5. Capitalization and uppercase (*templated*) – indicates whether the word is capitalized or entirely in uppercase (e.g., an acronym).

6. Acronym declension – indicates whether the word is a declension of an acronym (e.g., *"HOO-om"*, *"HDZ-a"*). Declension of acronyms in Croatian language follows predictable patterns [1].
7. Initials – indicates whether a token is an initial, i.e., a single uppercase letter followed by a period.
8. Cases – concatenation of all possible cases for the word, based on morpho-syntactic descriptors (MSDs) from the morphological lexicon. If the word has two or more MSDs with differing cases, we concatenate them in alphabetical order. We also add one Boolean feature for each individual case (*isNominative*, *isGenitive*, *isDative*, *isAcusative*, and *isInstrumental*).
9. Bigram features – concatenations of the previously described features computed for two consecutive tokens: *word bigram*, *lemma bigram*, *POS bigram*, *shape bigram*, and *cases bigram*.
10. Lemmas in window – all lemmas within a symmetric window of size 5 from the current token.
11. MSDs in window – all MSDs of the words within a symmetric window of size 5 from the current token.

Gazetteer-based features. Information about the presence of named entities from predefined gazetteers has been shown to be an important information for NERC [19]. We use several gazetteers: first names, surnames, ethnics, organizations, cities, streets, and countries gazetteers. The last four gazetteers have multi-word entries. The following is a list of gazetteer-based features.

1. Gazetteer match – indicates whether the lemma matches a gazetteer entry (used for gazetteers with single-word entries: names, surnames, and ethnics).
2. Starts gazetteer match – indicates whether there is any sequence of words starting with the current word that fully matches a gazetteer entry. E.g., in *"usluge Zavoda za javno zdravstvo"* (*services of the Public Health Department*), the word *"Zavoda"* would have this feature set to *true* because the organizations gazetteer contains *"Zavod za javno zdravstvo"*.
3. Stemmed gazetteer match – similar to the previous feature, but considers stems instead of lemmas. This feature is used only for the organizations gazetteer.
4. Gazetteer match length – the length (number of words) of the gazetteer entry whose first token matches the current token (e.g., for token *"Zavod"* in text *"usluge Zavoda za javno zdravstvo"*, the length would be 4).
5. Inside gazetteer match – indicates whether a word is inside the phrase that matches a gazetteer entry (e.g., true for tokens *"za"*, *"javno"*, and *"zdravstvo"* in organization entry *"Zavod za javno zdravstvo"*).

Both the text and the gazetteer entries were lemmatized before looking for matches. As gazetteers predominantly contain proper nouns, we needed to extend the morphological lexicon with the inflectional forms of proper names. We did this automatically with a set of rules following the paradigms for proper names declension [1]. We expanded both Croatian and foreign proper names.

Some simple preprocessing steps were applied for all gazetteers. All entries containing non-alphabetic characters were removed. We considered all words with more than 10% non-capitalized occurrences in the corpus to be common words and removed such entries. The rationale was to eliminate common word entries from the gazetteers in order to reduce the noise in the training set. For example,*"Luka"* is a very common personal name, but also a frequent common noun (*port*). Capitalization frequencies required for the above analysis were gathered from the *Vjesnik* corpus, a collection of 270,000 newspaper articles.

The major source of the Croatian names and surnames was the Croatian telephone directory. For English names, we used Stanford NER[1] to extract names from the NYT corpus[2] and Wikipedia. The compiled gazetteers for personal names and surnames contain 13,618 Croatian first names, 64,240 Croatian surnames, 70,488 foreign first names, and 228,134 foreign surnames. For locations we use three gazetteers – for streets, countries and cities. The street names (52,593 entries) were extracted from the Croatian telephone directory. Country names in Croatian (276 entries) were obtained from Wikipedia. The cities gazetteer (289,707 entries) was constructed using the telephone directory and internet sources. The organizations gazetteer (3035 entries) was created from several different sources, and includes names of institutions (e.g., *Ministry of Science*, *Louvre*), political parties (e.g., SDP, HDZ), international organizations (e.g., *UNESCO*, *NATO*), local and foreign companies, newspaper names, and sports teams. Finally, we compiled the ethnics gazetteer (940 entries) automatically from country names using the appropriate rules of Croatian grammar [1].

Numerical features. We used the following features to deal specifically with numbers (occurring in numexes and timexes):

1. Integer or decimal number – indicates whether the word is an integer or a decimal number;
2. Two/four digit integer – indicates whether the token is a two digit (useful for recognizing numexes) or a four digit integer (useful for recognizing years in dates);
3. Number followed by a period – indicates whether the token is an integer followed by a period (a good clue for dates and currencies);
4. Currency – indicates whether a token is a currency marker (e.g., *"$"* or *"EUR"*). We compiled a currency gazetteer that includes all major world currencies.

[1] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[2] The New York Times Annotated Corpus, (2008), LDC.

4.2 Document-level consistency

The CRF model predicts the sequence of B-I-O labels on the sentence level. It is therefore possible to have at the document level differing labels for the same named entity. The goal of the document-level label consistency postprocessing is to unify the labels of named entities on the document level. We experimented with incorporating document-level consistency into our model as both soft constraints (two-stage CRF) and hard constraints (handcrafted postprocessing rules).

Two-stage CRF. The two-stage CRF [12] is a model that accounts for non-local dependencies between named entities. The main idea is to employ a second CRF that uses both local features (same features the first CRF uses) and non-local features computed on the output of the first CRF. We use three document-level features computed from the output of the first CRF:

1. The most frequent lemma label – the most frequent label assigned to a given lemma in the document (e.g., *B_Person* or *I_Organization*);
2. The most frequent NE label – the most frequent label assigned to a given NE mention in the document;
3. The most frequent superentity label – a superentity is a mention of the same entity that contains two or more tokens (e.g., *"Ivan Horvat"* vs. *"Horvat"*, or *"Zavod za javno zdravstvo"* vs. *"Zavod"*). This feature represents the most frequent label assigned to all the superentities of a given entity within the document.

Postprocessing rules (PPR). We created two sets of postprocessing rules: one to enforce document-level consistency (hard constraint) and another one to improve the recall on numexes and timexes. The rules for enforcing document-level label consistency work as follows. First, we collect all the different named entities recognized by the CRF model and identify the most frequent label assigned to each of them. Then we correct (i.e., re-label) NE instances that were assigned a different label from the most frequently assigned one. In the second step, we search for the potential false negatives (i.e., mentions of named entities from the collection that were omitted by the CRF model). If found, omitted mentions are also assigned the most frequent label for the corresponding named entity.

The rules for improving the recall for numexes are in fact token-level regular expressions. For currencies and percentages the rules are defined as follows:

1. $[num][num|prep|conj]^*[currencyMarker]$ – the currency expression starts with a number, followed by either numbers, prepositions, or conjunctions, and ends with a currency clue. When written in words, numbers often contain conjunctions. E.g., in *"trideset i pet"* (*thirty five*), word *"i"* is a conjunction. Ranges are often expressed using prepositions; e.g., *"30 do 50 milijuna kuna"* (*30 to 50 million kuna*);
2. $[num][num|prep|conj]^*[percentClue]$ – the rule for percentages is similar to the rule for currencies, except for requiring that the phrase ends with a percent clue (*"posto"* or *"%"*) instead of a currency marker.

For timex (time) class we use the following three rules:

1. $[\text{u}][number][timeword]$ – captures phrases like *"u 12.30 sati"* (*at 12.30 o'clock*), where *number* is an appropriately formatted number and *timeword* is a word from a predefined list of time-related words, e.g., *"sati"* (*o'clock*);
2. $[mod]?[preposition]?[daytimeword][mod]?$ – captures phrases like *"rano u jutro"* (*early in the morning*). Here *mod* represents a modifying word, e.g.. *"rano"* (*early*);
3. $[modGen][daytimeword]$ – captures phrases like *"tijekom podneva"* (*during the afternoon*), where *modGen* is a modifier that governs a noun in genitive case; e.g., *"prije"* (*before*).

5 Evaluation

We measured the performance of four different models: single CRF (1-CRF), two-stage CRF (2-CRF), single CRF with postprocessing rules (1-CRF + PPR), and two-stage CRF with postprocessing rules (2-CRF + PPR). In Tables 2 and 3 we report the performance in terms of precision, recall, and F1 for MUC (allows for extent overlap instead of an exact extent match) and Exact (requires that both extent and class match) evaluation schemes [19], respectively. Results are reported separately for each NE class. We also report both micro- and macro-averaged overall performance for each of the four models. The results were obtained with 10-fold cross validation on the entire annotated corpus.

5.1 Result analysis

Regarding the enamex classes, the performance for organizations is significantly (5–7%) worse than for persons and locations. This is expected, because in Croatian many organization instances are multi-word expressions, whereas person and location mentions more often consist of only one or two words. The lower inter-annotator agreement (cf. Table 1) for organizations supports this assumption.

The results show that 2-CRF outperforms 1-CRF consistently on main enamex classes (*Person*, *Organization*, and *Location*); the improvement is between half a point (*Location*) and a full point (*Organization*). The 1-CRF + PPR model similarly outperforms 1-CRF (e.g., 0.8 point increase for *Person*). However, the 2-CRF + PPR model brings negligible gain when compared to either 2-CRF or 1-CRF + PPR (on average 0.1 point for enamex classes). This indicates that both the second stage CRF and postprocessing rules ensure document-level consistency in a similar fashion, hence combining them does not lead to significant performance improvements.

Table 2: CroNER MUC evaluation results

	1-CRF			2-CRF			1-CRF + PPR			2-CRF + PPR		
NE Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Person	91.31	92.12	91.71	91.76	93.26	92.50	91.13	93.58	92.34	91.62	93.68	**92.64**
Location	89.27	89.77	89.52	89.83	90.30	**90.06**	88.30	91.00	89.63	89.00	90.46	89.72
Organization	88.15	81.65	84.78	88.66	82.94	**85.71**	85.51	84.74	85.13	86.43	84.11	85.25
Ethnic	96.82	90.56	93.59	97.73	90.55	94.01	97.74	90.56	94.01	98.29	90.56	**94.27**
Date	93.72	82.35	87.67	93.48	82.02	87.38	93.55	83.05	**87.99**	93.56	82.47	87.67
Time	91.86	50.22	64.94	91.74	49.33	64.16	76.96	78.67	77.80	77.06	79.11	**78.07**
Currency	99.54	87.30	93.02	99.32	88.10	93.37	99.20	99.20	**99.20**	99.20	99.20	**99.20**
Percent	100.00	96.43	98.18	100.00	96.21	98.07	99.54	97.77	**98.65**	99.54	97.77	**98.65**
Overall Micro	90.67	87.21	88.91	91.07	87.99	89.51	89.48	89.43	89.45	90.09	89.09	**89.59**
Overall Macro	93.84	83.80	88.78	94.06	84.08	88.79	91.49	89.82	90.65	91.83	89.67	**90.73**

Table 3: CroNER Exact evaluation results

	1-CRF			2-CRF			1-CRF + PPR			2-CRF + PPR		
NE Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Person	89.42	90.22	89.81	89.92	91.38	**90.64**	89.06	91.46	90.24	89.62	91.64	90.62
Location	87.60	88.09	87.84	88.11	88.57	**88.34**	86.58	89.21	87.87	87.34	88.74	88.03
Organization	80.79	74.83	77.70	81.05	75.82	**78.35**	77.26	76.94	77.10	78.58	76.57	77.56
Ethnic	96.82	90.56	93.59	97.74	90.56	94.01	97.73	90.56	94.01	98.29	90.56	**94.27**
Date	86.19	75.73	80.62	85.98	75.44	80.37	85.73	76.10	**80.63**	85.95	75.77	80.54
Time	87.80	48.00	62.07	88.43	47.55	61.85	66.08	67.55	66.81	66.23	68.00	**67.10**
Currency	95.93	84.13	89.64	95.75	84.92	90.01	96.45	97.22	**96.84**	96.27	97.22	96.74
Percent	95.60	92.19	93.86	95.82	92.19	93.97	98.86	97.09	**97.97**	98.86	97.10	**97.97**
Overall Micro	86.84	83.53	85.15	87.19	84.24	**85.69**	85.30	85.36	85.33	86.08	85.17	85.62
Overall Macro	90.00	80.47	84.97	90.35	80.80	85.31	87.21	85.76	86.49	87.64	87.20	**87.42**

For numexes, the second CRF model seems not to improve the performance, whereas the postprocessing rules significantly improve the performance. This improvement is to be attributed to the use of extraction rules for numexes, implying that document-level consistency is not an issue for numexes. Postprocessing rules for currencies and percents increase the recall and keep the precision on the same level. For temporal expressions, however, increase in recall is accompanied by a proportional decrease in precision. Deeper inspection reveals that this is mostly due to inconsistent annotations of timexes, as confirmed by the very low inter-annotator agreement for these classes (cf. Table 1).

As expected, Exact evaluation results are generally lower than MUC results. However, for most classes the decrease in performance is not significant. Exceptions to this are *Organization*, *Date*, and *Time* classes, for which the decrease in performance is 7%, 7%, and 11%, respectively. Many organization instances consist of four or more words, and in such cases our models – though able to recognize the mention – often fail to exactly match its extent. The most common errors include omitting the last word or adding an extra word at the end. The performance on the three mentioned classes is also limited by the annotation quality; these classes are in fact the ones on which human annotators agreed the least (cf. Table 1).

Table 4 shows the performance of the best-performing model (2-CRF + PPR) depending on the size of the training set. (25%, 50%, 75%, and 100% of the training data). Expectedly, the performance generally improves as the size of the training set increases. However, the improvement from using 75% data to using 100% data is relatively small, suggesting that no significant increase in performance could be gained from annotating a larger corpus.

5.2 Discussion

Unfortunately, our results are not directly comparable to other reported results because of the differences in (1) language (though very similar, all Slavic languages have their own peculiarities), (2) NE types (e.g., some use only four classes: *Person*, *Location*, *Organization*, and *Miscellaneous*), or (3) evaluation methodology (non-adherence to standard evaluation methodology, such as in the work from [2]). Nonetheless, the comparison might still be informa-

Table 4: CroNER performance depending on the size of the training set (CRF-2 + PPR)

Evaluation	Size (tokens)	Person	Loc.	Org.	Ethnic	Date	Time	Curr.	Perc.	Micro	Macro
MUC	25% (75k)	92.51	82.69	79.95	92.30	79.46	78.74	100.00	98.99	86.01	88.08
	50% (155k)	92.56	87.56	82.60	93.70	85.01	76.40	99.62	98.64	88.05	89.51
	75% (230k)	92.19	88.81	85.00	94.87	87.30	76.84	99.59	98.77	89.07	90.42
	100% (310k)	92.64	89.72	85.25	94.27	87.67	78.07	99.20	98.65	**89.59**	**90.73**
Exact	25% (75)	90.17	79.50	69.53	92.30	71.57	59.84	96.97	98.32	80.65	82.28
	50% (155k)	90.59	85.04	73.66	93.70	76.47	62.17	97.51	97.74	83.35	84.61
	75% (230k)	90.06	86.71	77.25	94.87	79.45	65.40	97.24	97.84	84.80	86.10
	100% (310k)	90.62	88.03	77.56	94.27	80.54	67.10	96.74	97.97	**85.62**	**87.42**

tive to some extent. In [2], a 79% F1-score on persons, 89% on organizations, and 95% on locations is reported, although it must be noted that for the latter two classes the evaluation was limited to selected subsets of NE instances. Our results seem to be better than those reported for other Slavic languages: Polish – 82.4% F1, [21], Czech – 76% F1, Russian – 70.9% F1 [23]. Only the best reported results for Bulgarian are comparable to our results: 89.6% overall F1, persons 92.79%, locations 90.06%, organizations 89.73% [9]. These comparisons suggest that CroNER is a state-of-the-art NERC system when considering the Slavic languages.

5.3 Experiments with distributional features

It has been demonstrated [8, 24, 7] that NERC can benefit from distributional modelling of lexical semantics. Distributional semantics is based on the hypothesis that semantically similar words occur in similar contexts, therefore the meaning of a word can be represented by its context. Distributional representations can be used to compare and cluster together similar words, improving NERC performance. To determine if this also holds in our case, we performed preliminary experiments with semantic cluster features.

To obtain semantic word clusters we use Brown's algorithm [3, 14]. The algorithm takes as input a sequence of words (a corpus) and outputs semantic clusters for each word. The number of clusters k is a parameter of the algorithm. The algorithm works by assuming the probability of a word sequence is given as follows:

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i|C(w_i))p(C(w_i)|C(w_{i-1})) \quad (1)$$

where $C(w_i)$ is the cluster to which the i-th word w_i is assigned. It is assumed that the probability of an occurrence of a particular word w_i at position i depends only on its cluster $C(w_i)$, which, in turn, depends only on the cluster of the previous word $C(w_{i-1})$. The quality of a clustering is measured by how well the classes of adjacent words in the sequence predict each other. This is achieved by maximizing the log probability of the input sequence given by (1). In [14] it was demonstrated that this optimisation is equivalent to maximizing the sum of mutual information

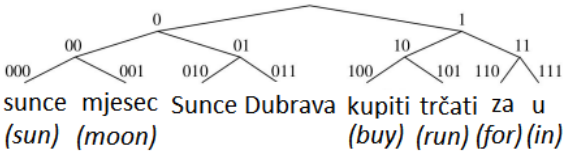


Figure 1: A clustering example for Croatian words (*"Sunce"* and *"Dubrava"* are proper names)

weights between all pairs of classes, and presented an efficient algorithm for computing the clusters.

In order to generate the sequence required as input to the algorithm, we took a sample from the HrWaC corpus (it was not possible to use the entire corpus due to its size). The sample consists of texts from three large internet news portals: *monitor.hr*, *slobodnadalmacija.hr*, and *vecernji.hr*. We chose news portals because they are of the same genre as our training and test data. Additionally, we expect the language used in news portals to be more standard and clean. The chosen texts were tokenized and lemmatized. The final input set for the algorithm had 351M tokens. To compute the clustering we used the freely available implementation from [14] with k set to 100. As a result, we obtained classes for each word as a bit string. The bit strings represent paths to each word in a binary tree whose leaves are clusters. An example of a good clustering is given in Fig. 1. An interesting property of this clustering is that we can control the generality of the clustering by looking only at a fixed length prefix of the bit string (e.g., it has been noted that prefixes of length four often correspond to POS tags).

We use clusters (in form of bit strings) as additional features for the CRF model. For each word w_i there are five features representing distributional clusters of words w_{i-2} to w_{i+2}. The number of possible distinct values for each of these features equals the number of clusters (100 in our case). We use the same procedure to include information about cluster prefixes of length two and four; in these cases the number of possible distinct values is smaller than the total number of clusters because all clusters beginning with the same prefix are merged. This approach is along the

Table 5: Comparison of CroNER performance with and without distributional features

Evaluation	Model	Person	Loc.	Org.	Ethnic	Date	Time	Curr.	Perc.	Ov. Macro
MUC	2-CRF + PPR	92.35	89.31	83.88	95.53	87.69	79.40	99.53	98.57	89.78
	2-CRF + PPR + dist.	93.37	89.17	85.21	95.46	87.26	78.5	99.4	98.57	**89.86**
Exact	2-CRF + PPR	90.20	87.82	76.81	95.53	81.19	67.15	96.93	98.0	86.00
	2-CRF + PPR + dist.	91.33	87.61	77.49	95.46	80.93	67.29	97.64	98.04	**86.25**

lines of the one proposed in [24].

Table 5 gives a comparison of performance with and without using distributional features, averaged over five cross validation folds on the entire data set. The use of distributional features leads to consistent improvements for Person and Organization classes. However, results for some of the other classes showed slight deterioration. This suggest that the distributional features are beneficial, but further experiments (with respect to the number of clusters and corpus size/choice) are required.

6 Conclusion and future work

We have presented CroNER, a NERC system for Croatian based on sequence labeling with CRF. CroNER uses a rich set of lexical and gazetteer-based features achieving good recognition and classification results. We have shown how enforcing document-level label consistency (either through postprocessing rules or a second CRF model capturing non-local dependencies) can further improve NERC performance. The experimental results indicate that, as regards the Slavic languages, CroNER is a state-of-the-art named entity recognition and classification system.

The work presented here could be extended in several ways. First, the annotated set should be revised, considering that the inter-annotator agreement is rather low on some classes. Secondly, a systematic feature selection (e.g., wrapper feature selection) may be performed in order to select an optimal subset of features. Thirdly, we plan to employ classification using more fine-grained NE labels. Finally, we intend to further explore the use of distributional semantic features.

7 Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under Grant 036-1300646-1986.

References

[1] S. Babić, B. Finka, and M. Moguš. *Hrvatski pravopis*. Školska knjiga, 1996.

[2] B. Bekavac and M. Tadić. Implementation of Croatian NERC system. In *Proc. of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 11–18, 2007.

[3] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, and J.C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[4] A. Cucchiarelli and P. Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1):123–131, 2001.

[5] J.F. Da Silva, Z. Kozareva, and GP Lopes. Cluster analysis and classification of named entities. In *Proc. Conference on Language Resources and Evaluation*, pages 321–324, 2004.

[6] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[7] M. Faruqui and S. Padó. Training and evaluating a German named entity recognizer with semantic generalization. *Semantic Approaches in Natural Language Processing*, page 129, 2010.

[8] Dayne Freitag. Trained named entity recognition using distributional clusters. In *Proceedings of EMNLP*, volume 4, pages 262–269, 2004.

[9] G. Georgiev, P. Nakov, K. Ganchev, P. Osenova, and K. Simov. Feature-rich named entity recognition for Bulgarian using conditional random fields. In *Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP'2009)*, pages 113–117, 2009.

[10] R. Grishman and B. Sundheim. Message Understanding Conference-6: A brief history. In *Proc. of COLING*, volume 96, pages 466–471, 1996.

[11] J. Kravalová and Z. Žabokrtskỳ. Czech named entity corpus and SVM-based recognizer. In *Proc. of the 2009 Named Entities Workshop: Shared Task on Transliteration*, pages 194–201, 2009.

[12] V. Krishnan and C.D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1121–1128, 2006.

[13] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML01*, 2001.

[14] P. Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.

[15] N. Ljubešić, T. Lauc, and D. Boras. Generating a morphological lexicon of organization entity names. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[16] M. Marcińczuk and M. Janicki. Optimizing CRF-based model for proper name recognition in Polish texts. *Computational Linguistics and Intelligent Text Processing*, pages 258–269, 2012.

[17] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191, 2003.

[18] A. Mikheev, C. Grover, and M. Moens. Description of the LTG system used for MUC-7. In *Proc. of 7th Message Understanding Conference (MUC-7)*, 1998.

[19] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[20] N. Okazaki. CRFsuite: a fast implementation of conditional random fields (CRFs), 2007.

[21] J. Piskorski. Extraction of Polish named entities. In *Proc. of the Fourth International Conference on Language Resources and Evaluation, LREC*, pages 313–316, 2004.

[22] T. Poibeau. The multilingual named entity recognition framework. In *Proc. of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 155–158, 2003.

[23] B. Popov, A. Kirilov, D. Maynard, and D. Manov. Creation of reusable components and language resources for named entity recognition in Russian. In *Proc. of the Fourth International Conference on Language Resources and Evaluation, LREC*, pages 309–312, 2004.

[24] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[25] J. Šnajder, B.D. Bašić, and M. Tadić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731, 2008.

[26] S. Yu, S. Bai, and P. Wu. Description of the Kent Ridge Digital Labs system used for MUC-7. In *Proc. of the Seventh Message Understanding Conference*, 1998.

Semi-Supervised Learning for Quantitative Structure-Activity Modeling

Jurica Levatić
Faculty of Science, Department of Mathematics, University of Zagreb, Zagreb, Croatia
Bijenička cesta 30, 10000 Zagreb, Croatia
E-mail: jurica.levatic@ijs.si

Sašo Džeroski
Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: saso.dzeroski@ijs.si

Fran Supek and Tomislav Šmuc
Division of Electronics, Institute Ruđer Bošković, Zagreb, Croatia
Bijenička cesta 54, 10000 Zagreb, Croatia
E-mail: fran.supek@irb.hr, smuc@irb.hr

Keywords: semi-supervised learning, supervised learning, QSAR, drug design, machine learning

Received: September 6, 2012

In this study, we compare the performance of semi-supervised and supervised machine learning methods applied to various problems of modeling Quantitative Structure Activity Relationship (QSAR) in sets of chemical compounds. Semi-supervised learning utilizes unlabeled data in addition to labeled data with the goal of building better predictive models than can be learned by using labeled data alone. Typically, labeled QSAR datasets contain tens to hundreds of compounds, while unlabeled data are easily accessible via public databases containing thousands of chemical compounds: this makes QSAR modeling an attractive domain for the application of semi-supervised learning. We tested four different semi-supervised learning algorithms on three different datasets and compared them to five commonly used supervised learning algorithms. While adding unlabeled data does help for certain pairings of dataset and method, semi-supervised learning is not clearly superior to supervised learning across the QSAR classification problems addressed by this study.

Povzetek: Metode delno-nadzorovanega učenja smo testirali na različnih podatkih iz domene kvantitativnega modeliranja razmerja med strukturo in aktivnostjo kemičnih spojin (angl. Quantitative Structure Activity Relationship, oziroma QSAR).

1 Introduction

Two major approaches to machine learning are supervised learning (e.g., classification, regression), where all the data are labeled, and unsupervised learning (e.g., clustering, dimensionality reduction) where all the data are unlabeled. The semi-supervised learning (SSL) paradigm [21] examines how merging both types of data (labeled and unlabeled) affects learning, aiming to benefit from the information that unlabeled data bring in the context of the supervised learning tasks.

SSL is of important practical value since the following scenario often holds true: labeled data are scarce and hard to get because they require human experts, expensive devices or time-consuming experiments, while, at the same time, unlabeled data abound and are easily obtainable. Real-world classification problems of this type include: phonetic annotation of human speech, protein 3D structure prediction, and spam filtering. Intuitively, SSL yields best results when there are few labeled examples as compared to unlabeled ones (i.e., large-scale labelling is not affordable). But, the setting where plenty of labeled data are available is also suitable for SSL, if even more unlabeled data are available. The other scenario where SSL can be applied is 'domain adaptation'; where we have labeled examples belonging to one domain, but we want to develop a model for another, related, domain.

Establishing a connection between biological effects and structural and/or physicochemical properties of chemicals is the task of quantitative structure-activity relationship or QSAR modeling. Formal studies of such relationships are the basis for the development of predictive models. The main value of a predictive QSAR model is the fact that it provides insight into the biological activity of a molecule without the need to

synthesize it. This leads to a number of benefits including savings in the cost and duration of product development (e.g., in the pharmaceutical or pesticide industries), reduction of the need for animal testing, prediction of unwelcome or toxic environmental impact, and overall improvement in the efficiency of drug design.

The application of SSL to the domain of QSAR modeling is particularly attractive since the premise: "labeled data are scarce, while unlabeled data abound" is generally satisfied in this domain. Public databases with (hundreds of) thousands of chemical compounds are available (e.g., the human tumor cell line screen database from the U.S. National Cancer Institute's Developmental Therapeutics program), while labeled datasets sizes typically range from tens to hundreds and rarely surpass a thousand molecules.

In this work, we empirically investigate whether we can successfully apply SSL (i.e., whether we can achieve better performance with SSL than with supervised learning) to build predictive QSAR models. To draw reliable conclusions, we use several SSL methods which embody different approaches, together with three QSAR datasets from various domains. We compare the SSL methods to several commonly used supervised learning methods. The results show that the improvements which SSL yields are selective - the degree to which unlabeled data help varies from notable to insignificant, depending on the dataset or SSL method used.

2 Semi-supervised learning

In this study, we are concerned with semi-supervised classification, while other forms of SSL, such as semi-supervised regression or semi-supervised clustering are not considered.

2.1 The task of semi-supervised classification

In supervised learning, we are given training data in the form of instance-label pairs, i.e., for each instance we know the desired prediction. The goal is to use the training data to infer a mapping, from instances to labels, which will provide (true) labels for future instances. If the domain of labels is discrete, such a mapping is called a classification function (or a classifier).

The task of *semi-supervised classification* is an extension to the task of supervised classification, where the training data, in addition to the labeled instances, contain a set of unlabeled instances. The goal is again to produce a classification function, which hopefully performs better than the classifier learned from the supervised data only classifier. Figure 1 shows a simple example how unlabeled data can help to induce a classifier that is better in separating the classes.

2.2 Major approaches to semi-supervised classification

In order for SSL to work, the knowledge we gain trough unlabeled data has to carry some information about the class labels. If this prerequisite is fulfilled, we can draw on unlabeled data by making certain assumptions about the behavior of labels with respect to the structure of unlabeled data. Different assumptions inspire different classes of algorithms; therefore, SSL methods can be grouped on the basis of the assumption(s) they implement as follows: low-density separation methods, graph-based methods, generative models, self-training and co-training.

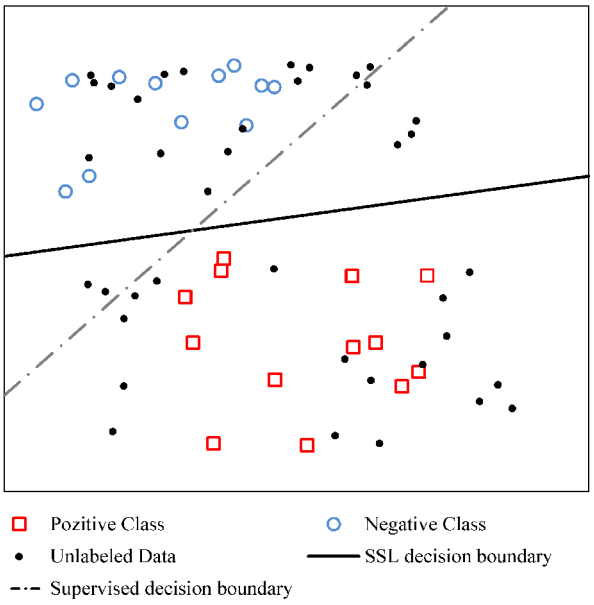


Figure 1. Semi-supervised linear SVMs use unlabeled data to find decision boundary which separates the two classes better than the decision boundary discovered by supervised SVMs.

Low-density separation methods assume that the decision boundary should lie in the region of low density of the data. For example, semi-supervised support vector machines try to find a labeling for the unlabeled data in a way that maximizes the margin of the decision boundary considering both labeled and unlabeled data. Equivalent to the low density separation assumption is *the cluster assumption*: the points belonging to the same cluster should be of the same class.

Graph-based methods use nodes for data representation (labeled and unlabeled) and edges (usually with weights representing the similarity of the data points) for propagation of the labels through the graph, assuming label smoothness over the graph (i.e, the label of the unlabeled instance should be similar to its neighbors in the graph). Here, unlabeled data help to "bridge" the points which would otherwise be unconnected. The construction of the graph is a critical step of graph-based methods – it should reflect the information which is not easily encoded in feature vectors.

Generative models assume a probabilistic model of the data and use unlabeled, together with labeled data, to estimate the most probable model parameters. The success of generative models depends largely on choosing a probabilistic model which is appropriate for the data. Once the probabilistic model is chosen (e.g.,

Gaussian mixture models), a maximum likelihood estimate (MLE) of the parameters can be calculated (e.g., by using the Expectation-Maximization algorithm), followed by a calculation of class distributions using Bayes' rule.

Self-training and *co-training* are two approaches that are often used by SSL algorithms, since they can be "wrapped" around any (supervised) learning algorithm. They iteratively use their own most reliable predictions in the training process (assuming they are correct), as additional data for learning. The main pitfall of these methods is the reinforcement of mistakes – a mistake once made can reinforce itself in the next iterations, leading to degradation of performance.

These assumptions are at the heart of SSL, but also present the main risk for bad performance of SSL: an inappropriate match of a problem structure to a method's assumption can cause severe degradation of performance when using unlabeled data [2]. This is a particularly relevant issue since it is not yet clearly understood which SSL method should be used for which problem, or whether a certain problem (or dataset) is suitable for SSL the use of at all. As mentioned before, unlabeled data has to carry useful information about the structure of the data with respect to the labels.

Zhang and Oles [19] tried to quantify the value of unlabeled data in a probabilistic framework by using regularized logistic regression as an approximation of support vector machines. They showed that, in the setting where labeled and unlabeled data do not share parameters, semi-supervised support vector machines are unlikely to be helpful in general, and are prone to maximize the "wrong margin". It should be noted that unlabeled data should not be used to compensate for the lack of labeled data, but to complement labeled data. In other words, the improvements based on SSL should not rely on the inability of supervised methods to learn anything useful at all due to the lack of data.

We tackled the difficulties of matching the problem structure with the right SSL method empirically, i.e., by selecting methods which differ in their basic approach. We tried to cover most of the groups of methods mentioned above. The SSL methods we used will be described in Section 4.

3 QSAR datasets

To better assess the performance of SSL algorithms in the domain of QSAR modeling, we extracted three different datasets from publicly available sources. These are the NCI, Mutagenicity and MUSK dataset. The datasets differ in terms of the biological activity they model, the number and type of molecular descriptors used to represent molecules, and the number of compounds (size of the dataset).

3.1 NCI dataset

The NCI datasets was extracted from the human tumor cell line screen database [11] of the National Cancer Institute's Developmental Therapeutics (NCI-DTP) program (October 2009 release). The NCI-DTP measures cytostatic activity of chemical compounds against 60 human tumor cell lines grown in cell culture. For representation of a compound's cytostatic activity we used GI_{50} measurements – the compound concentration that inhibits cell growth by 50%. Only compounds that have missing or default values for at most 20 cell lines were accepted. Additionally, cell lines with more than 20% of missing values were removed, leaving 49 cell lines in total. The compounds were thus described with the GI_{50} profiles across the 49 cell lines, and in addition with two other groups of attributes: (1) molecular descriptors describing the structure of a molecule (calculated with the DRAGON 3.0 web interface [18]), and (2) molecular charge densities and charge density-based electrostatic properties of a molecule (calculated with the RECON software [4]).

The subject of interest for the NCI dataset is to predict a compound's mechanism of action (MOA) – the biological process in which the molecule interacts with its molecular targets - proteins (enzymes or otherwise) or DNA. The type of MOA influences the pharmacological effects of a molecule; therefore, the drug discovery process benefits from an early detection of an appropriate MOA for a given use. The NCI dataset represents a multiclass classification problem, with 12 different MOA classes, where each molecule belongs to a single class. A very similar dataset has been used to find putative MOAs for new drug candidates [7, 16], and is essentially an updated and extended version of the dataset used in previous analyses of cytostatic activities and MOA in global computational analyses of the NCI database using self-organizing maps [13, 15].

3.2 Mutagenicity dataset

The Mutagenicity dataset [10] is the benchmark dataset for modeling of Ames mutagenicity. The Ames test is a standard microbiological assay for assessing the mutagenic potential of a chemical compound. A compound which is positive to the test causes mutations on the DNA (and consequently can be carcinogenic); avoiding mutagenicity is important for drug-candidates and other molecules with significant human exposure (e.g., cosmetics, food additives).

The mutagenicity dataset represents a binary classification problem where compounds are classified as mutagenic or non-mutagenic. Molecules from this dataset were represented by using DRAGON molecular descriptors [18].

3.3 MUSK dataset

The MUSK dataset was downloaded from the UCI machine learning repository [8]. Musk, a substance secreted by the Asian musk deer, is an expensive animal product heavily used by the perfume industry; therefore, synthetic compounds are often used instead. The prediction of the strength of such synthetic musk compounds has similarities to the prediction of biological drug activity – the molecules are similar in size and composition to the orally active drug molecules [5].

A single molecule can adopt multiple conformations – different shapes of the same molecule, when some of the internal bonds rotate. The features that describe compounds from the MUSK dataset depend on the exact shape (conformation) of a molecule ("distance features" and displacement of oxygen; a detailed description is given by Dietterich et al. [5]), where each molecule is represented by several feature vectors. This dataset was assembled by generating low-energy conformations of molecules, which were then filtered to remove highly similar conformations. The molecules from the MUSK dataset were categorized by human experts to be musk or non-musk.

4 Experimental setup

To evaluate the potential of SSL in a controlled manner (i.e., to be able to evaluate the methods thoroughly, and to make sure that the unlabeled data is relevant to the problem), our experiments were carried out using only labeled data. We simulated unlabeled data by temporarily ignoring the class label for a portion of the data. The relative amount of unlabeled and labeled data is a relevant factor when measuring the success of SSL methods: SSL should perform better when the labeled set is rather small and a lot of unlabeled data are available. Our experiments were aimed to test the former premise by creating situations where we have different ratios of labeled and unlabeled data.

The data were randomly split into a training and a test set. Both the supervised and the semi-supervised methods used the training set for learning and were then evaluated by using the test set. For the SSL methods, the test set served as unlabeled data during the learning process. Several different train/test splits were produced where labeled data ranges from 1% to 66% (i.e., unlabeled data ranges from 99% to 33%). The final results were averaged over 10 different train/test split repetitions, in order to obtain a more robust evaluation of the algorithms. We performed experiments using the Weka [9] machine learning environment and the R [17] environment for statistical computing.

4.1 Datasets

As described in Section 3, we conducted experiments on three different QSAR datasets. The NCI dataset contains 507 compounds, each described with: GI50 profiles (49 features in the form of -logGI50), DRAGON descriptors (1497 features) and RECON descriptors (248 features). The Mutagenicity dataset is the largest with 6512 compounds represented with 1497 DRAGON descriptors. The MUSK dataset has 166 features and 476 examples, which correspond to different conformations of 92 molecules.

4.2 Methods

We used publicly available implementations of several SSL algorithms. As mentioned in Section 2, we selected the SSL algorithms to cover different groups of SSL methods. The algorithms used are: Yet Another Two Stage Idea (YATSI), Co-training: Fitting the Fits (Co-FTF), Learning with Local and Global Consistency (LLGC) and TSVMLight.

The YATSI [6] algorithm, implemented in the Weka Collective Classifiers package, is similar to the self-training concept, since it can be wrapped around any classifier and it uses its own predictions in the training process. As the name implies, YATSI works in two steps. First, a base classifier is trained on the labeled data and then unlabeled data is "pre-labeled". This pre-labeled data is then given weights and used by the nearest neighbors classifier to improve on the initial classifier.

Co-FTF [3] is an implementation of the co-training algorithm in the R programing language. Co-FTF uses two different features sets (views) to train separate classifiers, which iteratively use their most confident predictions as additional labeled training data. It is assumed that views provide different, complementary information about the data. We applied Co-FTF only to the NCI dataset (the other datasets do not meet the prerequisite for different views) with the combination of the descriptors which proved to be the best: RECON and DRAGON. Other combinations: GI50 profiles coupled with RECON or DRAGON descriptors, achieved lower performances (not shown). The baseline classifier for Co-FTF was the random forests classifier with 500 trees.

LLGC [20] is a graph-based method implemented in the Weka Collective Classifiers package. LLGC first performs spectral clustering and then propagates labels through the graph using a spreading activation network.

TSVMLight [12] is a representative of the low-density separation methods. It implements a semi-supervised version of support vector machines by finding the locally optimal solution.

The supervised machine learning methods that we compared with SSL methods were taken from Weka: decision trees (J48), k-nearest neighbors (KNN), Naive Bayes (NB), support vector machines (SMO from Weka, and the stand-alone version of SVMLight) and random forests (RF).

We used the J48, NB and SMO methods with their default parameters and RF with 500 trees. For the KNN method, the 'crossValidate' option was used to select an appropriate number K of neighbours. For YATSI and LLGC, we used the Weka Experimenter Environment to search for the parameter values which produce the best classification accuracy. The parameters for (T)SVMLight were tuned manually.

5 Results and discussion

In this section we present the experimental comparison of performance of semi-supervised and supervised machine learning methods. In Tables 1-3, the predictive accuracies for different ratios of labeled and unlabeled data are presented. The best result for each ratio is shown in bold, and whether YATSI exhibited improvement in accuracy over the baseline classifier is marked with an upward (improvement) or downward (deterioration) arrow. The baseline classifier for YATSI is given in

brackets. The number of neighbors for the KNN algorithm is indicated (e.g., 1NN).

Semi-supervised methods behave differently over the three datasets. Improvements of semi-supervised over supervised learning are most notable for the NCI dataset with a small percentage of labeled data (≤10%), where LLGC achieves the best overall predictive accuracy and YATSI significantly improves the baseline classifier in most cases. YATSI consistently deteriorates the performance of SMO for all amounts of labeled data.

For the other two datasets, Mutagenicity and MUSK, semi-supervised and supervised algorithms show very similar performance with small improvements of SSL over supervised learning in some cases. Generally, the improvements achieved by YATSI over the baseline classifier are more frequent and significant for the less complex classifiers (KNN, J48, NB), while for classifiers with greater capacity for learning (RF, SMO) the improvements are not so regular and are sometimes negative, i.e., the usage of YATSI even deteriorates their predictive accuracy (Figure 1).

Driessens et al. [6] performed an extensive testing of YATSI over 29 different datasets with several different base classifiers and made similar observations: YATSI behaves somewhat differently when using RF and SMO as base classifiers, as compared to the other algorithms (including J48 and KNN). In most cases, YATSI lost some of the accuracy achieved by RF, and performed equal to SMO, while it improved other base classifiers (with most notable improvements when little labeled data were available).

In the setting of supervised learning, robust methods, such as support vector machines or random forests are known to perform well on a wide range of classification tasks, and can be successfully used without specific domain knowledge. The results obtained on the datasets considered in this study confirm this: the SMO and RF

	Algorithm	Percentage of labeled data				
		5%	10%	20%	33%	66%
Supervised learning	J48	45.93	57.98	69.05	76.31	81.92
	1NN	47.45	67.19	78.14	82.73	86.80
	NB	42.41	51.19	66.13	74.49	84.14
	SMO	62.80	73.15	**83.42**	**87.41**	**92.69**
	RF	56.24	66.32	78.29	84.14	88.64
Semi-supervised learning	YATSI(J48)	55.37↗	68.27↗	78.54↗	83.31↗	84.87↗
	YATSI(1NN)	58.87↗	70.89↗	79.95↗	82.79↗	86.50↘
	YATSI(NB)	54.70↗	65.96↗	75.61↗	81.67↗	83.03↘
	YATSI(SMO)	62.06↘	72.69↘	81.99↘	84.76↘	87.59↘
	YATSI(RF)	58.76↗	68.44↗	79.11↗	83.46↘	86.32↘
	LLGC	**66.50**	**74.95**	82.46	85.46	88.29
	Co-FTF	-	35.78	51.16	65.43	76.45

Table 1: Predictive accuracies of semi-supervised and supervised learning methods on the NCI dataset

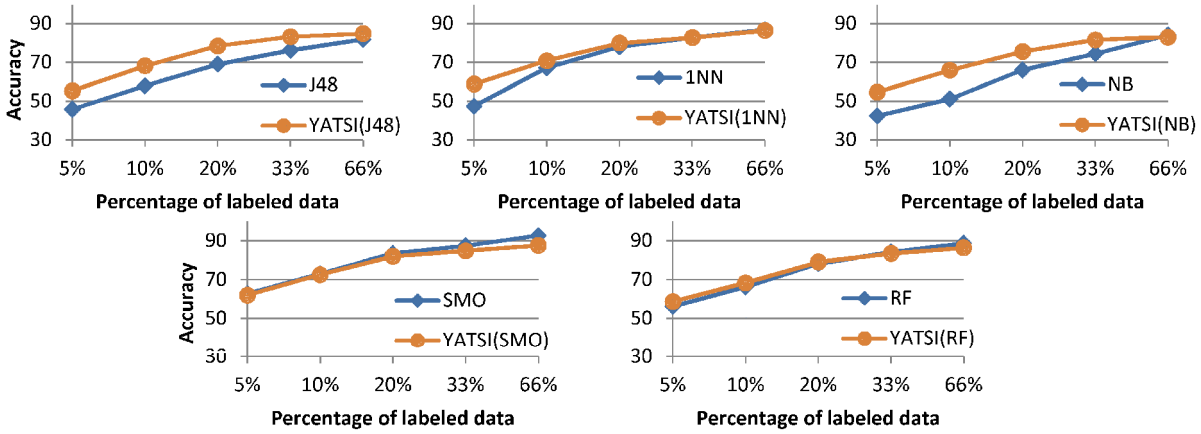


Figure 1. Comparison of learning curves for YATSI and baseline algorithms on the NCI dataset show that YATSI improves the performance of the less complex classifiers (J48, KNN, NB), but not the more complex classifiers (SMO and RF). The improvements in accuracy which unlabeled data brings gradually decrease with the increase of the relative amount of labeled data.

	Algorithm	Percentage of labeled data			
		1%	5%	10%	20%
Supervised learning	J48	58.40	63.98	67.08	69.97
	1NN	58.32	64.24	64.71	67.86
	NB	57.80	60.40	61.10	60.79
	SMO	61.86	68.95	72.41	**75.41**
	RF	62.13	68.68	71.27	73.68
	SVM^{Light}	**62.73**	69.29	72.52	75.16
Semi-supervised learning	YATSI(J48)	58.85↗	65.78↗	68.19↗	70.88↗
	YATSI(1NN)	58.45↗	64.57↗	66.77↗	69.30↗
	YATSI(NB)	57.85↗	62.71↗	64.62↗	65.02↗
	YATSI(SMO)	61.53↘	67.84↘	70.35↘	72.73↘
	YATSI(RF)	59.50↘	66.36↘	67.89↘	70.62↘
	$TSVM^{Light}$	61.24	**69.65**	**72.85**	**75.41**
	LLGC	58.75	62.70	63.65	64.86

Table 2: Predictive accuracies of semi-supervised and supervised learning methods on the Mutagenicity dataset

	Algorithm	Percentage of labeled data		
		5%	10%	20%
Supervised learning	2NN	63.89	71.15	77.97
	SMO	66.01	71.65	76.03
	RF	62.70	71.84	78.77
	SVM^{Light}	**69.49**	75.18	**81.02**
Semi-supervised learning	YATSI(2NN)	65.12↗	71.51↗	78.05↗
	YATSI(SMO)	67.83↗	74.42↗	78.07↗
	YATSI(RF)	63.57↗	73.25↗	78.48↘
	$TSVM^{Light}$	66.69	**75.25**	80.50
	LLGC	65.34	73.11	80.39

Table 3: Predictive accuracies of semi-supervised and supervised learning methods on the MUSK dataset.

classifiers consistently outperform the other (supervised) methods. However, if we compare SSL methods across the three datasets (Tables 1-3) we do not have a clear winner. For example, the LLGC algorithm performs better than the other SSL methods on the NCI dataset, but it is outperformed on the Mutagenicity and MUSK datasets.

Similar observations have been made by other scientists: Chawla and Karakoulas [1] performed an extensive empirical study of SSL techniques over various domains (not including QSAR modeling), using real-world and artificial datasets to investigate the conditions under which SSL can perform well. They observed that SSL methods behave very differently depending on the nature of the datasets, and that no single SSL method consistently performs better than supervised learning.

In practice, it is not easy to assess in advance how certain SSL method will behave given the task at hand. Several method/problem combinations are known to work well together (e.g., semi-supervised SVMs and text classification, [12]), but there are no clear strategies how to verify the model assumptions against certain problem structure. Specific domain knowledge and understanding of SSL algorithms should be used to couple the problem at hand with an appropriate method. Currently, scientists in this area are dealing with the question of how to make SSL safe, i.e., how to make sure that SSL performs at least as well as supervised learning, and how to make SSL usable by non-experts on realistic tasks [14].

6 Conclusion and future work

In this study, we performed an empirical comparison of several semi-supervised and supervised machine learning methods on three different QSAR datasets under different experimental conditions (amount of unlabeled data relative to labeled data). Our results show that SSL can achieve better predictive performance than supervised learning (typically when a small portion of the data is labeled), but the improvements depend on the dataset and method used. We cannot claim clear superiority of semi-supervised over supervised learning on the QSAR classification problems addressed by this study. However, the large improvements (in general and relative to the baseline classifier) in classification accuracy in certain cases suggest that it is worthwhile to

take SSL into consideration when dealing with problems of QSAR modeling.

Semi-supervised learning is a more delicate task than supervised learning, where more (labeled) data generally a means better and more robust model. While more unlabeled data can help, it is not guaranteed to do so. We have pointed out the difficulties that one can encounter when dealing with the task of semi-supervised learning, as compared to supervised learning.

In further work, we would like to systematically investigate which features of a dataset make it suitable for the use of SSL. In addition, we would like to extend our experiments and use data which are truly unlabeled. This would enable us to exploit the vast amount of information readily available within public compound databases.

References

[1] Chawla, N.V. and Karakoulas, G. 2005. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*. 23 (1), 331–366.

[2] Cozman, F.G. et al. 2002. Unlabeled data can degrade classification performance of generative classifiers. In *Proc of the Fifteenth International Florida Artificial Intelligence Research Society Conference* (2002), 327–331. AAAI Press.

[3] Culp, M. and Michailidis, G. 2009. A co-training algorithm for multi-view data with applications in data fusion. *Journal of Chemometrics*. 23 (6), 294–303.

[4] Curt M. Breneman et al. 2003. *RECON version 5.5/5.3*. Rensselaer Polytechnic Institute.

[5] Dietterich, T.G. et al. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*. 89 (1–2), 31–71.

[6] Driessens, K. et al. 2006. Using weighted nearest neighbor to benefit from unlabeled data. In *Proc of the Knowledge Discovery and Data Mining*, 60–69. Springer.

[7] Ester, K. et al. 2012. Putative mechanisms of antitumor activity of cyano-substituted heteroaryles in HeLa cells. *Investigational New Drugs*. 30 (2), 450–467.

[8] Frank, A. and Asuncion, A. 2010. *UCI Machine Learning Repository:* University of California, Irvine, School of Information and Computer Sciences. *http://archive.ics.uci.edu/ml.* Accessed: January, 2011.

[9] Hall, M. et al. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 11 (1), 10–18.

[10] Hansen, K. et al. 2009. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *Journal of Chemical Information and Modeling*. 49 (9), 2077–2081.

[11] Holbeck, S.L. 2004. Update on NCI in vitro drug screen utilities. *European Journal of Cancer*. 40 (6), 785–793.

[12] Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proc of the Sixteenth International Conference on Machine Learning*, 200–209. Morgan Kaufmann.

[13] Rabow, A.A. et al. 2002. Mining the National Cancer Institute's Tumor-Screening Database: Identification of Compounds with Similar Cellular Activities. *Journal of Medicinal Chemistry*. 45 (4), 818–840.

[14] Xiaojin Zhu, Semi-Supervised Learning for Non-Experts: *http://pages.cs.wisc.edu/~jerryzhu/ssl/*. Accessed: August, 2012.

[15] Supek, F. et al. 2005. A prototype structure-activity relationship model based on National Cancer Institute cell line screening data. *Periodicum Biologorum*. 107 (4), 451.

[16] Supek, F. et al. 2008. Atypical cytostatic mechanism of N-1-sulfonylcytosine derivatives determined by in vitro screening and computational analysis. *Investigational New Drugs*. 26 (2), 97–110.

[17] Team, R.C. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing

[18] Tetko, I.V. et al. 2005. Virtual computational chemistry laboratory--design and description. *Journal of Computer-aided Molecular Design*. 19 (6), 453–463.

[19] Zhang, T. and Oles, F. 2000. A probability analysis on the value of unlabeled data for classification problems. In *Proc of the Seventeenth International Conference on Machine Learning*, 1191–1198. Morgan Kaufmann.

[20] Zhou, D. et al. 2004. Learning with local and global consistency. *Advances in Neural Information Processing Systems*. 16, 321–328.

[21] Zhu, X. and Goldberg, A.B. 2009. *Introduction to Semi-Supervised Learning*. Morgan and Claypool.

Quaternion Based Fuzzy Neural Network Classifier for MPIK Dataset's View-invariant Color Face Image Recognition

Wai Kit Wong and Gin Chong Lee
Faculty of Engineering and Technology, Multimedia University
75450 JLN Ayer Keroh Lama, Melaka, Malaysia.
E-mail: wkwong@mmu.edu.my, gclee@mmu.edu.my

Chu Kiong Loo, and Raymond Lock
Faculty of Computer Science and Information Technology, University of Malaya
50603 Lembah Pantai, Kuala Lumpur, Malaysia.
E-mail: ckloo.um@gmail.com

Keywords: image processing, face recognition, fuzzy neural network classifier, quaternion correlation

Received: January 12, 2013

This paper presents an effective color image processing system view-invariant person face image recognition for Max Planck Institute Kybernetik (MPIK) dataset. The proposed system can recognize face images of view-invariant person by correlating the input face images with the reference face image and classifying them according to the correct persons' name/ID indeed. It has been carried out by constructing a complex quaternion correlator and a max-product fuzzy neural network classifier. Two classification parameters, namely discrete quaternion correlator output (p-value) and the peak to sidelobe ratio (PSR), were used in classifying the input face images, and to categorise them either into the authentic class or non-authentic class. Besides, a new parameter called G-value is also introduced in the proposed view-invariant color face image recognition system for better classification purpose. Experimental results shows that the proposed view-invariant color face image recognition system outperforms the conventional NMF, BDNMF and hypercomplex Gabor filter in terms of consumption of enrollment time, recognition time and accuracy in classifying MPIK color face images which are view-invariant, noise influenced and scale invariant.

Povzetek: Predstavljena je metoda prepoznavanja obrazov, testirana na domeni Max Planck Institute Kybernetik.

1 Introduction

Face recognition has been applied in many areas such as face search in databases, authentication in security system, smart user interfaces, robotics and etc. Conventional face recognition methods normally focus on grayscale face image recognition. However in recently, there are many researchers focus on color information of the face images to improve the performance of recognition algorithm due to the reasons that color face images offer more information for face recognition task in contrast to grayscale face images.

A simple color face recognition system was first proposed by Torres et. al. in [1] based on the PCA (principal component analysis) method. The method is based on the representation of the facial images using eigenfaces. The information of three different channels (R, G, B) of color face images are first represented in the form of eigenvectors respectively; and the recognition is implemented separately on each color channel. However, it is found out that the information of different color channels that utilized separately would destroy the color information's structural and make it hard to learn the facial features (variation in expression, poses and illuminations). Rajapakse et. al. [2] presented a parallel work based on NMF (Non-negative matrix factorization) for color face recognition. In their work, color information on face images of different channels were processed separately too. Some observed advantages of NMF method on face recognition are this method is more robust to occlusion, variation of expressions and poses. However, since NMF method also treats information of different color channels separately, just like the PCA method, it would also destroy the structural integrally of color information and the correlation among the color information.

In order to preserve the integrally of color information on different channels in color face recognition system, Wang et. al. [3, 4] proposed a supersede NMF method, which is the block diagonal non-negative matrix factorization (BDNMF). Inspired by the NMF method, BDNMF also separated color information into different color channels, but it uses block diagonal matrix to simultaneously encode color information of different channels, hence preserving the integrally of color information. However, BDNMF method has the demerit of complex enrolment/training stage. In BDNMF, unsupervised multiplicative learning

rules are used iteratively to up-date the parameters such as basis image matrix (W) and encoding image (H). Therefore, longer enrolment time is required for this method. Another demerit of BDNMF is that an additional coined block diagonal constraint is imposed on the factorization part to construct the BDNMF algorithm. This makes the computation more complex compare to the conventional NMF method.

Another recently developed color face recognition method is the use of hypercomplex Gabor filters [5]. Conventional Gabor filters are used in many face recognition applications [6-8] and they are proven to obtain good recognition performance due to its inherent merits of insensitivity to illumination and pose variation. In [5], the author further extended conventional Gabor filter into hypercomplex (quaternion) domain to perform color based feature extraction. Experimental results in [5] show that the conventional Gabor filter feature extraction achieved significant improvement in face matching accuracy over the monochromatic case. However, hypercomplex Gabor filter required a large number of different kernels, and hence the length of the feature vectors in quaternion domain would increase dramatically. Also, hypercomplex Gabor filter is twice the size of filter structure comparing to those used in the conventional Gabor filters.

Most of the proposed algorithms for color face recognition treat the three color channels (R, G, B) separately and apply grayscale face recognition methods [9, 10] to each of the channels and then combine the results at last. But with the quaternion correlation techniques [11], it processes all color channels jointly by using its quaternion numbers. Quaternion numbers are the generalization of complex numbers. It is a number which consists of one real part and three orthogonal imaginary parts. A RGB color face image can be represented using quaternion representation by inserting the value of three color channels into the three imaginary parts of the quaternion number respectively. Therefore, in this paper, the concept of quaternion is proposed for view-invariant color face image recognition system.

An advanced correlation filter named as unconstrained optimal trade-off synthetic discriminant (UOTSDF) [12, 13], is also applied in the proposed view-invariant color face image recognition system. The goal of the filter is to produce sharp peak that resemble 2-D delta type correlation outputs when the input face image belongs to the class of the reference face image that were used to train the input face image; and, this provides automatic shift-invariance. A strong and sharp peak can be observed in the output correlation plane when the input face image comes from the authentic class (input face image matches with a par-ticular training/reference face image stored in database) and no discernible peak if the input face image comes from the imposter class (input face image does not matches with the particular reference face image).

Three classification parameters are in concern in classifying whether an input face image belongs to the authentic class or not. They are the real to complex ratio of the discrete quaternion correlator output (ρ-value) [11], peak to sidelobe ratio (PSR) [14] and the max product fuzzy neural network classifier value (G-value). ρ-value has been introduced in [11], which is used in measuring the correlation output between the colors, shape, size and brightness of input image and a particular reference image. PSR is another parameter introduced in [14] for a better recognition due to the reason that it is more accurate if we consider the peak value with the region around the peak value, rather than a single peak point. The higher the value of PSR, the more likely the input face image belongs to the referenced image class. In this paper, both the ρ-value and the PSR are combined, normalized and applied with Gaussian distribution function in the max-product fuzzy neural network classifier. This technique generates a parameter, so-called the G-value. This parameter as well as the algorithm is applied in view-invariant color face image recognition system for better classification purposes. The same technique was applied in the machine condition monitoring [15] and it yields high success rate in classifying machine conditions. It is good to be implemented for color face image recognition.

In this paper, quaternion based fuzzy neural network classifier is proposed for MPIK dataset's view-invariant color face image recognition. 10,000 repeated images generated/collected from the 7 different position color face images of 200 people in MPIK dataset were used to evaluate the system performance. Among the 10,000 repeated color face images, 5000 are normal MPIK color face images; 2500 are normal MPIK color face images embedded with noise features such as "salt and pepper", "poisson", "speckles noise" as provided in Matlab image processing toolbox; and, 2500 are normal MPIK color face images with scale invariant (shrink or dilation). The performance of the proposed quaternion based fuzzy neural network classifier is compared to NMF, BDNMF and hypercomplex Gabor filter. Experimental results show that the quaternion based fuzzy neural network classifier outperforms conventional NMF, BDNMF and hypercomplex Gabor filter in terms of enrolment time, recognition time and accuracy in classifying view-invariant, noise influenced and scale invariant MPIK color face images.

The paper is organized as follows: Section 2 briefly comments on the proposed view-invariant color face image recognition model and the quaternion based color face image correlator. Section 3 describes the enrolment stage and recognition stage for the algorithm of the proposed quaternion based color face image correlator. Then in section 4, the structure of fuzzy max-product neural network classifier will be described. Section 5 contains the experimental results. Finally, in section 6, the work is summarized and some future work is planned.

2 View-invariant color face image recognition system model

The proposed view-invariant face recognition system model considered in this paper is shown in Figure 1.

The view-invariant input color face image is first supplied to the quaternion based color face image correlator. The quaternion based color face image

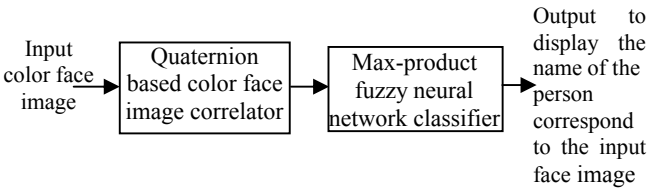


Figure 1: View-invariant color face image recognition system.

correlator is used to obtain correlation plane for each correlated input face image with reference face images stored in a database to calculate out some classification characteristics such as the real to complex ratio of the discrete quaternion correlator (DQCR) output, *p*-value and the peak-to-sidelobe ratio, PSR. These classification characteristic will later be input to the max-product fuzzy neural network to perform classification. Detailed discussion on the quaternion based color face image correlation will be discussed below.

The referenced face image after performing discrete quaternion Fourier transforms (DQFT) [11]:

$$I(\mathrm{m},\mathrm{n}) = I_R(\mathrm{m},\mathrm{n}).i + I_G(\mathrm{m},\mathrm{n}).j + I_B(\mathrm{m},\mathrm{n}).k \quad (1)$$

where m, n are the pixel coordinates of the reference face image. R, G, B parts of reference face image are represented by $I_R(\mathrm{m},\mathrm{n})$, $I_G(\mathrm{m},\mathrm{n})$ and $I_B(\mathrm{m},\mathrm{n})$ respectively, and *i*-, *j*-, *k*- are the imaginary parts of quaternion complex number [15] and the real part of it is set to zero. Similarly, $\mathrm{h_i}(\mathrm{m},\mathrm{n})$ is used for representing input face image. Then, we can produce output $\mathrm{b}(\mathrm{m},\mathrm{n})$ to conclude whether the input face image matches the reference face image or not. If $\mathrm{h_i}(\mathrm{m},\mathrm{n})$ is the space shift of the reference face image:

$$\mathrm{h_i}(\mathrm{m},\mathrm{n}) = I(\mathrm{m}-\mathrm{m_0},\mathrm{n}-\mathrm{n_0}) \quad (2)$$

Then after some mathematical manipulation,

$$\mathrm{Max}\left(\mathrm{b_r}(\mathrm{m},\mathrm{n})\right) = \mathrm{b_r}(-\mathrm{m_0},\mathrm{n_0}) \quad (3)$$

where $\mathrm{b_r}(\mathrm{m},\mathrm{n})$ means the real part of $\mathrm{b}(\mathrm{m},\mathrm{n})$ and

$$\mathrm{b_r}(-\mathrm{m_0},\mathrm{n_0}) = \sum_{\mathrm{m}=0}^{\mathrm{M}-1}\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\left|I(\mathrm{m},\mathrm{n})\right|^2 \quad (4)$$

where M, N is the image x-axis, y-axis dimension. At the location ($-\mathrm{m_0}$, $\mathrm{n_0}$), the multiplier of *i*-, *j*-, *k*- imaginary part of $\mathrm{b}(-\mathrm{m_0}, \mathrm{n_0})$ are equal to zero:

$$\mathrm{b_i}(-\mathrm{m_0},\mathrm{n_0}) = \mathrm{b_j}(-\mathrm{m_0},\mathrm{n_0}) = \mathrm{b_k}(-\mathrm{m_0},\mathrm{n_0}) = 0 \quad (5)$$

Thus, the following process [11] can be modified for face image correlation:

1.) Calculate energy of reference face image $I(\mathrm{m},\mathrm{n})$:

$$E_I = \sum_{\mathrm{m}=0}^{\mathrm{M}-1}\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\left|I(\mathrm{m},\mathrm{n})\right|^2 \quad (6)$$

Then we normalized the reference face image $I(\mathrm{m},\mathrm{n})$ and the input face image $\mathrm{h_i}(\mathrm{m},\mathrm{n})$ as:

$$I_{\mathrm{a}}(\mathrm{m},\mathrm{n}) = I(\mathrm{m},\mathrm{n})\Big/\sqrt{E_I} \quad (7)$$

$$\mathrm{H_a}(\mathrm{m},\mathrm{n}) = \mathrm{h_i}(\mathrm{m},\mathrm{n})\Big/\sqrt{E_I} \quad (8)$$

2.) Calculate the output of discrete quaternion correlation (DQCR):

$$\mathrm{g_a}(\mathrm{m},\mathrm{n}) = \sum_{\tau=0}^{\mathrm{M}-1}\sum_{\eta=0}^{\mathrm{N}-1} I_{\mathrm{a}}(\tau,\eta).\overline{\mathrm{H_a}(\tau-\mathrm{m},\eta-\mathrm{n})} \quad (9)$$

where ‘ $\overline{\ }$ ’ means the quaternion conjugation operation and perform the space reverse operation:

$$\mathrm{g}(\mathrm{m},\mathrm{n}) = \mathrm{g_a}(-\mathrm{m},\ -\mathrm{n}) \quad (10)$$

3.) Perform inverse discrete quaternion Fourier Transform (IDQFT) on (10) to obtain the correlation plane $\mathrm{P}(\mathrm{m},\mathrm{n})$.

4.) Search all the local peaks on the correlation plane and record the location of the local peaks as ($\mathrm{m_s}$, $\mathrm{n_s}$).

5.) Then at all the location of local peaks ($\mathrm{m_s}$, $\mathrm{n_s}$) found in step 4, we calculate the real to complex value of the DQCR output:

$$p = \frac{\left|\mathrm{P_r}(\mathrm{m_s},\mathrm{n_s})\right|}{\left|\mathrm{P_r}(\mathrm{m_s},\mathrm{n_s})\right| + \left|\mathrm{P_i}(\mathrm{m_s},\mathrm{n_s})\right| + \left|\mathrm{P_j}(\mathrm{m_s},\mathrm{n_s})\right| + \left|\mathrm{P_k}(\mathrm{m_s},\mathrm{n_s})\right|} \quad (11)$$

where $\mathrm{P_r}(\mathrm{m_s},\mathrm{n_s})$ is the real part of $\mathrm{P}(\mathrm{m_s},\mathrm{n_s})$. $\mathrm{P_i}(\mathrm{m_s},\mathrm{n_s})$, $\mathrm{P_j}(\mathrm{m_s},\mathrm{n_s})$ and $\mathrm{P_k}(\mathrm{m_s},\mathrm{n_s})$ are the *i*-, *j*- *k*- parts of $\mathrm{P}(\mathrm{m_s},\mathrm{n_s})$ respectively. If $p \geq \mathrm{d_1}$ and $\mathrm{c_1} < \left|\mathrm{P}(\mathrm{m_s},\mathrm{n_s})\right| < \mathrm{c_2}$, then we can conclude that at location ($\mathrm{m_s}$, $\mathrm{n_s}$), there is a face image that has the same shape, size, color and brightness as the reference face image. $\mathrm{d_1} < 1$, $\mathrm{c_1} < 1 < \mathrm{c_2}$ and $\mathrm{c_1}$, $\mathrm{c_2}$ and $\mathrm{d_1}$ are all with values near to 1. The value of *p* decays faster with the color difference between matching the input face image to the reference face image.

Another classification characteristic that can be applied in quaternion based color face image correlation is the peak-to-sidelobe ratio (PSR). A strong peak can be observed in the correlation output if the input face image comes from imposter class. A method of measuring the peak sharpness is the peak-to-sidelobe ratio (PSR) which is defined as below [14, 17]:

$$\mathrm{PSR} = \frac{peak - mean(sidelobe)}{\sigma(sidelobe)} \quad (12)$$

where *peak* is the value of the peak on the correlation output plane. *sidelobe* refers a fixed-sized surrounding

area off the peak. *mean* is the average value of the sidelobe region. σ is the standard deviation of the sidelobe region. Large PSR values indicate the better match of the input face image and the corresponding reference face image.

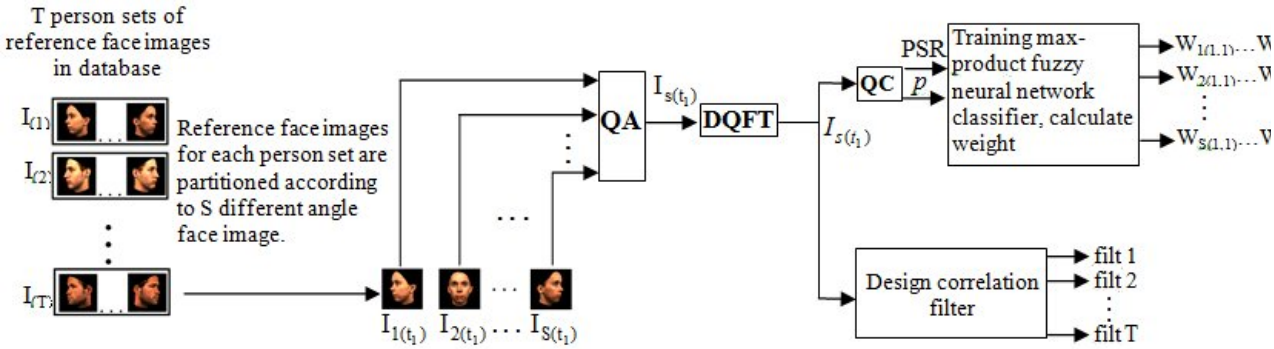


Figure 2: Schematic of enrollment stage.

The quaternion based color face image correlator involved 2 stages: 1. Enrolment stage and 2. Recognition stage. During the enrollment stage, one or multiple face images of each person in database are acquired. These multiple reference face images have the variability in the angle of turning faces (for e.g. 90º to left, 60º to left, 30º to left 0º facing in front, 30º to right, 60º to right, 90º to right and etc). The DQFT of the reference face images are used to train the fuzzy neural network and determine correlation filter coefficients for each possible person's set. During recognition stage, sample face images will be input and the DQFT of such images are correlated with the DQFT form of the reference face images stored in the database together with their corresponding filter coefficients, and the inverse DQFT of this product results in the correlation output for that filter. Enrollment stage and recognition stage are discussed in detail in the following section.

3 Enrolment stage and recognition stage for quaternion based color face image correlator

This section will describes the enrollment stage and recognition stage for the algorithm of the proposed quaternion based color face image correlator.

3.1 Enrolment Stage

The schematic of enrollment stage is shown in Figure 2. During the enrollment stage, the reference face images for each person set in database are partitioned according to S different angle face image. These partitioned reference face images are then encoded into a two dimensional quaternion array (QA) as follows:

$$\mathrm{I}_{s(t_1)} = \mathrm{I}_{sr(t_1)} + \mathrm{I}_{sR(t_1)}.i + \mathrm{I}_{sG(t_1)}.j + \mathrm{I}_{sB(t_1)}.k \qquad (13)$$

where $t_1 = 1, 2, \ldots, T$ represents the number of person subscribe to the database, $\mathrm{I}_{sr(t_1)}$ represents the real part of quaternion array of s-th face image for person set t_1, s = 1, 2, …, S represents the number of face images in different angle for a particular person. $\mathrm{I}_{sR(t_1)}$, $\mathrm{I}_{sG(t_1)}$ and $\mathrm{I}_{sB(t_1)}$ each represents the *i*-, *j*-, *k*- imaginary part of s-th face image for person t_1 respectively.

The quaternion array in (13) is then undergoes discrete quaternion Fourier transform (DQFT) to transform the quaternion image to the quaternion frequency domain. A two-side form of DQFT has been proposed by Ell [18, 19] as follows:

$$I_{s(t_1)}(m,n) = \sum_{\tau=0}^{M-1}\sum_{\eta=0}^{N-1} e^{-\mu_1 2\pi(m\tau/M)}.\mathrm{I}_{s(t_1)}(\tau,\eta).e^{-\mu_2 2\pi(n\eta/N)} \qquad (14)$$

where *e* is exponential term, μ_1 and μ_2 are two pure quaternion units (the quaternion unit with real part equal to zero) that are orthogonal to each other [20]:

$$\mu_1 = \mu_{1,i}.i + \mu_{1,j}.j + \mu_{1,k}.k \qquad (15)$$

$$\mu_2 = \mu_{2,i}.i + \mu_{2,j}.j + \mu_{2,k}.k \qquad (16)$$

$$\mu_{1,i}^2 + \mu_{1,j}^2 + \mu_{1,k}^2 = \mu_{2,i}^2 + \mu_{2,j}^2 + \mu_{2,k}^2 = 1$$

$$(\text{i.e}: \ \mu_1^2 = \mu_2^2 = -1) \qquad (17)$$

$$\mu_{1,i}.\mu_{2,i} + \mu_{1,j}.\mu_{2,j} + \mu_{1,k}.\mu_{2,k} = 0 \qquad (18)$$

The output of DQFT, $I_{s(t_1)}$ is used to train the max-product fuzzy neural network classifier and design the correlation filter.

3.1.1 Quaternion Correlator (QC)

To train the max-product fuzzy neural network classifier, the output of the DQFT is first passed to a quaternion correlator (QC) as shown in Figure 3. The function of the QC is summarized as below: For DQFT output of s-th face image, perform discrete quaternion correlation (DQCR) [21, 22] on reference face image $I_{s(t_1)}$ with reference face image $I_{s(t_2)}$ and multiply with corresponding filter coefficients ($\mathrm{filt}_{(t_2)}$):

$$g_{s(t_1,t_2)}(m,n)=\sum_{\tau=0}^{M-1}\sum_{\eta=0}^{N-1} I_{s(t_1)}.\overline{I_{s(t_2)}(\tau-m,\eta-n)}.\,filt_{(t_2)} \tag{19}$$

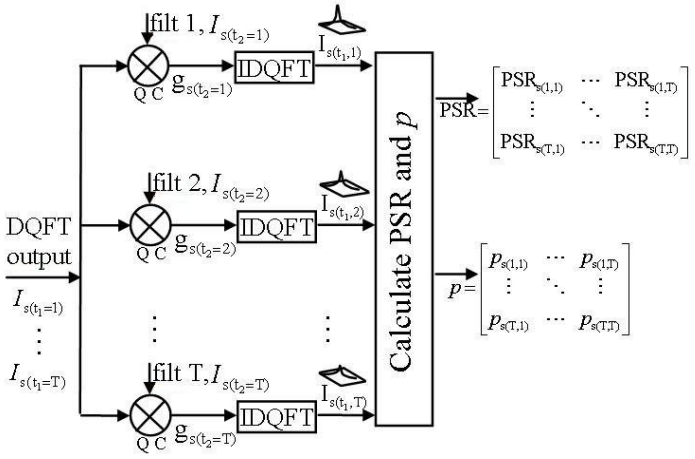


Figure 3: Quaternion correlator (QC)

where t_1, t_2 = 1, 2, …, T are the number of person subscribe to the database. After that, (19) is performing inverse DQFT to obtain the correlation plane function:

$$P_{s(t_1,t_2)}(m,n)=\frac{1}{4\pi^2}\sum_{\tau=0}^{M-1}\sum_{\eta=0}^{N-1} e^{-\mu_1 2\pi(m\tau/M)}.g_{s(t_1,t_2)}(m,n).\; e^{-\mu_2 2\pi(n\eta/N)} \tag{20}$$

The correlation plane is a collection of correlation values, each one obtained by performing a pixel-by-pixel comparison (inner product) of two images ($I_{s(t_1)}$ and $I_{s(t_2)}$). A sharp peak in the correlation plane indicates the similarity of $I_{s(t_1)}$ and $I_{s(t_2)}$, while the absence or lower of such peak indicate the dissimilarity of $I_{s(t_1)}$ and $I_{s(t_2)}$.

Calculate $p_{s(t_1,t_2)}$ and $PSR_{s(t_1,t_2)}$ from the correlation plane as in (20) using (11) and (12) respectively. $p_{s(t_1,t_2)}$ means p-values of reference face image $I_{(t_1)}$ correlate on reference face image $I_{(t_2)}$ in s-th angle, while $PSR_{s(t_1,t_2)}$ means PSR values of reference face image $I_{(t_1)}$ correlate on reference face image $I_{(t_2)}$ in s-th angle. These values are then feed into max-product fuzzy neural network classifier to perform training and calculate weight, which will be discussed in section 4.

3.1.2 Correlation Filter

Conventional filtering methods [23] are emphasizing on applying matched filters. Matched filters are optimal for detecting a known reference image in additive white Gaussian noise environment. If the input image changes slightly from the known reference image (scale, rotation and pose invariant), the detection of the matched filters degrades rapidly. However the emerge of correlation filter designs [24] have developed to handle such types of distortions. The minimum average correlation energy (MACE) filters [25] are one of such design and show good results in the field of automatic target recognition and applications in biometric verification [14, 26]. MACE filters different from conventional matched filters that more than one reference image are used to synthesize a single filter template, therefore making its classification performance invariant to shift of the input image [24].

There are two types of MACE filters in general, namely: 1.) Conventional MACE filter [25] and 2.) Unconstrained MACE (UMACE) filter [27], both with the goal to produce sharp peaks that resemble two dimensional delta-type correlation outputs when the input image belongs to the authentic class and low peaks in imposter class. Conventional MACE filter [25] minimizes the average correlation energy of the reference images while constraining the correlation output at the origin to a specific value (usually 1), for each of the reference images. Lagrange multiplier is used for noise optimization, yielding:

$$filt_{MACE}=D^{-1}X(X'D^{-1}X)^{-1}c \tag{21}$$

This equation is the closed form solution to be the linear constrained quadratic minimization. D is a diagonal matrix with the average power spectrum of the reference images placed as elements along diagonal of the matrix. X contains Fourier transform of the reference images lexicographically re-ordered and placed along each column. As an example, if there are T sets of reference face images, each with size $256\times 1{,}792$(=458,752), then X will be a $458{,}792\times T$ matrix. X' is the matrix transpose of X. c is a column vector of length T with all entries equal to 1.

The second type of MACE filter is the unconstrained MACE (UMACE) filter [27]. Just like conventional MACE filter, UMACE filter also minimizes the average correlation energy of the reference images and maximizes the correlation output at the origin. The different between conventional MACE filter and UMACE filter is the optimization scheme. Conventional MACE filter is using Lagrange multiplier but as for UMACE filter, it is using Raleigh quotient which lead to the following equation:

$$filt_{UMACE}=D^{-1}m \tag{22}$$

where D is the diagonal matrix which is the same as that in conventional MACE filter. m is a column vector containing the mean values of the Fourier transform of the reference images.

Besides MACE filters, there is a type of correlation filter, namely the unconstrained optimal tradeoff synthetic discriminant filter (UOTSDF) shown by Refreiger [28] and Kumar et al [12] has yielding good verification performance. The UOTSDF is by:

$$filt_{UOTSDF}=(\alpha D+\sqrt{1-\alpha^2}C)^{-1}m \tag{23}$$

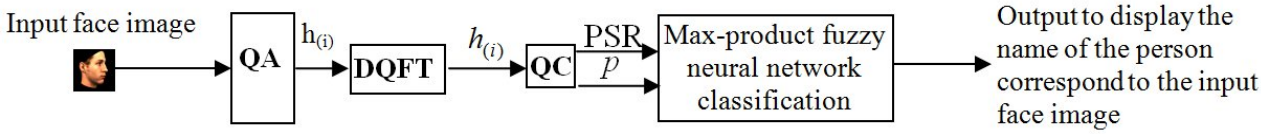


Figure 4: Schematic of recognition stage.

where D is a diagonal matrix with average power spectrum of the training image placed along the diagonal elements. m is a column vector containing the mean values of the Fourier transform of the reference images. C is the power spectral density of the noise. For most of the applications, a white noise power spectral density is for assumption since white noise is dominant in image; therefore C reduces to the identity matrix. α term is typically set to be close to 1 to achieve good performance even in the presence of noise , but it also helps improve generalization to distortions outside the reference images.

By comparing the three correlation filters listed above, conventional MACE filter is complicated to implement whereby it requires many inversion of $T \times T$ matrices. UMACE filter is simpler to implement from a computational viewpoint as it involves inverting diagonal matrix only, and the performance are close to the conventional MACE but poorer than UOTSDF. Therefore, we plan to extend UOTSDF in our quaternion based face image correlator for the recognition of view invariant person face since it is less complicated in computational viewpoint than conventional MACE filter and achieve good performance.

3.2 Recognition stage

The schematic of recognition stage for classification of color face image by quaternion correlation is shown in Figure 4. During the recognition stage, an input view invariant face image is first encoded into two dimensional quaternion array (QA) as follows:

$$\mathrm{h}_{(\mathrm{i})} = \mathrm{h}_{\mathrm{r(i)}} + \mathrm{h}_{\mathrm{R(i)}}.i + \mathrm{h}_{\mathrm{G(i)}}.j + \mathrm{h}_{\mathrm{B(i)}}.k \qquad (24)$$

where i represents the input face image, $\mathrm{h}_{\mathrm{r(i)}}$ represents the real part of quaternion array for input face image i. $\mathrm{h}_{\mathrm{R(i)}}$, $\mathrm{h}_{\mathrm{G(i)}}$ and $\mathrm{h}_{\mathrm{B(i)}}$ each represents the *i*-, *j*-, *k*-imaginary part for input face image i respectively.

Performing DQFT to transforms the quaternion image to the quaternion frequency domain. A two-side form of DQFT is used:

$$h_{(\mathrm{i})}(\mathrm{m},\mathrm{n}) = \sum_{\tau=0}^{\mathrm{M}-1}\sum_{\eta=0}^{\mathrm{N}-1} e^{-\mu_1 2\pi(\mathrm{m}\tau/\mathrm{M})}.\mathrm{h}_{(\mathrm{i})}(\tau,\eta).e^{-\mu_2 2\pi(\mathrm{n}\eta/\mathrm{N})} \qquad (25)$$

where *e* is exponential term, μ_1 and μ_2 are two pure quaternion units as shown in (15) and (16) respectively. The output of the DQFT, $h_{(\mathrm{i})}$ is cross correlated with every quaternion correlation filter in the database using the quaternion correlator (QC) just as the one shown in Figure 3, but the DQFT output is now $\mathrm{h}_{(\mathrm{i})}$. In QC, performs quaternion correlation on $h_{(\mathrm{i})}$ with reference face images $I_{\mathrm{s}(\mathrm{t}_2)}$ from database, and multiply with corresponding filter coefficients ($\mathrm{filt}_{(\mathrm{t}_2)}$):

$$\mathrm{g}_{\mathrm{s(i,\,t_2)}}(\mathrm{m},\mathrm{n}) = \sum_{\tau=0}^{\mathrm{M}-1}\sum_{\eta=0}^{\mathrm{N}-1} h_{(\mathrm{i})}.\overline{I_{\mathrm{s}(\mathrm{t}_2)}(\tau-\mathrm{m},\eta-\mathrm{n})}.\mathrm{filt}_{(\mathrm{t}_2)} \qquad (26)$$

After that, (26) is performing inverse DQFT to obtain the correlation plane function:

$$\mathrm{P}_{\mathrm{s(i,\,t_2)}}(\mathrm{m},\mathrm{n}) = \frac{1}{4\pi^2}\sum_{\tau=0}^{\mathrm{M}-1}\sum_{\eta=0}^{\mathrm{N}-1} e^{-\mu_1 2\pi(\mathrm{m}\tau/\mathrm{M})}.\mathrm{g}_{\mathrm{s(i,\,t_2)}}(\mathrm{m},\mathrm{n}).\ e^{-\mu_2 2\pi(\mathrm{n}\eta/\mathrm{N})} \qquad (27)$$

Calculate $p_{\mathrm{s(i,t_2)}}$ and $\mathrm{PSR}_{\mathrm{s(i,t_2)}}$ from the correlation plane as in (27) using (11) and (12) respectively. $p_{\mathrm{s(i,t_2)}}$ means *p*-values of input face image $h_{(\mathrm{i})}$ correlate on s-th reference face image in $I_{(\mathrm{t}_2)}$, while $\mathrm{PSR}_{\mathrm{s(i,t_2)}}$ means PSR values of input image $h_{(\mathrm{i})}$ correlate on s-th reference face image in $I_{(\mathrm{t}_2)}$. These values are then feed into max-product fuzzy neural network classifier to perform classification for view invariant face images, which will be discussed in next section.

4 Max-product fuzzy neural network classifier

Fuzzy logic is a type of multi-valued logic that derived from fuzzy set theory to deal with approximate reasoning. Fuzzy logic provides high level framework for approximate reasoning that can appropriately handle both the uncertainty and imprecision in linguistic semantics, model expert heuristics and provide requisite high level organizing principles [13]. Neural network in engineering field refer to a mathematical/computational model based on biological neural network. Neural network provides self-organizing substrates for low level representation of information with adaptation capabilities. Fuzzy logic and neural network are complementary technologies. Therefore, it is plausible and justified to combine both these approaches in the design of classification systems. Such integrated system is referring to as fuzzy neural network classifier [13].

There are various fuzzy neural network classifiers have been proposed in the literature [29-32], and there has been much interest of many fuzzy neural networks applying max-min composition as functional basis [33-35]. However, in [36], Leotamonphong and Fang

mention that the max-min composition is "suitable only when a system allows no compensability among the elements of a solution vector". He proposed to use max-product composition in fuzzy neural network rather than max-min composition. Bourke and fisher in [37] also comment that the max-product composition gives better results than the traditional max-min operator. Therefore, efficient learning algorithms have been studied by others [38, 39] using the max-product composition afterwards.

In this paper, a fuzzy neural network classifier using max-product composition will be proposed for view invariant color face image classification system. The max-product composition is the same as a single perceptron except that summation is replaced by maximization, and in the max-min threshold unit, min is replaced by product.

4.1 Define T classes, for T person's sets of view invariant face images

The reference face images for all T persons in database will be assigned with a *Unique Number* started from 1 till (T×S), where S is the number of S different angle face image specified in section 3. *Class* number is assigned starting from 1 till T. The same person's face images in different angle of view will be arrange in sequence according to the unique number assigned and classified in the same *Class* number.

4.2 Training Max-Product Fuzzy Neural Network Classifier

The max-product fuzzy neural network classifier is training with 4 processes as listed below:

1.) $\mathrm{PSR}_{s(t_1,t_2)}$ and $p_{s(t_1,t_2)}$ output from the quaternion correlator of the enrollment stage are fuzzified through the activation functions (Gaussian membership function):

$$\mathrm{G}_{\mathrm{PSR}_{s(t_1,t_2)}} = \exp\left[\frac{-\left(\mathrm{PSR}_{s(t_1,t_2)}-1\right)^2}{\sigma^2}\right] \tag{28}$$

$$\mathrm{G}_{p_{s(t_1,t_2)}} = \exp\left[\frac{-\left(p_{s(t_1,t_2)}-1\right)^2}{\sigma^2}\right] \tag{29}$$

where σ is the smoothing factor, that is the deviation of the Gaussian functions.

2.) Calculate the G-value, which is the product value for s-th reference face image of the fuzzy neural network classifier at each correlated images:

$$\mathrm{G}_{s(t_1,t_2)} = \mathrm{G}_{\mathrm{PSR}_{s(t_1,t_2)}} \times \mathrm{G}_{p_{s(t_1,t_2)}} \tag{30}$$

3.) Gather and store the product values in an array:

$$\mathrm{X}_{\text{s training}} = \begin{bmatrix} \mathrm{G}_{s(1,1)} & \mathrm{G}_{s(1,2)} & \cdots & \mathrm{G}_{s(1,T)} \\ \mathrm{G}_{s(2,1)} & \mathrm{G}_{s(2,2)} & \cdots & \mathrm{G}_{s(2,T)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{G}_{s(T,1)} & \mathrm{G}_{s(T,2)} & \cdots & \mathrm{G}_{s(T,T)} \end{bmatrix} \tag{31}$$

4.) The output will be set so that it will output 1 if it is authentic class and 0 if it is imposter class, and it is in an array $\mathrm{Y}_{\text{identity}}$, whereby it is an identity matrix of dimension $\mathrm{T}\times\mathrm{T}$. To calculate the weight w for s-th angle face image, the equation is:

$$w_{\mathrm{s}} = \mathrm{X}_{\text{straining}}{}^{-1}\mathrm{Y}_{\text{identity}} \tag{32}$$

4.3 Max-Product Fuzzy Neural Network Classification

The max-product fuzzy neural network classification is with 7 steps:

1.) $\mathrm{PSR}_{s(i,t_2)}$ and $p_{s(i,t_2)}$ output from the quaternion correlator of the recognition stage are fuzzified through the activation functions (Gaussian membership function):

$$\mathrm{G}_{\mathrm{PSR}_{s(i,t_2)}} = \exp\left[\frac{-\left(\mathrm{PSR}_{s(i,t_2)}-1\right)^2}{\sigma^2}\right] \tag{33}$$

$$\mathrm{G}_{p_{s(i,t_2)}} = \exp\left[\frac{-\left(p_{s(i,t_2)}-1\right)^2}{\sigma^2}\right] \tag{34}$$

2.) Calculate the product value of the fuzzy neural network classifier at input face image on the training face images in database:

$$\mathrm{G}_{s(i,t_2)} = \mathrm{G}_{\mathrm{PSR}_{s(i,t_2)}} \times \mathrm{G}_{p_{s(i,t_2)}} \tag{35}$$

3.) Gather and store the product values in an array:

$$\mathrm{X}_{\text{s classification}} = \begin{bmatrix} \mathrm{G}_{s(i,1)} & \mathrm{G}_{s(i,2)} & \cdots & \mathrm{G}_{s(i,T)} \end{bmatrix} \tag{36}$$

4.) Obtain the classification outcomes by multiplying (36) with the weight trained at (32):

$$\mathrm{Y}_{\text{classification}} = \mathrm{X}_{\text{s classification}} \times w_{\mathrm{s}} \tag{37}$$

5.) Classify the input face image with the person it belongs to by using max composition:

$$\mathrm{Output} = \max\left\{\mathrm{Y}_{\text{classification}}\right\} \tag{38}$$

6.) Determine whether the face image is in the database or not:

If normalized Output ≤ $\mathrm{Thres}_{\text{output}}$

Then conclude: "The face is not in the database".

Else determine which element in $\mathrm{Y}_{\text{classification}}$ matrix match with Output :

ψ = the position number of element in $Y_{classification}$ matrix which has the equal value with Output . (39)

$Thres_{output}$ is the threshold value of an output to indicate that a face is not in the database.
ψ corresponds to the assigned number of reference image in database.

7.) Based on T sets of fuzzy IF-THEN rules, perform defuzzification:

R^l: IF ψ is match with the *Unique Number* stored in *Class l*, THEN display the name of the person correspond to *Class l*. (40)

where l = 1, 2, … T.

5 Experimental results

In this section, the application of quaternion based face image correlator together with max-product fuzzy neural network classifier for view invariant face recognition system will be briefly illustrated. Here, some experimental results are used to prove the algorithms' efficiency introduced in section 3 and 4.

5.1 Database of reference face images for 200 persons

A database with view-invariant color face images provided by the Max-Planck Institute for Biological Cybernetics in Tuebingen Germany [40] is use to test the proposed view-invariant color face image recognition system. The database contains color face images of 7 views of 200 laser-scanned (Cyberware TM) heads without hair. These modeling 200 persons' sets of color face images each with view-invariant/angle of different: facing 90º to left, facing 60º to left, facing 30º to left, facing 0º in-front, facing 30° to right, facing 60° to right and facing 90° to right. Hence, S=7 since there are 7 view-invariant images for 1 person set. An example of a person set with view-invariant face images are shown in Figure 5. The dimension of each image is 256 x 256 pixels.

5.2 Quaternion based face image correlation using unconstrained optimal trade-off synthetic discriminant filter (UOTSDF)

In the evaluation experiment, T=180 MPIK persons' faces are used to train the system during the enrollment stage. T x S = 1260 reference face images are use in database to synthesize a single UOTSDF using (23). D and m are calculated from the reference images and C is an identity matrix of dimension 1260 x 1260 and α set to 1. These values are substituted into (23) to calculate out the filter coefficients. Then in enrollment stage, for each filter line as in Figure 3, perform cross-correlations of all

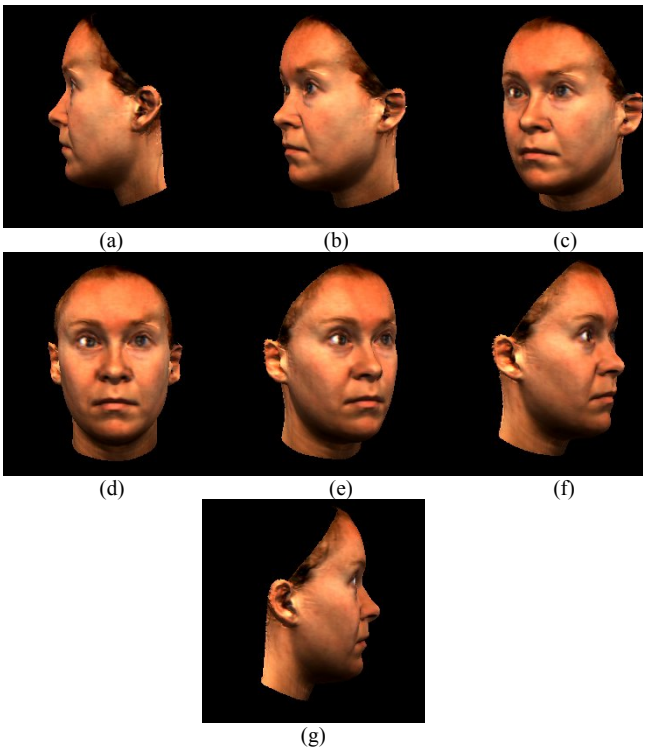


Figure 5: An example of a person set with view-invariant face images (a) facing 90º to left, (b) facing 60º to left, (c) facing 30º to left, (d) facing 0º in-front, (e) facing 30° to right, (f) facing 60° to right and (g) facing 90° to right.

the DQFT form of reference face images in database $I_{s(t_1)}$ with the DQFT form of reference face image in database as well $I_{s(t_2)}$, and multiply the output value with corresponding filter coefficients respectively, where t_1 , t_2 = 1, 2, … , 180; s = 1, 2, …, 7. In recognition stage, for each filter line, performed cross correlation of the DQFT form of input face image ($h_{(i)}$) with the DQFT form of reference face images in database ($I_{s(t_2)}$) and multiply the output value with corresponding filter coefficient respectively. For authentic case (good match in between two face images), the correlation plane should have sharp peaks and it should not exhibit such strong peaks for imposter case (bad match in between two face images). These two cases will be investigated below:

Authentic case: Figure 6 shows the samples correlation plane for input face image matching with the exact reference face image of the same person in the database. Since both the face images are in good match, the observed correlation plane is having smooth and sharp peak.

Imposter case: Figure 7 show the sample correlation plane for input face image matching with one of the reference face image of different person in the database. Since both the face images are not in good match, the observed correlation plane is having lower and round peak as compare to those in good match.

Table 1 shows the PSR and p-value for both authentic and imposter case as in Fig. 6 and Fig.7. Note that the sharp correlation peak resulting in large

normalized PSR and *p*-value in authentic case, whereas small PSR and *p*-value exhibiting in the imposter case.

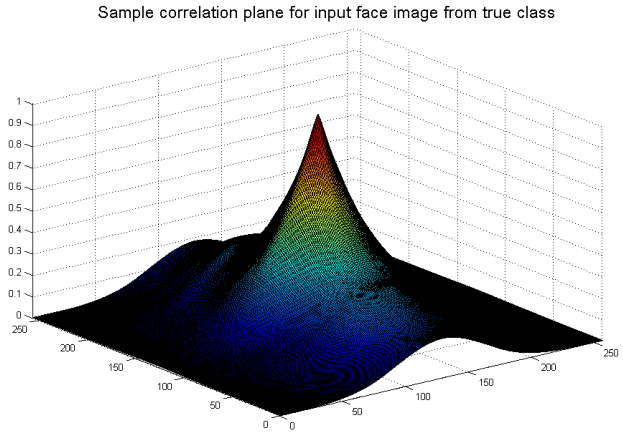


Figure 6: Sample correlation plane for input face image matching with the exact reference face image of the same person class in the database (authentic case).

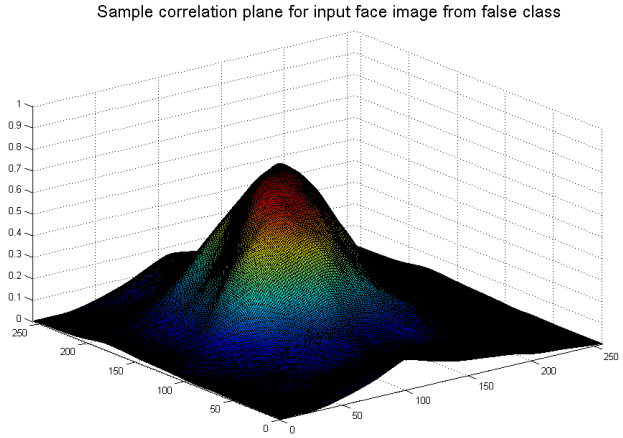


Figure 7: Sample correlation plane for input face image matching with one of the reference face image of different person in the database.

Case	**Normalized PSR**	Normalized *p*-value
Authentic case	1.0000	0.7905
Imposter case	0.7894	0.5083

Table 1: normalized PSR and p-value for both authentic and imposter case.

To indicate that a face is not in the database, a threshold, $\text{Thres}_{\text{output}}$ is implemented on the normalized Output value at Step 6 in section 4.3. The 20 persons' faces samples excluded from the training database are input to the trained system to run for accuracy test on different normalized Output value ranges from 0.05 to 1.0. The plot is shown in Figure 8. From the plot, the optimum $\text{Thres}_{\text{output}}$ is at 0.6.

5.3 Efficiency of the view-invariant color face image recognition system

The view-invariant color face image recognition system was evaluated with respect to random picking 10,000 repeated input face images from database (with mixing up the trained T=180 peoples' faces plus 20 more peoples' faces excluded from the training database sets)

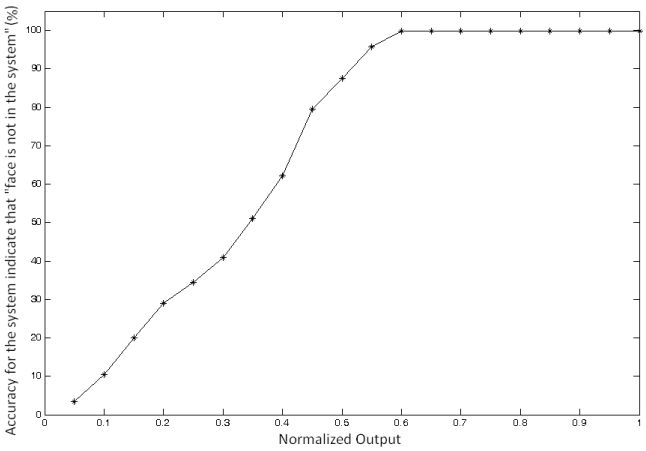


Figure 8: plot of accuracy versus normalized Output.

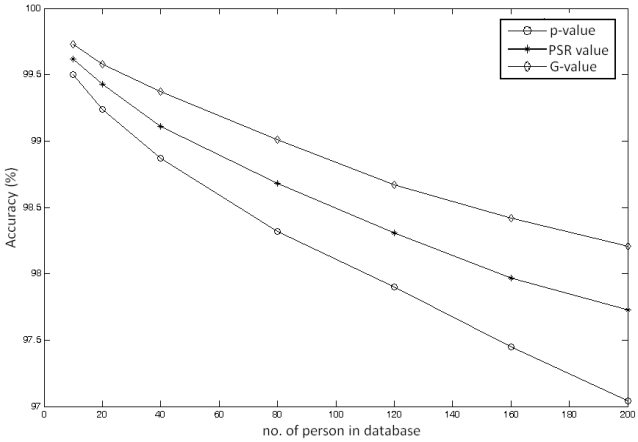


Figure 9: plot of accuracy versus no. of person sets in database.

and input to the view-invariant color face image recognition system to run test. The graph of accuracy versus no. of person sets in database is plotted in Figure 9.

From the plot, it can be observed that as number of person in database increases, the performance drop is actually not much if G-value is applying in the proposed view-invariant color face image recognition system. Among the 10,000 input face images, 9,973were tracked perfectly (output human names/ID agreed by the input images) in a database of 10 persons, i.e. an accuracy of 99.73%, while 9,821 were tracked perfectly in a database of 200 persons, i.e. an accuracy of 98.21%. The performance drop is increase if the proposed face recognition system is only applying PSR value (no G-value and *p*-value) whereby, the accuracy is 99.62% in a database of 10 persons, but 97.73% in a database of 200 persons. It is almost 0.24 fold more the performance drops compares to the system that applying G-value. The performance drops in the face recognition system, that applying only *p* -value (no G-value and PSR-value), is rather significant. Whereby, the accuracy is 99.50% in a database of 10 persons, but 97.04% in a database of 200 persons. It is almost 0.62 fold more the performance drop compares to the system that applying G-value. From the experiment results, it can be concluded that with the

implementation of G-value and the fuzzy neural network classifier, it helps boost up the accuracy of view-invariant color face image recognition.

5.4 Comparative study with parallel method

For comparative study, the proposed quaternion based fuzzy neural network classifier is compared with conventional NMF, BDNMF and hypercomplex Gabor Filter. For conventional NMF, the reference face images as in section 5.1 database has been extracted and used. Seven training sets, each set exclusively containing the color face images for every person in different position (facing 90º to left, facing 60º to left, facing 30º to left, facing 0º in-front, facing 30° to right, facing 60° to right and facing 90° to right). For each training set, three different basis matrices and encodings were extracted for each color channels in the RGB scheme, F^l where $l \in \{R, G, B\}$ is constructed such that each color channel, l, of training color face images occupies the columns of F^l matrices. The rank r of factorization is generally chosen so that [41]:

$$r < \frac{nm}{n+m} \tag{41}$$

In this case, $n = 7$ and $m = 180$, $r < 6.74$. Hence, r is set to 6. The experiment was carried out to test the enrollment stage time consumption and the classification stage time consumption. The recorded time consumption is normalized and recorded in Table 2. For the recognition accuracy, a total of 10,000 randomly selected and repeated MPIK color face images with mixing up the trained T=180 persons' faces plus 20 more persons' faces excluded from the training database sets are tested. These also distributed to 5000 are normal MPIK color face images, 2500 are normal MPIK color face images embedded with noise features such as "salt and pepper", "poisson", "speckles noise" as in Matlab image processing toolbox, and 2500 are normal MPIK color face images with scale invariant (shrink or dilation), some examples are shown in Figure 10. The recognition accuracy (Percentage of total correct recognized images /10,000 tested images) is recorded in Table 2.

Figure 10: An example of a person set (a) original image (b) embedded with mild salt and pepper noise (c) embedded with heavy salt and pepper noise (d) shrink (e) dilation.

For the BDNMF method, to evaluate the performance on different color spaces, the color faces are separate into RGB spaces and face recognition experiment using BDNMF algorithm is conducted. The rank of factorization r is set to 6 as well. In the experiment, all the seven color face images of each person in different position are used to constitute the training/enrollment set. For the testing/classification set, a total of 10,000 face images used in testing the conventional NMF method are used. The results of the identification test including the enrollment stage time consumption (normalized), classification stage time consumption (normalized) and matching accuracy are shown in Table 2.

To evaluate the effectiveness of the hypercomplex Gabor filter proposed by [5] for feature extraction used in this comparative study, the hypercomplex Gabor filter is operated on all the MPIK RGB dataset color face images as in section 5.1. Each color face image was analyzed at a total of 24 landmark location, determined by the statistical analysis of the MPIK face image population according to [5]. Since Mahalanobis distance applied in [5] yields higher accuracies in compare to normal Euclidean distance approach, matching in this comparative study was performed using Mahalanobis distance classification. The jets extracted at the chosen face landmark locations was used for face matching and the Mahalanobis distance was computed using the global covariance matrix for all the color face landmarks. During classification, jets derived from color face images in database were matched against models consisting of jets extracted from a set of 10,000 color face images as use in testing both conventional NMF and BDNMF above, for accuracy measurement. The results of the identification test including the normalized enrollment stage time consumption, normalized classification stage time consumption and matching accuracy are shown in Table 2.

From the experimental results in Table 2, it is observed that quaternion based fuzzy neural network classifier has the fastest enrollment time and classification time. This follow by hypercomplex Gabor filter using Mahalanobis distance classification, conventional NMF and the slowest BDNMF. Conventional NMF and BDNMF are slow due to the reason that they required an iterative training stage for enrollment, which is time consuming and in compare to the proposed quaternion based fuzzy neural network classifier and hypercomplex Gabor filter. Comparing between conventional NMF and BDNMF, BDNMF algorithm imposes an additional constraint, which is the block diagonal constraint, on the base image matrix and coefficient matrix slowing down the enrollment processes. However, with the block diagonal constraint, BDNMF can preserve the integrity of color information better in different channels for color face representation, hence achieves higher accuracy in color face

Color face classification method	Enrollment stage normalized time consumption (for training all datasets in database)	Classification stage normalized time consumption (for matching 10,000 tested image)	Accuracy (output human names/ID match with the corresponde nce input images)
Conventional NMF	2.76	1.39	80.18%
BDNMF	3.51	1.55	83.37%
Hypercomplex Gabor Filter (Mahalanobis Distance Classifiction)	1.36	1.20	86.13%
Quaternion based Fuzzy Neural Network Classifier	1.00	1.00	92.06%

Table 2: normalized enrollment stage time consumption, normalized classification stage time consumption and matching accuracy for different color face classification method.

recognition. In compare to fuzzy neural network, Gabor filter required a large number of different kernels, and hence the length of the feature vectors in quaternion domains would increase dramatically. Therefore, Gabor filter required more time in enrollment and classification comparing to fuzzy neural network. In terms of recognition accuracy, the proposed quaternion based fuzzy neural network outperform hypercomplex Gabor filter, conventional NMF and BDNMF in recognizing view-invariant, noise influenced and scale invariant MPIK color face images.

6 Conclusion

This paper presents a system capable of recognizing view-invariant color face images from MPIK dataset, using quaternion based color face image correlator and max-product fuzzy neural network classifier. One of the advantage of using quaternion correlator rather than conventional correlation method is that quaternion correlation method deals with color images without converting them into grayscale images. Hence important color information can be preserved. Also the proposed Max-product fuzzy neural network provides high level framework for approximate reasoning, since it is best suitable to apply in face image classification. Our experimental results show that the proposed face recognition system's performs well with a very high accuracy of 98% from a dataset of 200 persons each with 7 view-invariant images. In comparative study with parallel work, experimental results also show that the proposed face recognition system outperforms conventional NMF, BDNMF and hypercomplex Gabor filter in terms of consumption of enrolment time, recognition time and accuracy in classifying view-invariant, noise influenced and scale invariant color face images from MPIK. Since artificial dataset (MPIK) was used in the experiments which might be impractical, this work creates a number of avenues for further work. Direct extensions of this work may fall into three main sorts in future. Firstly, more rigorous work is necessary on investigating the system performance in realistic environment and the system should be extended to consider variations include translation, facial expression, and illumination. Real face images such as FERET dataset might be employed in the training as well as empirical tests. Secondly, facial image pre-processing mechanisms, mainly eye detection, geometric and illumination normalization might be employed to ease the image acquisition. A large scale of facial images acquisition and storage of facial data might raise security concerns in terms of identity theft. Third extension might fall in the employment of cancellable face data as a step to reinforce the system security.

References

[1] L. Torres, J. Y. Reutter and L. Lorento, (1999). "The importance of the color information in face recognition", *Proc. Int. Conf. on Systems, Man and Cyberbetics,* Vol. 3, p.p. 627-631.

[2] M. Rajapakse, J. Tan, J. Rajapakse (2004). "Color Channel Encoding With NMF for Face Recognition", *2004 Int. Conf. on Image Processing (ICIP 2004)*, p.p. 2007-2010.

[3] C. Wang, X. Bai (2009). "Color Face Recognition Based on Revised NMF Algorithm", *2nd Int. Conf. on Future Information Technology and Management Engineering*, p.p. 455-458.

[4] X. Bai, C. Wang (2009). "Fisher diagonal NMF based color face recognition", *2010 Chinese Control and Decision Conference* (CCDC), p.p. 4158-4162.

[5] C. Jones III, A. L. Abbott (2006)."Color Face Recognition by Hypercomplex Gabor Analysis", *Proc. Of the 7th Int. conf. on Automatic Face and Gesture Recognition* (FGR' 06), p.p. 1-6.

[6] L. skott, J. Fellous, N. Kruger and C. V. D. Malsburg (1999). "Face recognition by elastic bunch graph matching", Intelligent Biometric Techniques in Fingerprint and Face Recognition, CRC Press, p.p. 355-396.

[7] B. Due, S. Fischer, and J. Bigun (1999). "Face authentication with Gabor information on deformable graphs", IEEE Trans. On Image Processing, Vol. 8, No. 4, p.p. 504-516.

[8] C. Liu and H. Wechsler (2001). "A Gabor feature classifier for face recognition", Proc. Eight IEEE Int. Conf. on Computer Vision, Vol. 2, p.p. 270-275.

[9] S. Lawrence, C. L. Giles, A. C. Tsoi and A. Back (1997). "Face Recognition: A Convolutional Neural Network Approach", *IEEE Trans. On Neural Networks*, Vol. 8, No. 1, p.p. 98-113.

[10] I. Paily, A. Sachenko, V. Koval, Y. Kurylyak (2005). "Approach to Face Recognition Using Neural Networks", *IEEE Workshop on Intelligent*

Data Acquisition and Advanced Computing Systems: Technology and Applications, Sofia, Bulgeria, p.p. 112-115.

[11] S. C. Pei, J. J. Ding and J. Chang (2001). "Color pattern recognition by quaternion correlation", *Proc. of Int. Conf. on Image Processing*, Vol.1, p.p. 894-897.

[12] B.V.K. Kumar, D.W. Carlson, and A. Mahalanobis (1994). "Optimal trade-off synthetic discriminant function filters for arbitrary devices", *Optics Letters*, Vol. 19, No. 19, p.p. 1556-1558.

[13] S. Kumar (2004). *Neural Networks: A Classroom Approach*, McGraw Hill, Int. Ed.

[14] B.V.K. Vijaya Kumar, M. Savvides, K. Venkataramani and C. Xie (2002)."Spatial frequency domain image processing for biometric recognition", *Proc. Of Int. Conf. on Image Processing*, Vol.1, p.p. I53-I56.

[15] W. K. Wong, C. K. Loo, W.S. Lim and P. N. Tan (2009). "Quaternion based thermal condition monitoring system" ,*Fourth International Workshop on Natural Computing (IWNC 2009)*, Himeiji, Japan, p.p.317-327.

[16] W. R. Hamilton (1866). *Elements of Quaternions*, London, U.K.: Longmans, Green.

[17] C. Xie, M. Savvides and B.V.K. Vijaya Kumar (2005). "Quaternion correlation filters for face recognition in wavelet domain", *Int. Conf. on Accoustic, Speech and Signal Processing* (ICASSP 2005), p.p.II85- II88.

[18] T.A. Ell (1993). "Quaternion-Fourier transforms for analysis of two-dimensional linear time-invariant partial differential systems", *Proc. of 32nd Conf. Decision Contr.*, p.p. 1830-1841.

[19] T.A. Ell (1992). "Hypercomplex spectral transforms", PhD dissertation, Univ. Minnesota, Minneapolis.

[20] S.C. Pei, J.J. Ding and J.H. Chang (2001). "Efficient implementation of quaternion Fourier transform convolution and correlation by 2-D Complex FFT", *IEEE Trans. on Signal Processing*, Vol. 49, No. 11, p.p. 2783-2797.

[21] S.J. Sangwine and T.A. Ell (1999). "Hypercomplex auto- and cross-correlation of colour images", *Proc. of Int. Conf. on Image Processing,* (ICIP 1999) p.p. 319-323.

[22] T.A. Ell and S.J. Sangwine (2000). "Colour – sensitive edge detection using hypercomplex filters", (EUSIPCO 2000) p.p. 151-154.

[23] A Vanderlugt (1964). "Signal detection by complex spatial filtering", *IEEE Trans. Inf. Theory,* Vol. 10, p.p.139-145.

[24] M. Saviddes, K. Venkataramani and B.V.K. Vijaya Kumar (2003). "Incremental updating of advanced correlation filters for biometric authentication systems", *Proc. of Int. Conf. on Multimedia and Expo,* Vol. 3 (ICME 2003) p.p. 229-232.

[25] A. Mahalanobis, B.V.K. Vijaya Kumar and D. Casasent (1987). "Minimum average correlation energy filters", *Applied Optics*, Vol. 26, p.p. 3633-3640.

[26] M. Savvides, B.V.K. Vijaya Kumar and P. Khosla (2002). "Face verification using correlations filters", *Procs of 3rd IEEE Automatic Identification Advanced Technologies,* Tarrytown, N.Y., p.p. 56-61.

[27] A. Mahalanobis, B.V.K. Vijaya Kumar, S.R.F. Sims and J.F. Epperson (1994). "Unconstrained correlation filters", *Applied Optics*, Vol. 33, p.p. 3751-3759.

[28] P. Refreiger (1990). "Filter design for optical pattern recognition: multi-criteria optimization approach", Optics Letters, Vol. 15, p.p. 854-856.

[29] J.J. Buckley and Y. Hayashi (1994). "Fuzzy neural networks: A survey", *Fuzzy Sets and Systems*, 66, p.p. 1-13.

[30] C.T. Lin and C.S.G. Lee (1996). *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice Hall, Upper Saddle River, N.J.

[31] D. Nauck, F. Klawonn and R. Kurse (1997). *Foundations of Neuro-Fuzzy Systems*, Wiley, Chichester, U.K..

[32] S.K. Pal and S. Mitra (1999). *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*, Wiley, Chichester, U.K.

[33] R. Ostermark (1999). "A Fuzzy Neural Network Algorithm for Multigroup Classification", *Elsevier Science, Fuzzy Sets and Systems*, 105, p.p. 113-122.

[34] H.K. Kwan and Y. Cai (1997). "A Fuzzy Neural Network and its Application to Pattern Recognition", *IEEE Trans. on Fuzzy Systems*, 2(3), p.p. 185-193.

[35] G.Z. Li and S.C. Fang (1998). "Solving interval-valued fuzzy relation equations", *IEEE Trans. on Fuzzy Systems,* Vol. 6, No. 2, p.p. 321-324.

[36] J. Leotamonphong and S. Fang (1999). "An efficient solution procedure for fuzzy relation equations with max product composition", IEEE Trans. on Fuzzy Systems, Vol. 7, No. 4, p.p. 441-445.

[37] M.M. Bourke and D.G. Fisher (1996). "A predictive fuzzy relational controller", *Proc. of the Fifth Int. Conf. on Fuzzy Systems,* p.p. 1464-1470.

[38] M.M. Bourke and D.G. Fisher (1998). "Solution algorithms for fuzzy relational equations with max-product composition", *Fuzzy Sets Systems,* Vol. 94, p.p. 61-69.

[39] P. Xiao and Y. Yu (1997). "Efficient learning algorithm for fuzzy max-product associative memory networks", *SPIE*, Vol. 3077, p.p. 388-395.

[40] N. Troje and H. H. Bulthoff (1996). "Face recognition under varying poses: The role of texture and shape." Vision Research 36, P.p. 1761-1771. Redirected from http://faces.kyb.tuebingen .mpg.de/

[41] D.D. Lee, H. S. Seung (1999). " Learning the parts of objects by non-negative matric factorization", Nature 401, p.p. 788-791.

Vector Disambiguation for Translation Extraction from Comparable Corpora

Marianna Apidianaki
LIMSI-CNRS
Rue John von Neumann, F-91403, ORSAY CEDEX, France
E-mail: marianna@limsi.fr, http://www.limsi.fr/~marianna/

Nikola Ljubešić
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
E-mail: nikola.ljubesic@ffzg.hr, http://www.nljubesic.net/

Darja Fišer
Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
E-mail: darja.fiser@ff.uni-lj.si, http://lojze.lugos.si/darja

Keywords: word sense disambiguation, sense clustering, comparable corpora

Received: January 7, 2013

We present a new data-driven approach for enhancing the extraction of translation equivalents from comparable corpora which exploits bilingual lexico-semantic knowledge harvested from a parallel corpus. First, the bilingual lexicon obtained from word-aligning the parallel corpus replaces an external seed dictionary, making the approach knowledge-light and portable. Next, instead of using simple one-to-one mappings between the source and the target language, translation equivalents are clustered into sets of synonyms by a cross-lingual Word Sense Induction method. The obtained sense clusters enable us to expand the translation of vector features with several translation variants using a cross-lingual Word Sense Disambiguation method. Consequently, the vector features are disambiguated and translated with the translation variants included in the semantically most appropriate cluster, thus producing less noisy and richer vectors that allow for a more successful cross-lingual vector comparison than in previous methods.

Povzetek: V prispevku predstavljamo pristop za izboljšanje luščenja prevodnih ustreznic iz primerljivih korpusov z dodatnim virom leksiko-semantičnega znanja, izluščenega iz vzporednega korpusa.

1 Introduction

Due to the scarcity of general language parallel corpora, extracting translation information from comparable corpora has become a very active area of research in the past two decades. Identifying translation correspondences in comparable corpora offers low-resourced language pairs and domains a fast and affordable way to construct bilingual lexica and provides information useful for training Statistical Machine Translation systems (Munteanu and Marcu, 2005; Snover et al., 2008). The main idea behind translation extraction from comparable corpora is the assumption that a source word and its translation appear in similar contexts. n order to compare the context similarity of source and target words the same vector has to be produced, which means that the vectors of the one language have to be translated in the other language. Feature vector translation generally presupposes the availability of a bilingual dictionary (Fung, 1998; Rapp, 1999), which is however not the case for many language pairs or domains.

Another problem with the traditional approach to bilingual lexicon extraction and most of its extensions (Shao and Ng, 2004; Otero, 2007; Yu and Tsujii, 2009; Marsi and Krahmer, 2010) is that they neglect polysemy and consider a translation candidate as correct if it is an appropriate translation for at least one possible sense of the source word. This often corresponds to the most frequent sense of the word due to the way context vectors are built. An alternative to this consists in considering all translations provided for a source word in a bilingual dictionary but weighting them by their frequency in the target language (Prochasson et al., 2009; Hazem and Morin, 2012). The high quality of the information exploited by both these methods – generally found in hand-crafted resources – combined with the skewed distribution of the translations corresponding to different senses of the words, often leads to satisfying results. Nevertheless, this approach limits the usability of the proposed methods to languages and domains where such resources are available. We believe that relying on

minimal resources that can be easily obtained for any language pair and domain, and combining them with automatic disambiguation of the features in the context vectors can lead to the production of cleaner vectors and, consequently, to higher quality results during lexicon extraction from comparable corpora.

The goal of this paper is twofold: first, we wish to eliminate the need for an external knowledge source by automatically extracting a bilingual lexicon from a parallel corpus. Second, we propose a way for disambiguating polysemous features in the context vectors, as these features may be translated differently according to the sense in which they are used in a given context.

The rest of the paper is organized as follows: In the next section, we present some related work on the subject. In Section 3, we present the resources that were used in our experiments. In Section 4, we describe the approach and the experimental setup in detail. The obtained results are presented and discussed in Session 5, after which the paper is wrapped up with some concluding remarks and ideas for future work.

2 Related work

The need to bypass pre-existing dictionaries has been addressed in several works on translation information extraction from comparable corpora. Koehn and Knight (2002) build the initial seed dictionary automatically, based on identical spelling features between the two languages (English and German). Cognate detection has also been used by Saralegi et al. (2008) for extracting word translations from English-Basque comparable corpora. The cognate and the seed lexicon approaches have been successfully combined by Fišer and Ljubešić (2011) who showed that the results with an automatically created seed lexicon that is based on language similarity can be as good as with a pre-existing dictionary. But all these approaches work on closely-related languages and cannot be used as successfully for language pairs with little lexical overlap, such as English (EN) and Slovene (SL), which is the case in this experiment.

As for vector comparison, we believe we can produce less noisy vectors and improve their comparison across languages by using contextual information to disambiguate their features. This is done by a cross-lingual data-driven Word Sense Disambiguation method which assigns to each feature a cluster of semantically similar translations in the other language (Apidianaki, 2009). A similar idea has been implemented by Kaji (2003) who performed word clustering to extract sets of synonymous translation equivalents from from English-Japanese comparable corpora using pre-defined bilingual dictionaries. In addition, instead of providing one translation for each disambiguated feature, we translate it with all translation equivalents that belong to the assigned cluster similar to Déjean et al. (2005) who used a bilingual thesaurus instead of a lexicon.

The contribution of the work presented in this paper is a language independent and fully automated corpus-based approach to bilingual lexicon extraction from comparable corpora that does not rely on any external knowledge sources to determine word senses or translation equivalents.

3 Resources used

3.1 Comparable corpus

In this work, lexicon extraction is performed from a custom-built English-Slovene comparable corpus consisting of a collection of popular health and lifestyle articles from healthy-living magazines and the Internet. The core part of the corpus was collected manually from the Slovene reference corpus FidaPLUS (Arhar et al. 2007), already part-of-speech tagged and lemmatized. All the articles from the Slovene monthly health and lifestyle magazine (Zdravje) published between 2003 and 2005 have been included, amounting to one million words. For English, an equivalent amount of articles from the Health Magazine has been included. We PoS-tagged and lemmatized the English part of the corpus with TreeTagger (Schmid, 1994).

We then automatically extended the corpora from the two billion-word ukWaC (Ferraresi et al., 2008) and the 380 million-word slWaC (Ljubešić and Erjavec, 2011) that were constructed by crawling the .uk and .si domains. We took into account all the documents that pass a document similarity threshold with respect to the core corpus that was experimentally set in Fišer et al. (2011). The part of the extended corpus used in this experiment consists of 1 million words in each language.

3.2 Parallel corpus

3.2.1 Data

The information needed for applying our data-driven approach to the translation of source language vectors comes from an English-Slovene parallel corpus. Instead of an external seed lexicon used in most previous work, we translate source language vector features by exploiting the output of a cross-lingual WSD method (Apidianaki, 2009). The WSD method exploits the results of a cross-lingual Word Sense Induction (WSI) method that identifies word senses by clustering their translations in a parallel corpus. In the current setting, the English translations of Slovene words in a parallel corpus are clustered and the obtained sense clusters describe the senses of the source words.

The corpus used for sense induction is composed of the Slovene-English part of Europarl (release v6) (Koehn, 2005) and the Slovene-English part of the JRC-Acquis corpus (Steinberger et al., 2006), amounting to approximately 35M words per language.

So, the parallel corpus used for sense induction comes from a different domain than the comparable corpus described in Section 3.1. This is not the ideal scenario given that domain adaptation is important for the type of semantic processing we want to apply. There must be a noticeable shift in the senses present in the two corpora which makes the disambiguation stage harder and, in some cases, less interesting as true ambiguities become less frequent. The main reasons we opt for this configuration in this initial set of experiments are that there are very few large parallel corpora for the English-Slovene language pair, and that a comparable corpus and a gold standard needed for evaluation are available for

the health domain. Furthermore, the combination of the two EU corpora provides sufficient material for training the unsupervised word sense induction and disambiguation methods that we intend to use. We should however note that, although the corpora pertain to different domains, they do contain a lot of general vocabulary. This is the case for both the EU corpus and the health domain corpus which is not medical (in the technical sense) but more popular, built from health and lifestyle magazines.

3.2.2 Pre-processing

Prior to being used for sense induction, the parallel corpus is subject to several pre-processing steps. We first eliminate sentence pairs with a great difference in length (i.e. cases where one sentence is more than three times longer than the corresponding sentence in the other language). Next, the corpus is lemmatized and PoS-tagged with the TreeTagger (for English) and the ToTaLe tool (for Slovene) (Erjavec et al., 2010). ToTaLe uses the TnT tagger (Brants, 2000) and was trained on MultextEast corpora (Erjavec, 2012). Two part-of-speech lexicons are built containing the PoS with which each word appears in the corpus. Next, the corpus is word-aligned with GIZA++ (Och and Ney, 2003) and two bilingual lexicons are extracted from the alignment results, one for each translation direction (EN–SL/SL–EN).

Several filters are then applied to clean the lexicons from noisy alignments. The translations are filtered on the basis of their alignment score (threshold: 0.01) and their PoS, keeping for each word only translations pertaining to the same grammatical category. We retain the intersecting alignments and use for clustering only translations that translate a source word more than 10 times in the training corpus. Even if this threshold leaves out some translations of the source words, it has the double merit of reducing data sparseness issues and eliminating erroneous translations which may be found in the lexicons because of spurious alignments. The filtered EN-SL lexicon contains entries for 6,384 nouns, 2,447 adjectives and 1,814 verbs with more than three translations in the training corpus. This lexicon is exploited for Word Sense Induction, as will be explained in Section 4.

3.2.3 Gold standard

We evaluate the results of the different experiments we carry out for extracting bilingual lexicons from comparable corpora by comparing them to a gold standard lexicon, which was comparable corpus and manually inspected. The gold standard lexicon contains 187 domain terms (nouns) that are present in the source language corpus with a minimum frequency of 50. Twenty-three of these terms have two attested translations in the corpus (e.g. EN *rectum* → SL *danka, rektum)* while the rest have just one (e.g. EN *breast* → SL *dojka)*.

4 Experimental setup

4.1 Cross-lingual sense clustering

4.1.1 Vector building from the parallel corpus

The translations retained for each English target word (*w*) from the parallel corpus after the filtering process described in Section 3.2.2, are clustered on the basis of source language distributional information. Each Slovene translation *(T_i)* of *w* is characterized by a vector built from the co-occurrences of *w* in English. The vector contains the lemmas of content words (nouns, verbs and adjectives) that co-occur with *w* in the source side of the aligned sentences where it is translated by T_i, and their frequency counts. Using these vectors, pairwise similarities between the translations of *w* are calculated by a variation of the Weighted Jaccard measure (Grefenstette, 1994; Apidianaki, 2008).

For each translation T_i of *w*, let *N* be the number of features retained from the corresponding source context. Each feature F_j ($1 \leq j \leq N$) receives a total weight $tw(F_j,T_i)$ with translation T_i defined as the product of the feature's global weight, $gw(F_j)$, and its local weight with that translation, $lw(F_j,T_i)$:

$$\mathrm{tw}(F_j, T_i) = \mathrm{gw}(F_j) \cdot \mathrm{lw}(F_j, T_i)$$

The global weight of a feature F_j depends on its dispersion in the contexts of *w*. More precisely, the global weight of the feature is a function of the number N_i of translations (T_i's) to which F_j is related, and of the probabilities (p_{ij}) that F_j co-occurs with instances of *w* translated by each of the T_i's:

$$\mathrm{gw}(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{ij})}{N_i}$$

Each of the p_{ij}'s is computed as the ratio between the co-occurrence frequency of F_j with *w* when translated as T_i, denoted as $cooc\_frequency(F_j,T_i)$, and the total number of features (*N*) seen with T_i:

$$p_{ij} = \frac{\mathrm{cooc\_frequency}(F_j, T_i)}{N}$$

Finally, the local weight $lw(F_j,T_i)$ between F_j and T_i directly depends on their co-occurrence frequency:

$$\mathrm{lw}(F_j, T_i) = \log(\mathrm{cooc\_frequency}(F_j, T_i))$$

4.1.2 Similarity calculation

The weights assigned to the features by the Weighted Jaccard measure reflect their relevance for calculating the similarity of the translation vectors. The score assigned to a pair of vectors indicates the degree of similarity of the corresponding translations. Translation pairs with a score above a threshold defined locally for each *w,* and dependent on the similarity scores assigned to its pairs of translations, are considered as semantically related.

Language	PoS	Source word	Slovene sense clusters
EN-SL	Nouns	sphere	{krogla} (*geometrical shape)* {sfera, področje} (*area*)
		address	{obravnava, reševanje, obravnavanje} (*dealing with*) {naslov} (*postal address*)
		portion	{kos} (*piece*) {obrok, porcija} (*serving*) {delež} (*share*)
		figure	{številka, podatek, znesek} (*amount*) {slika} (*image*) {osebnost} (*person*)
	Verbs	seal	{tesniti} (*to be water-/airtight)* {zapreti, zapečatiti} (*to close an envelope or other container)*
		weigh	{pretehtati} (*consider possibilities)* {tehtati, stehtati} (*check weight)*
		educate	{poučiti} (*give information)* {izobraževati, izobraziti} (*give education)*
		consume	{potrošiti} (*spend money/goods)* {uživati, zaužiti} (*eat/drink*)
	Adjectives	mature	{zrel, odrasel} (*adult*) {zorjen, zrel} (*ripe*)
		minor	{nepomemben} (*not very important)* {mladoleten, majhen} (*under 18 years old)*
		juvenile	{nedorasel} (not *adult/biologically mature yet)* {mladoleten, mladoletniški} (*not 18/legally adult yet)*
		remote	{odmaknjen, odročen} (*far away and not easily accessible)* {oddaljen, daljinski} (*controlled from a distance)*

Table 1: Examples of nominal, verbal and adjectival entries from the English-Slovene sense cluster inventory.

The similarity threshold is set following the method proposed in Apidianaki and He (2010). This iterative procedure permits to define a local threshold for each *w* and to avoid using a static threshold that might not be appropriate for different words. The threshold (*T*) for a word *w* is initially set to the mean of the scores (above 0) of the translation pairs of *w*. The translation pairs of *w* are then divided into two sets (G_1 and G_2) according to whether they exceed, or are inferior to, the threshold. Then, the average of the scores of the translation pairs in each set is computed (m_1 and m_2) and a new threshold is created that is the average of m_1 and m_2 ($T = (m_1 + m_2)/2$). The new threshold serves to separate once again the translation pairs into two sets, a new threshold is calculated and the procedure is repeated until convergence is reached.

The similarity threshold calculated in this way permits to estimate the semantic proximity of the translations. Once this is done, the clustering algorithm groups the semantically similar Slovene translations into 'sense-clusters' describing the senses of the corresponding English words.

4.1.3 Translation clustering

The clustering algorithm takes as input the list of translations of the English word, their similarity scores and the similarity threshold, and outputs clusters of semantically related translations of the word in the target language. The clustering is performed in two steps. First, each translation pair with a similarity score exceeding the threshold is considered to have a pertinent relation and forms a cluster. The obtained two-element clusters might be enriched, during the second clustering step, by additional translations that are semantically related to all the translations already in the cluster. The clustering stops when all translations are included in some cluster and all their relations have been checked. All the elements in the final clusters are linked to each other by strong semantic relations, similar to cliques in undirected graphs.

Table 1 provides examples of clusters for English words of different PoS with clear sense distinctions in our training corpus. For each English word, we give the obtained clusters of Slovene translations, including a description of the sense described by each cluster.

For instance, the translations *krogla, sfera* and *področje* of the word *sphere* are grouped into two sense-clusters {*krogla*} and {*sfera, področje*} which describe the two senses of *sphere* observed in the corpus: "*geometrical shape*" and "*area*". Similarly, the translations retained for the adjective *minor* from the training corpus (*nepomemben, mladoleten* and *majhen*) are grouped into two clusters describing its two senses*:* {*nepomemben*} - "*not very important*" and {*mladoleten, majhen*} - "*under 18 years old*". The resulting cluster inventory contains 13,352 clusters in total, for 8,892 words. 2,585 of the words (1518 nouns, 554 verbs and 513 adjectives) have more than one cluster.

4.2 Vector building from the comparable corpus

Context vectors in both the source and the target language are built for nouns occurring at least 50 times in the comparable corpus. This frequency threshold is

required in order to obtain enough contextual data and ensure minimally reliable results in the lexicon extraction process.

As features in context vectors, we use three content words to the left and to the right of the retained nouns, stopping at the sentence boundary. The position of each content word is not taken into account, i.e. the context is seen as a bag of words. Our previous research (Fišer and Ljubešić, 2011; Ljubešić et al., 2011) has shown that encoding feature positions is mostly useful only when extracting translation candidates between closely related, syntactically similar languages.

Feature weights are calculated by the TF-IDF measure. TF is calculated as the relative frequency of a content word feature regarding all content word features in a specific context vector. IDF weights are calculated on the whole ukWaC and slWaC corpora in a typical IR manner by obeying document boundaries. Our previous research (Ljubešić et al., 2011) has shown that TF-IDF feature weights perform as good as the more complex log-likelihood weighting and better than pure relative frequency. These feature weights serve additionally to filter out 'weak' features that are shown not to be useful for the lexicon extraction task (see Section 5.2).

4.3 Vector disambiguation

4.3.1 A data-driven approach

In order to identify the translations of the source words in the target language side of the comparable corpus, the vectors built in the two languages must be compared. This comparison serves to quantify the similarity of the source and target language words represented by the vectors, and the highest ranked pairs are proposed as entries for the lexicon.

As the vectors have been built from monolingual corpora, the source language vectors must first be translated into the target language. As explained above, in most previous work on bilingual lexicon building from comparable corpora, the vectors were translated using external seed dictionaries. The first translation proposed for a word in the dictionary was used to translate all the instances of the word in the vectors irrespective of their sense, and no disambiguation was performed.

The use of external resources ensures the quality of the translations used for translating the source vectors. Moreover, the selection of the most frequent translation often results in good translations because of the skewed distribution of the translations corresponding to different senses of the words. Nevertheless, this technique limits the usability of the proposed lexicon extraction methods to languages and domains where such resources are available.

In this work, instead of using an external bilingual dictionary, we translate the source language vectors using the data-driven cross-lingual WSD method proposed by Apidianaki (2009). The method exploits the sense clusters acquired from parallel corpora by the sense induction method described in Section 4.1. This property extends the applicability of the method to languages lacking large-scale lexical resources but for which parallel corpora are available.

4.3.2 Cross-lingual WSD

The sense clusters of translations obtained during sense induction (cf. Section 4.1) represent the candidate senses of the English words in the parallel corpus. We exploit this sense inventory for disambiguating the features in the English vectors that were extracted from the comparable corpus. More precisely, the WSD method has to select for each feature in the vectors built from the comparable corpus, the cluster that correctly translates its sense in the target language.

In the current setting, the selection is performed by comparing information from the context of the vector features to the distributional information that served to estimate the semantic similarity of the clustered translations. The context of a feature to be disambiguated corresponds to the rest of the vector where it appears. Inside the vectors, the features are ordered according to their weight (calculated as explained in Section 4.1). The feature weights serve to filter out the *weak* features (i.e. features with a score below a threshold) which were shown not to be useful for the lexicon extraction task. The threshold was experimentally set at 0.01. The retained features are then considered as a bag of words.

On the clusters side, the information used for disambiguation is found in the source language vectors built from the parallel corpus which revealed the semantic similarity of the clustered translations. If common features (*CF*'s) are found between the context of a feature and just one cluster, this cluster is selected to describe the feature's sense. Otherwise, if there exist *CF*'s with more than one cluster, then a score is assigned to each 'cluster-feature' association. This weight corresponds to the mean of the weights of the *CF's* relative to the clustered translations (weights assigned to each feature during clustering). In the following formula, CF_j is the set of *CF*'s found between the cluster and the new context and N_{CF} is the number of translations T_i in the cluster characterized by a *CF*:

$$assoc\_score = \frac{\sum_{i=1}^{N_{CF}} \sum_j w(T_i, CF_j)}{N_{CF} \cdot |CF_j|}$$

The highest scored cluster is selected and assigned to the feature as a sense tag. The features are also tagged with the most frequent (MF) translation of the word in the parallel training corpus, which sometimes already exists in the cluster selected during WSD.

In Table 2, we present some examples of disambiguated vector features of different PoS. For each case, we provide: the headword entry to which the vector corresponds; a feature from the vector that has been disambiguated (a noun, a verb and an adjective, respectively, in the three examples); and the context that was used for disambiguation, which consists of the other strong features found in the same vector (i.e. features with a weight above the threshold). From the candidate clusters available for the feature (given in column 4), the WSD method selects the most appropriate one (in boldface) to describe the feature's sense in this context. In the last column of the table, we provide the most frequent sense/translation (MF) for the feature.

Headword	Feature (PoS)	Context	Candidate clusters	MF alignment
infertility	treatment (n)	*doctor, diabetes, health, emergency, check, ...*	- {**zdravljenje, obdelava, obravnavanje, obravnava, ravnanje}** (*treat an illness)* - {čiščenje} (*treat a person/animal)* - {raba} (*usage*)	obravnava
clot	seal (v)	*block, heart, vessel, pressure, infection, ...*	- {**tesniti**} (*to be waterproof or airtight)* - {zapreti, zapečatiti} (*to close)*	zapečatiti
arrhythmia	irregular (a)	*heart, abnormal, monitor, failure, risk, ...*	- {**nepravilen, nereden}** (*not regular)* - {ilegalen} (*illegal*)	nepravilen

Table 2: Disambiguation results.

We observe that the MF translation may already exist in the cluster selected by the WSD method, like in the first example where *obravnava* is already in the selected cluster. The inverse, i.e. that the MF is not found in the proposed cluster, is also possible as is the case with the *zapečatiti* translation of the verb *seal.*

The disambiguation of source language features using cross-lingual sense clusters constitutes the main contribution of this work and presents several advantages. First, the method performs disambiguation by using sense descriptions derived from the data, which extends its applicability to resource-poor languages. This procedure clearly differentiates our method from previous approaches where the first translation in a dictionary – which is often the most frequent one – was selected for translating each vector feature. An additional advantage is that the sense clusters assigned to features may contain more than one translation. This property is important in this setting as it provides supplementary material for the comparison of the vectors in the target language.

Cross-lingual vector comparison

The translation of the source vectors into the target language, performed as described in the previous section, makes possible the comparison of the vectors in the same vector space. We experiment with three different ways of translating features:

1. by keeping the translation a feature was most frequently aligned to in the parallel corpus (MF);
2. by keeping the most frequent translation from the cluster assigned to the feature during disambiguation (CLMF); and
3. by using the same cluster as in the second approach, but producing features for all translations in the cluster with the same weight (CL).

The first approach is used as a baseline since instead of the sense clustering and WSD results, it just uses the "most frequent sense/alignment" heuristic. In the first batch of the experiments, we noticed that the results of the CL approach heavily depend on the part-of-speech of the features. So, we divided the CL approach into three sub-approaches:

1. translate only nouns with the clusters and other features with the MF approach (CL-n);
2. translate nouns and adjectives with the clusters and verbs with the MF approach (CL-na); and
3. translate all PoS with the clusters (CL-nav).

The distance between the translated source and the target-language vectors is computed by the Dice metric which has proven to be very efficient when combined with the TF-IDF weighting (Ljubešić et al., 2011).

During our experiments, we noticed that discarding the weakest features from the context vectors in the source language significantly improves the results. So, we also experiment with a minimum feature weight threshold and call this parameter the 'minimum feature weight threshold' (mfwt). By comparing the translated source vectors to the target language ones, we obtain a ranked list of candidate translations for each gold standard entry.

5 Evaluation and discussion of the results

5.1 Evaluation setting

The final result of our method consists in ranked lists of translation candidates for gold standard entries. We evaluate this output by the mean reciprocal rank (MRR) measure which takes into account the rank of the first good translation found for each entry. Formally, MRR is defined as

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where $|Q|$ is the length of the query, i.e. the number of gold standard entries we compute translation candidates for, and $rank_i$ is the position of the first correct translation in the candidate list.

Since most of the entries in our gold standard contain just one translation, we did not consider using more advanced evaluation measures for ranked results, like mean average precision (MAP).

5.2 Results and discussion

The results of our final experiment are shown in Figure 1. The *x* axis shows the minimum feature weight threshold (mfwt) while on the *y* axis the evaluation measure MRR is plotted.

The phenomenon that is first observed in the graph is the one for which we have introduced the minimum feature weight threshold parameter: the best results are obtained when discarding all features that have a TF-IDF weight score lower than 0.01. This is something we had not noticed before and that we intend to explore more thoroughly in a new set of experiments, by measuring its consistency when different weight measures, distance measures, seed lexicons, language pairs and comparable corpora are used.

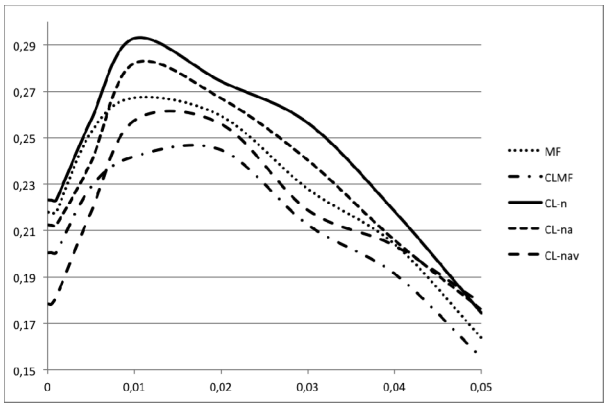


Figure 1: Evaluation of different approaches to lexicon extraction.

The lowest results are consistently obtained when using the CLMF approach, which consists in using only the most frequent translation from the cluster chosen through the WSD procedure. A possible reason for this is the fact that alignment frequencies used for finding the most frequent translation in the cluster were calculated on a corpus of a different domain than our comparable corpus (Europarl vs. health corpus).

The baseline which always uses the most frequent translation of the feature from the parallel corpus, without sense clustering and WSD, achieves a medium result. The baseline is outperformed by the CL-n and the CL-na approaches but performs better than the CL-nav approach, which shows that taking verbs into account deteriorates the quality of the results.

The different CL approaches yield somewhat expected results. The biggest gain is obtained from clustering and WSD information calculated on nouns, nouns and adjectives scored second and the lowest results are obtained when verbs are added to the mix. This is probably due to the fact that the verbal clusters are noisier than the nominal and adjectival ones. We intend to further explore this issue.

Since our gold standard is quite small, we checked the statistical significance of the difference in the results of the baseline MF approach and the winning CL-n approach. We used the approximate randomization procedure with R = 1000 (i.e. 1000 random assignments were done without replacement of the two sets of results). The resulting *p-value* is 0.091, which is higher than the commonly used 0.05 threshold.

These results show that in our future experiments we will need a larger gold standard to draw safer conclusions on the statistical significance of the results. However, since the p-value is below 0.1 and is accompanied by a consistent increase in performance throughout a large number of experiments, we are rather confident that this increase is not the result of random variation.

The main conclusions that can be drawn from the reported results here are the following:

- extending the feature set with multiple translations obtained by sense clustering and word sense disambiguation of features is beneficial to the lexicon extraction procedure;
- the most valuable information obtained from the clustering and WSD approach comes from nouns;
- using just the most frequent translation inside the cluster selected during WSD does not yield good results; and
- further investigation of the improvement that occurs when weak features are discarded is needed.

6 Conclusions and future work

We presented an approach that allows the use of lexico-semantic knowledge acquired from parallel corpora to improve the extraction of translation equivalents from comparable corpora. A parallel corpus served as the source of the seed dictionary, so that no external knowledge source is needed for the translation of features in context vectors. In addition, the seed dictionary was enhanced with clusters of translation variants obtained from the parallel corpus in an unsupervised way. The cross-lingual clusters were used to disambiguate the features in the context vectors, reducing noise, and allowed for a more accurate comparison of source and target vectors. Furthermore, the tagging of the vector features with clusters during disambiguation increased the translation information available for each feature and, therefore, facilitated the comparison of context vectors across languages.

The results show that lexico-semantic knowledge derived from a parallel corpus can help to circumvent the need for an external seed dictionary, traditionally considered as a pre-requisite for bilingual lexicon extraction from parallel corpora. Moreover, disambiguating the vectors improves the quality of the extracted lexicons and manages to beat the simpler, if powerful, most frequent sense/alignment heuristic.

These encouraging results pave the way towards pure data-driven methods for bilingual lexicon extraction from comparable corpora. This knowledge-light approach can be applied to languages and domains that do not dispose of large-scale seed dictionaries but for which parallel corpora are available. Moreover, the use of a data-driven cross-lingual WSD method, such as the one proposed in this paper, can contribute to obtain less noisy translated vectors, which is important especially when lexicon extraction is performed from general language comparable corpora.

The experiments carried out till now focus on a health comparable corpus. Although this is not a very specialized corpus but a rather popular one, cases of true polysemy are still less frequent than in a general corpus.

We would thus like to extend this work by applying the method to a more general comparable corpus, for instance a corpus built from Wikipedia texts. We expect that the effect of applying the WSD method on a general corpus will be highly beneficial, as ambiguity problems will be more prevalent.

We also want to explore the use of second order co-occurrences for disambiguation. For the moment, the context used to disambiguate vector features consists of other features that appear in the same vector. However, these features are direct co-occurrences of the headword, which does not necessarily mean that the features themselves co-occur with each other in the corpus. We consider that it would be preferable to replace this context with the co-occurrences of the features in the corpus for disambiguation, which would correspond to the second order co-occurrences of the English words, and investigate the effect of using this type of context on lexicon extraction.

References

[1] Marianna Apidianaki (2008) Translation-oriented Word Sense Induction based on Parallel Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

[2] Marianna Apidianaki and Yifan He (2010) An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT),* Paris, France, 219–226.

[3] Marianna Apidianaki (2009) Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece, 77–85.

[4] Špela Arhar, Vojko Gorjanc and Simon Krek (2007) FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools. In *Proceedings of the Corpus Linguistics conference*, Birmingham, UK.

[5] Thorsten Brants (2000) TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP),* Seattle, WA.

[6] Hervé Déjean, Eric Gaussier, Jean-Michel Renders and Fatiha Sadat (2005) Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine,* 33(2):111–124.

[7] Tomaž Erjavec (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

[8] Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. (2010) The JOS linguistically tagged corpus of Slovene. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC)*,Valletta, Malta.

[9] Adriano Ferraresi, Eros Zanchetta, Marco Baroni and Sylvia Bernardini (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4): Can we beat Google?,* Marrakech, Morocco, 47–54.

[10] Darja Fišer and Nikola Ljubešić (2011) Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP),* Hissar, Bulgaria, 125–131.

[11] Darja Fišer, Nikola Ljubešić, Špela Vintar and Senja Pollak (2011) Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th BUCC Workshop: Comparable Corpora and the Web,* Portland, Oregon, USA, 19–26.

[12] Pascale Fung (1998) Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora. *Lecture Notes in Artificial Intelligence*, Springer, Vol. 1529, 1–17.

[13] Gregory Grefenstette (1994) *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Norwell, MA.

[14] Amir Hazem and Emmanuel Morin (2012) ICA for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 5th Building and Using Comparable Corpora (BUCC) workshop,* Istanbul, Turkey, 126-133.

[15] Hiroyuki Kaji (2003) Word sense acquisition from bilingual comparable corpora. In *Proceedings of HLT-NAACL*, Edmonton, Canada, 32–39.

[16] Philipp Koehn and Kevin Knight (2002) Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition,* 9–16.

[17] Philipp Koehn (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X,* Phuket, Thailand, 79–86.

[18] Nikola Ljubešić and Tomaž Erjavec (2011) hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Proceedings of Text, Speech and Dialogue (TSD),* Lecture Notes in Computer Science (LNCS) Vol. 6836, Springer, 395–402.

[19] Nikola Ljubešić, Darja Fišer, Špela Vintar and Senja Pollak (2011) Bilingual lexicon extraction from comparable corpora: A comparative study. In *Proceedings of the International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia.

[20] Erwin Marsi and Emiel Krahmer (2010) Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING),* Beijing, China, 752–760.

[21] Dragos Stefan Munteanu and Daniel Marcu (2005) Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

[22] Franz Josef Och and Hermann Ney (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* 29(1):19–51.

[23] Pablo Gamallo Otero (2007) Learning bilingual lexicons from comparable English and Spanish corpora. In *Proceedings of Machine Translation (MT) Summit XI*, Copenhagen, Denmark, 191–198.

[24] Emmanuel Prochasson, Emmanuel Morin and Kyo Kageura (2009) Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Machine Translation Summit (MT Summit XII)*, Ottawa, Ontario, Canada, 284-291.

[25] Reinhard Rapp (1999) Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL),* College Park, Maryland, USA, 519–526.

[26] Xabier Saralegi, Iñaki San Vicente, Antton Gurrutxaga (2008) Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the 1st Building and Using Comparable Corpora (BUCC) workshop,* Marrakech, Morocco.

[27] Helmut Schmid (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing,* Manchester, UK, 44–49.

[28] Li Shao and Hwee Tou Ng (2004) Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING),* Geneva, Switzerland, 618–624.

[29] Matthew Snover, Bonnie Dorr and Richard Schwartz (2008) Language and Translation Model Adaptation using Comparable Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawai, 857-866.

[30] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş and Dániel Varga (2006) The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2142–2147.

[31] Kun Yu and Junichi Tsujii (2009) Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. In *Proceedings of NAACL/HLT 2009*, Boulder, Colorado, USA, 121–124.

Design Science Perspective on NFC Research: Review and Research Agenda

Mehmet N. Aydin
Faculty of Engineering and Natural Sciences, Kadir Has University, Istanbul
Kadir Has Caddesi Cibali / Istanbul 34083, Turkey
E-mail: mehmet.aydin@khas.edu.tr

Busra Ozdenizci
Department of Information Technologies
Işık University, Istanbul, Turkey
E-mail: busraozdenizci@isikun.edu.tr

Keywords: near field communication, design science research, review

Received: January 23, 2013

Near Field Communication (NFC), as one of the emerging and promising technological developments, provides means to short range contactless communication for mobile phones and other devices alike. NFC has become an attractive design science research area for many academicians due to its exploding growth and its promising applications and related services. A better understanding of the current status of NFC research is necessary to maintain the advancement of knowledge in NFC research and to identify the gap between theory and practice. In this paper, we present a literature review on NFC. To facilitate the analysis of the literature, we propose a research framework and organize the NFC literature into four major categories (theory and development, applications and services, infrastructure, ecosystem). We contend that due to the nature of NFC (industry high stakes, multidisciplinary research, artifacts development), the design science research paradigm serves an appropriate ground to investigate an extent to which relevance and rigor is achieved. By employing the proposed research framework and design science perspective, we set up a research agenda (research directions and promising research questions) which may help practitioners and academics to achieve a substantial progress in NFC.

Povzetek: Predstavljen je strokovni okvir za NFC, komunikacijo kratkega dosega.

1 Introduction

Today the rapid development and adoption of information technologies (IT) is changing the way of doing business significantly. The growing interest on electronic commerce to perform business transactions brought vital improvements, especially in wireless technologies [80]. Near Field Communication (NFC) has become one of the promising wireless technological developments in the information and communication industry. NFC technology is a short-range, high frequency, low bandwidth radio technology. It allows us to transfer data within few centimeters. As shall be discussed later on, along with three operating modes (reader/writer, peer-to-peer and card emulation [81]), key advantages of NFC over other wireless technologies include simplicity and inherent security [13, 19]. The integration of NFC technology into mobile devices offers many reliable applications such as payment, ticketing, loyalty services, identification, access control, content distribution, smart advertising, peer-to-peer data/money transfers, and set-up services [85].

NFC has become an attractive research area for many academics due to its exploding growth and its promising applications and related services. Noticeably, for the last few years, there has been a considerable amount of increase in the number of research papers and activities concerning NFC. However, a better understanding of the current status of NFC research area is necessary to maintain the advancement of knowledge in NFC research and to identify the progress of NFC research. Thus, a literature research framework is necessary to fulfill the needs. In the present research, such a framework is established and used to make sense of NFC endeavors and to propose promising research directions with a number of research questions.

Scholars, including [83], address a relevance issue in information systems (IS) research and emphasize an importance of studying information technology (IT) artifacts as design science research (DSR). [92] maintains that DSR enables a focus on the IT artifact with a high priority on relevance in an application domain. In this regard, NFC as an innovative artifact

exemplifies the central role of the IT artifacts in IS research. As shall seen later on, most of the research on NFC yields such artifact types as constructs, models, methods and instantiations [86]. Thus, to examine the progress of NFC research one needs to examine how well rigor and relevance is achieved and what research issues need to be addressed. We contend that due to the nature of NFC (industry high stakes, interdisciplinary research, artifacts development), design science research paradigm serves an appropriate ground to investigate an extent to which relevance and rigor is achieved. By employing the proposed research framework and design science perspective, we set up a research agenda which may help practitioners and academics to achieve a substantial progress in NFC.

The contribution to this study is two-fold. First, it goes beyond a typical literature review and establishes a framework by which we articulate the status of Body-of-Knowledge for NFC. By employing a DSR perspective, the paper proposes a research agenda and brings up promising research questions. Second, the paper contributes to IS research by showing how DSR can be used to examine the progress of NFC as an emerging research field.

The paper is organized as follows. First, we clarify what the basis of this research is and what relevant research is used to explicate the research rationale in this paper. Second, we present the research approach and method adopted to establish the framework and set up the research agenda. Third, the framework is proposed and used to explicate the BoK for NFC. Fourth, the DSR perspective with a number of criteria is used to examine the progress of NFC. Fifth, the research agenda is provided to support academics and practitioners and, finally, the conclusion is drawn.

2 Relevant Research and Methodology

2.1 Organizing Frameworks in Relative Research Areas

[84] maintain that an effective review is essential to create a solid foundation for advancing knowledge. Such a foundation provides a reference Body-of-Knowledge (BoK) and facilitates both an academic progress and effective use of research outcomes.

Reviewing academic literature for an emerging research area like NFC is a challenge because the accumulated knowledge may not be mature enough for synthesis. On the other hand, it is necessary as to one can make sense of existing research endeavors and relate ongoing research to the BoK. Such a review work about the NFC research area has not been performed so far rigorously. The present research attempts to fulfill this need.

While we determine the basis of our review, we need to look into those review studies which can contribute to establishment of the NFC framework. Thus, we examine review studies in IS in general, NFC relevant reviews in particular. As shall seen in the next section, we used the former to determine a review approach, that is what review approach should be adopted for examining NFC. The latter includes review studies on electronic commerce (e-commerce), RFID or any wireless technology related topics, and is important to elaborate in a such way that subject-specific insights can be gained and may help in determining the organizing framework. We shall discuss briefly what and how frameworks have been established in representative studies.

Regarding with electronic commerce (e-commerce), it is broader, yet helps in identifying a relevant research area to NFC which provides a sense of organizing boundary for implications of NFC with respect to such perspectives as business, organization, technology. Indeed, one can find several review studies and frameworks on electronic commerce in terms of essential concepts along with these perspectives. For instance, in [3] and [4], the proposed research frameworks are based on four dimensions (applications, technology, support, and implementation along with other issues).

Likewise, mobile commerce (m-commerce) literature reviews are also good sources for understanding the implications of mobile technology on modifying existing e-commerce frameworks [3]. [2] identified the gaps between theory and practice and future research directions for m-commerce papers through a well structured classification framework and analyses.

[87] conduct one of the prominent survey studies on wireless technologies. The organizing framework maintains high level conceptions on underlying notions, characterization of types of applications, design principles and architecture issues. Such an overarching survey concludes with summarizing existing research attempts and the very need of this technology for further development and use in practice.

Regarding review studies on Radio Frequency Identification (RFID), as a related technology to NFC, [1] organized studies as "technological issues, applications areas, policy and security issues, and other issues". As stated in [3], such a study is considered to be a reference study for those researchers interested in this area.

While the examined review studies are useful for determining an organizing framework, the key questions still remain as follows: how to develop such a framework for NFC? What theoretical perspective helps to examine and facilitate the progress of NFC?

2.2 NFC as Design Science Research

Upon the establishment of a research framework, one needs to examine the progress and opportunities for NFC. To do this, we seek to identify an appropriate research perspective. Thanks to recent discussions on prominent orientations in conducting IS research, which is about behavior- versus design oriented research [92]. The discussions appear to be escalated in recent issues of top IS journals such as MISQ, ISR, and EJIS where scholars, including [83], argue origination and values of

design-oriented IS. This present research does not delve into philosophical argumentations for research perspectives, rather aims to get the most out of the rich discussion on how appropriate research perspective may benefit examining progress of NFC.

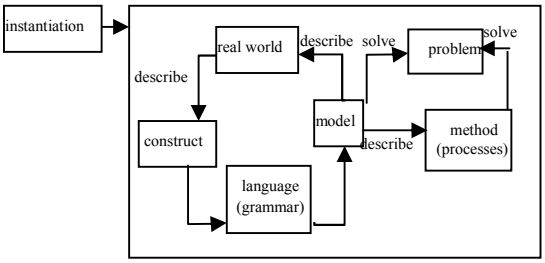


Figure1: IT artifact instantiation (adopted from [86])

In seeking an appropriate research perspective, it appears that NFC is an excellent research area to exemplify design-oriented research characteristics. First and foremost characteristic is the research focus on IT artifacts. It can been seen that BoK for NFC is mainly dominated by studies focusing on emerging and innovative artifacts (see section 3). This is no surprise since prevailing research motivation in NFC research area is problem-solving oriented and results in type of artifact - that is, constructs, models, methods and instantiations (See figure 1). [82] shows how instantiation inherits complex relations among *construct* (deriving from and a real world phenomenon and leading too language), *model* (formulated by a language and representing the problem under investigation), and *method (*explicating the process of achieving the solution). We shall explicate these relations with some illustrative examples, but there is one thing to note that a real world is primarily a triggering source. In the context of NFC research, this source was evident that industry leaders such as Nokia, Philips and Sony jointly developed NFC as an alternative or complementary communication model to overcome issues with wireless technologies such as RFID, Bluetooth [11].

Noticeably, the focus on IT artifacts has a lot to do with a design rationale and aims to solve a particular problem which brings up the value of relevant research. This is the secondary characteristic in that design oriented research strives for high relevance by examining an extent to which proposed artifacts meet expected utility. In recent years including [81, 82, 83], scholars in the IS research domain have raised the issue of lacking relevance. Regarding the need for relevance, the difference of behavioral science and design science research should not be considered as dichotomy, but complementary approaches with differing research rationale. [86] suggests that while behavioral IS research aims at 'truth', i.e., at the exploration and validation of generic cause–effect relations, IS design science research aims at 'utility', i.e., at the construction and evaluation of generic means–ends relations. That is, the notion of relevance is equally important matter for design and behavior research. Thirdly, design science research may benefit from a systematic process of IT artifact development (e.g., deductive or inductive) at higher abstraction, which in turn contributes substantially to the structuring and integration of the body of knowledge.

Guideline	Description	DSR Cycles	Key Questions
Guideline 1. Design as an artifact	Design science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation	Design Cycle	What is the artifact? How is the artifact represented?
Guideline 2. Problem relevance	The objective of design science research is to develop technology-based solutions to important and relevant business problems	Relevance Cycle	What is the research question (design requirements)? Has the research question been satisfactorily addressed?
Guideline 4. Research contributions	Effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies	Rigor Cycle	What new knowledge is added to the knowledge base and in what form (e.g..peer-reviewed literature, meta-artifacts, new theory, new method)?
Guideline 3. Design evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods	Relevance	How is the artifact introduced into the application environment and how is it field tested? What metrics are used to demonstrate artifact utility and improvement over previous artifacts?
Guideline 5. Research rigor	Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact	Rigor Cycle and Design Cycle	What design processes (search heuristics) will be used to build the artifact? How are the artifact and the design processes grounded by the knowledge base? What, if any, theories support the artifact design and the design process?
Guideline 6. Design as a search process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment		
Guideline 7. Communication of research	Design science research must be presented effectively to both technology-oriented and management-oriented audiences	Relevance Cycle	No specific questions identified.

Table 1: DSR Guidelines, Cycles and Checklist (adopted from [92])

As stated in [92], design science is inherently a problem solving process that creates and evaluates IT artifacts intended to solve identified organizational problems. They provide seven critical guidelines for researchers to achieve effective design-science research in Information Systems (IS). Later on, [92] introduces three cycles and checklist questions to make the guidelines more operational in empirical sense.

The relevance cycle refers to how research is initiated in light of application context so that the requirements for the research as inputs and as well as for acceptance criteria are explicitly defined. The rigor cycle is concerned with knowledge related to both experience and expertise defining the state of art in the application domain and artifacts, processes. The design cycle indicates actual artifact development and its evaluation. In Table 1, we relate guidelines to cycles and checklist questions.

As the research cycles indicate, knowledge and understanding of design science research guidelines is the critical part of our research study. In fact, these guidelines are not mutually exclusive. In accordance with [92], the first requirement is that design science research has to provide an innovative, purposeful *design artifact* in the form of a construct, a model, a method, or an instantiation. The design artifact has to solve a specific problem or to develop technology based solutions which is refers to *problem relevance* as the second requirement. Indeed, these two guidelines generally mentioned in a typical design science paper due to their nature. *Design evaluation* as the third requirement maintains the evaluation of utility, quality, and efficiency of the proposed design artifact through observational, analytical, experimental, testing or descriptive methods [92]. In our assessments, we mainly focused on which techniques for design evaluation were used in detail, and the quality of the design evaluations.

In essence, the design artifact itself must be *rigorously* defined, formally represented. *Applicability and generalizability of the artifact* has to be mentioned explicitly which is the sign of research rigor. Such a rigorous research work with clear contributions and efficient *design evaluations* has to facilitate a search process (i.e. the search for the best or optimal design artifact). Furthermore, the proposed design artifact must be presented both to *technology-oriented as well as management-oriented audiences* [92]; each side needs sufficient detail about the design artifact. Such communication of design science research provides repeatability of the proposed artifact and further research works for technology oriented audiences. At the same time, management oriented audiences appreciate such an artifact's nature, make assessments within their specific organizational context.

In later sections, these guidelines and checklist questions are used to examine NFC studies and induce a research agenda in light of three cycles. For an illustration purpose and contextualizing design science guidelines, consider the following three NFC studies which are examined from the design science guidelines (see Table 2).

2.3 Research Methodology

Building a research framework requires identification of essential characteristics for NFC. The literature review and relevant organizing framework studies serve a good basis to induce a framework. [84] state that the literature review is expected to answer questions such as: What are the key theories, concepts and ideas?, How is knowledge on the topic structured and organized? What are the major issues and debates about the topic? How have approaches to these questions increased our understanding and knowledge? In IS literature, several examples such as [81] can be found where reviews are often concept centric, which is also the case in this study.

Namely, core concepts underlying the research matter are used to determine the organizing framework.

Since NFC is a rather emerging technology, research papers on NFC are relatively recent. First NFC related papers appear in the scientific publication in 2005. Thus

Guidelines	**Keywords**	**[14]**	**[79]**	**[80]**
Guideline 1. Design as an Artifact	Constructs, Models, Practices, Representations, Methods, Instantiations, Prototypes	Platform to securely manage smartcard applications in NFC devices	Prototype of a snowboarder community platform	NFC application to support health monitoring
Guideline 2. Problem Relevance	Problem Solving, Optimization, Profit Maximization	Clearly mentioned; need for secure management	Mentioned: for social interaction and provides product information	Clearly mentioned: the requirements; providing accurate measurement devices
Guideline 3. Design Evaluation	Observational (Case Studies), Analytical, Experimental, Functional or Structural Testing, Descriptive (Scenarios)	Not evaluated, only implications of the platform	Not evaluated; implications of use cases are mentioned	Not clearly mentioned
Guideline 4. Research Contributions	New Metrics, System Development Methodologies, Design Tools, Prototypes or Improvement of Existing Foundations	Not explicit	Not explicit	Clear and verifiable contributions are provided
Guideline 5. Research Rigor	Applicability, Generalizability, Appropriateness, Feasibility of the Design Artifact, Well Design Evaluations	To some degree	Not explicit	Not explicit
Guideline 6. Design as a Search Process	Iterative Process, Searching for The Best, Optimal Design, Future Work or Studies	Facilitates search process	Facilitates search process	Facilitates search process, needing more future technical study
Guideline 7. Communication of Research	Communication to both audiences; Managerial and Technology Oriented Audiences	Communicates all types of audiences	Communicates all types of audiences	Communicates all types of audiences

Table 2: Exemplary NFC Studies Examined by a Design Science

the scope of this survey is limited to the time frame of 2005-December 2011; this period is considered as the representative NFC literature.

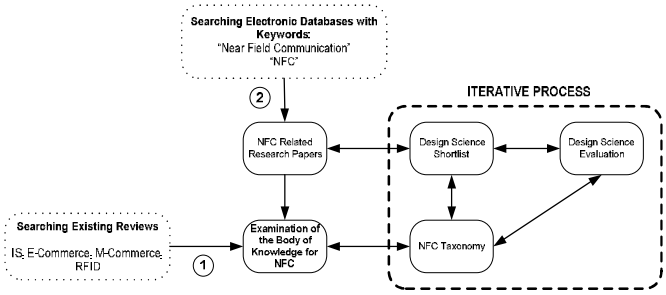


Figure 2: Search Strategy

Our literature review includes articles in journals and conference proceeding papers (especially, published by IEEE, ACM, and other academic associations). We exclude master's theses, doctoral dissertations, textbooks, unpublished working papers, and white papers. Researchers and practitioners often use journal papers to acquire information and to disseminate new research findings [4], thus most of the existing literature reviews exclude conference proceeding papers, too. However, we did not exclude conference papers in our literature review as the proceeding papers provide also a high level of research, both in width and breadth after journals. At the same time, we exclude some writings those are published as editorials, industry and news reports or book reviews.

After performing the search for the papers as defined above, we have found 202 articles (see figure 2, Step 2). The literature search was based on two descriptors; "NFC" and "Near Field Communication". It was conducted using the following electronic databases:

1. IEEE/IEE Electronic Library
2. Association for Computing Machinery
3. ISI Web of Knowledge
4. Academic Search Complete
5. Computer and Applied Science Complete
6. Science Direct
7. Emerald Full Text

By using the academic sources above, we listed all studies related to NFC along with their relevance. After the collection of 202 NFC related papers, a shortlist from these studies is created for a design science evaluation; 25 studies were selected by two researchers. Two strategies were followed during the selection of studies for a design science evaluation; elimination of similar papers in terms of *topic coverage, varieties* and selection of the papers which cover the subjects *in-depth*. To illustrate how strategies have been implemented, consider [48] [73]. These two studies basically focus on NFC applications in health care. Thus, in terms of topic coverage and specific aspects of NFC, they are concerned about similar research issues though their coverage varies. To make use of an extent to which the subject is examined, we look at a degree to which in-depth articulation and re-contextualization of underlying theories or accounts. Nevertheless, our shortlist also gives information about title, author, source, domain and key research issues of the papers. 25 NFC related papers were reviewed from the design science point of view.

In accordance with Design Science Research Guidelines [92], two researchers conducted separate evaluations of these papers to see any discrepancy with their evaluations. The papers in question were examined and evaluated again to ensure more objective, systematic and rigor assessments. Meanwhile, with the collection of NFC related papers, two researchers started to work on the taxonomy of NFC research and categorization of each study. The research strategy followed for this study was an iterative process, backward strategy (see figure 1) while working on the classification of the NFC literature. We tried to find and add new studies about NFC to our review and design science shortlist. In doing so, we are able to provide academicians and practitioners with a comprehensive base for better understanding of NFC research.

The distribution of the papers by their publication year is presented in figure 3. As shown in figure 3, research on NFC as a promising design science research area grew significantly in recent years, especially after 2008.

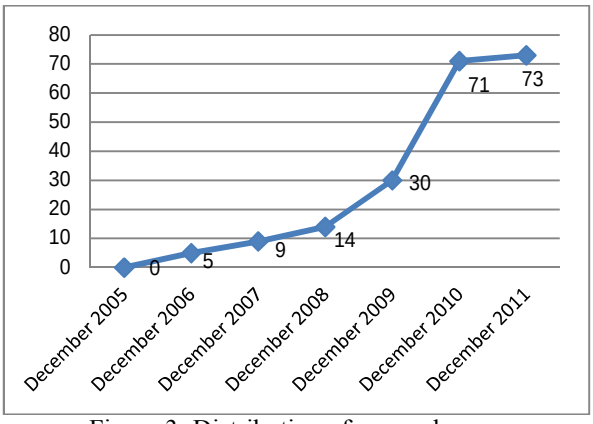


Figure 3: Distribution of papers by year

We should also note that the research methodology that is employed for this academic literature review has some limitations. The first limitation is about the limited number of journal papers found for the literature review. Due to its characteristics, NFC research results are yet to be mature enough, so this limitation is naturally inevitable. The second possible limitation is that the evaluation of 25 research papers through design science guidelines was done by human-reasoning with articulations. This is also due the fact that the adopted evaluations criteria are aimed to facilitate our examination without quantitative measures even two researcher did separate evaluations and compare their results with a number of review cycles.

3 Framework for Research on NFC

The proposed framework is based on a concept-centric literature review [84]. Concepts are consolidated in terms of subject categories. We identify four major categories (see figure 4) and bidirectional relationships between categories.

These are NFC Theory and Development, NFC Infrastructure, NFC Applications and Services and NFC Ecosystem. In the following, we shall describe them and their sub-categories with corresponding studies.

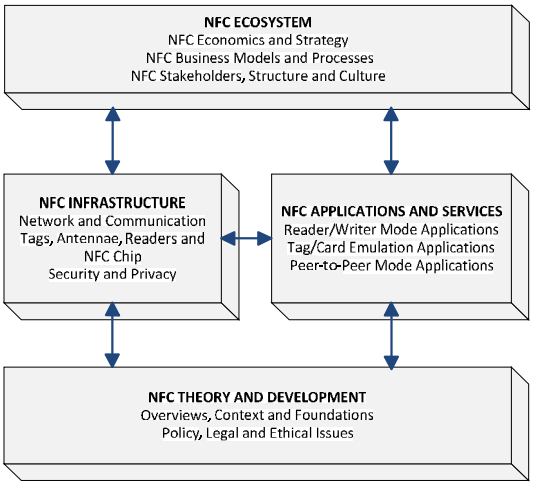


Figure 4: Classification Framework for NFC Research

3.1 NFC Theory and Development

This is the fundamental level of the proposed NFC research framework. It includes the studies related with the development of NFC technology and applications. We examine this level along with two aspects. The first one is "Overviews, Context and Foundations" which includes general introductions, assessments, reviews about NFC, foundations or standards on NFC technology, performance analysis and measurements and new guidelines for the development of NFC enabled applications or services. The second one is "Policy, Legal and Ethical issues" such as security and privacy issues, regulations, and legal requirements. These papers generally focus on more behavioural issues and behavioural sciences which seek to develop and justify theories, rather than developing a design artifact. It is true that these theories underpin and are affected by design decisions [92]. NFC Development papers dealing with this level influences upper levels that focus on design science in NFC research.

3.2 NFC Infrastructure

In fact, this intermediate level is introduced as NFC technology which is examined in terms of three major aspects; "Network and Communication" issues (e.g. data aspect, new communication protocols, OTA transactions), hardware issues dealing with "Tags, Antennae, Reader and NFC Chip", "Security and Privacy" issues (e.g. vulnerability analysis, availability, confidentiality, integrity, authentication, authorization, non-repudiation) that focus on developing design artifact rather than behavioural issue. This layer is positioned with pre-defined business related with to existing technology infrastructure, applications and existing ecosystem. That is, the proposed framework shows the direct linkages of "NFC Infrastructure" with other categories. Moreover, NFC infrastructure related research facilitates new business needs due to the search process nature of NFC.

3.3 NFC Applications and Services

Another middle level of NFC framework as NFC enabled Applications and Services. This is influenced from other three categories and provides a problem space or new business needs. NFC technology covers a wide range of applications and these applications provides real implementations or prototypes with rigor design artifact evaluations such as experimental, testing or field studies etc. We investigate NFC applications from the standpoint of NFC operating modes. "Reader/Writer Mode Applications" provides NFC devices to read and modify data stored in NFC compliant passive (without battery) transponders, "Card Emulation Mode Applications" provides NFC devices to behave like a standard smartcard (e.g. payment and ticketing applications), "Peer-To-Peer Mode Applications" enables two NFC devices to establish a device to device link-level communication to exchange contacts or any other kind of data [81]. Indeed, design artifacts which propose composed applications or services operating in two or more modes can be seen in NFC literature.

3.4 NFC Ecosystem

NFC Ecosystem as the highest level of the NFC Research Framework can be also referred as a part of the problem space or environment of NFC research, the improvements or changes in middle and fundamental layers affect NFC Ecosystem significantly. We examined NFC ecosystem in three major categories. "NFC Economics and Strategy" and "NFC Business Models and Processes" are about business requirements, analysis and managerial sides of the NFC technology. Third aspect is the "NFC Stakeholders, Structure and Culture" which deals with more social sides of NFC technology such as roles, characteristics and capabilities (e.g. user acceptance, usability, adoption, reliability, manageability) of stakeholders (e.g. Mobile network operators, service providers, end users), cultural context of NFC enabled services. Stakeholders play a crucial role in facilitating the NFC research and development. In accordance with [2], in a NFC ecosystem, there are the goals, tasks, problems, and opportunities that define business needs as they are perceived by the stakeholders. These perceptions are shaped by the roles and capabilities. The characteristics of stakeholders are evaluated within the context of economics and strategies, structure and culture, business models and processes.

4 Framework for Research on NFC

4.1 Findings from the literature

A total of 74 studies were classified with respect to our proposed framework. These articles were analyzed by year of publication and by topic area. At the same

time, 25 design science research papers which are selected from these 74 papers were evaluated through design science guidelines. These two particular analyses will provide us promising guidelines for pursing rigorous and business relevant research on NFC and its applications, services.

A majority of NFC research papers (186 out of 202 or %92 of the total) were published in conferences or symposiums, even though in the last two years more journal publications are available. This shows that there is a clear need for more rigorous NFC research articles to be published in journals. Once the progress of NFC research is reached to more established results, academics and practitioners may benefit from this mature Body-of-Knowledge.

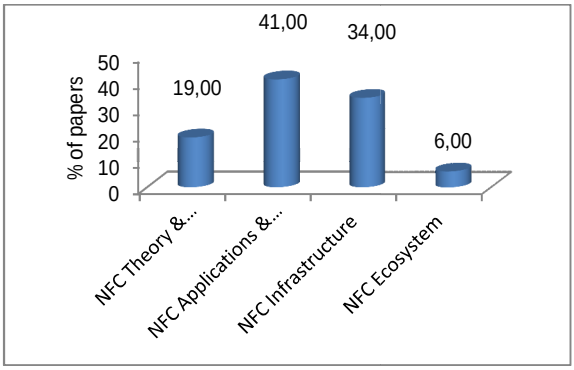


Figure 5: Distribution of Papers by Categories

The distribution of NFC research papers by subject is shown in figure 5. A majority of the NFC research is related to NFC Applications and NFC application development, while a few of them were on "NFC Ecosystem", covering only 12 published papers out of 202.

Table 3 indicates the status of existing Body of Knowledge with respect to the proposed framework. As mentioned before, the majority of NFC research as "NFC Applications & Services" (41%) is examined in a standpoint of operating modes of NFC, in three broad topics. More than half of the academic papers in this category deal with applications and services of NFC that is operating in reader/writer mode (41 academic papers). At the same time, the academic literature related with "Reader/Writer Mode Applications" is the largest proportion (20 %) of the NFC literature (e.g. retailing, health, education, supply chain management, museums, social networking, shopping, electronic voting, multimedia controller, smart posters etc.).

The second largest topic is "Card Emulation Mode Applications" (e.g. payment, mobile coupons, ticketing, electronic key) with 20 academic papers out of total. The fewest number of papers were on the "Peer-to-Peer Mode Applications".

The second largest category of NFC literature is related to "NFC Infrastructure" (34%) which provides "Tags, Antennas, Readers and NFC Chip" issues made up the largest topic (38%) within this category. The other topics discussed were "Security" (26%) and "Network and Communication" (19%). In fact, within this category distribution of NFC Infrastructure literature among topics is quite proportional.

The third category as "NFC Theory and Development" is examined in two broad topics. "NFC Overviews, Context and Foundations" with 27 related academic papers is the large proportion of this category. The other topic on theory and development discussed in NFC literature is "NFC Policy, Ethical and Legal Issues" (11 academic papers). These findings reflects the fact

Classification Criteria	# of Papers	Some References	% by subject	% by all subject
NFC Theory and Development				
NFC Overview, Context and Foundations	28	[11, 16, 19, 22, 43, 47, 49, 54, 64, 67, 72,74]	69	13
NFC Policy, Ethical and Legal Issues	11	[8, 9, 40, 91]	31	6
Total	**39**		**100**	**19**
NFC Applications and Services				
Reader / Writer Mode Applications	41	[15, 17, 21, 25, 30, 31, 32, 33, 35, 48, 50, 51, 58, 61, 66, 73, 75, 77, 78]	60	20
Tag Emulation Mode Applications	20	[14, 20, 34, 45, 55, 57, 59, 60, 62, 79]	30	10
Peer-to-Peer Mode Applications	7	[65]	10	3
Total	**68**		**100**	**33**
NFC Infrastructure				
Network and Communication	19	[27, 28, 36, 39, 44, 46, 70]	24	9
Tags, Antennas, Readers and NFC Chip	38	[7, 24, 41, 52, 53, 69, 71]	45	19
Security and Privacy	26	[12, 18, 23, 38, 26, 42, 68, 76]	31	13
Total	**83**		**100**	**41**
NFC Ecosystem				
NFC Economics and Strategy	1	[90]	0.09	0.4
NFC Business Models and Processes	5	[6, 13, 37, 89]	4.1	2
NFC Stakeholders, Structure and Culture	6	[10, 29, 57, 63]	50	3
Total	**12**		**100**	**5.4**

Table 3: Classification of the reviewed NFC literature.

that NFC is relatively a new, promising research area, so that there is a clear need for more academic study on regulations, privacy, and legal issues surrounding NFC to sustain its development.

As seen in Table 3, there were relatively fewer academic research papers on "NFC Ecosystem" (5,4% out of the total). This category is examined in three broad topics, unfortunately there were not any "specific" academic paper dealing with NFC Economics and Strategy for NFC technology's development, improvement. There were research papers mostly that are surrounding "NFC Business Models and Processes" (5 research papers out of 202) and "NFC Stakeholders, Structure and Culture" (6 research papers out of 202).

In fact, most of NFC related papers contribute to new ideas, such as on security, hardware or business models while proposing a new, unique NFC enabled application or a new Communication Protocol. In such situations, we tried to discover the paper's main contribution, focus point, and made the appropriate classification scheme. Table 3 gives a summary of all of the reviewed academic papers clearly according the proposed classification scheme. This table should be beneficial and helpful resource for anyone who is searching for NFC related papers on a specific area. Meanwhile, Table 3 includes a representative study for each sub-category of the NFC Framework except for the category of "NFC Economics and Strategy", which is not present in the literature yet.

Based on the descriptive findings above, we shall induce some insights in the following:

- It is not surprising that most of the academic research papers were related to "NFC Applications and Services", especially operating in reader/writer mode. The reason of this model is that development and implementation of such services or applications are viable than developing applications operating in other modes. Unfortunately we did not find many rigorous research papers on "Peer-to-Peer Mode Applications".

- The second largest proportion of the papers is related with the "NFC Infrastructure". Our review shows the importance of focusing on technical issues of a new technology again, rather than issues related to realizing economics, business values or strategies for NFC development, dissemination and marketing. As seen in Table 1, literature dealing with technical issues on NFC is useful for anyone who is studying on "NFC Infrastructure". We expect more specific research to be conducted on business issues, economics of NFC technology.
- While developing new NFC enabled applications or services, ecosystem of NFC technology clearly needs to be considered. Such new applications or services can bring new business models, processes with new players. Especially the capabilities, characteristics and roles of stakeholders need to be evaluated and modified when necessary, in order to satisfy the requirements of new business models and processes. Cultural differences on adopting NFC enabled technologies could be an interesting area for investigation.
- In terms of theory and development, most of the research papers those are published in journals were overviews and assessments on NFC technology rather than proposing a new design artifact. The articles in journals that we found are not sufficient for development of NFC literature. We expect more rigorous design science research on NFC to be published in journals. Policy, ethical and legal problems which can be referred as societal and behavioral issues were another important and demanding research areas for development of a new, emerging technology. However, it is hard to find papers dealing with the public policy or legal problems (e.g. taxation problems, trust, fraud, privacy issues for internet privacy, financial privacy). [91] provide a review of the regulations and policies governing NFC in Europe and Asia and related incentives. We agree with [91] that "for NFC to thrive, privacy must be considered in the design of the technology, the platforms, and the services". Indeed, this should prompt academic researchers to adopt design science research paradigm to investigate this area.

4.2 Findings from the DSR Perspective

Based on the aforementioned design criteria, Table 4 shows the evaluations of representative papers for each NFC research category. For the rest of the short-listed papers, the complete evaluations can be found in the Appendix.

The findings from the design science guideline evaluations show that most of the NFC design science papers propose an artifact which provides an utility for a specific and relevant business problem. These two requirements for a design science research are sufficiently considered and explained in the research papers. Needlessly to say that explicitly emphasized business problems will be more beneficial and useful for interested researchers and practitioners.

As mentioned before utility and efficiency of the proposed artifact must be demonstrated with appropriate methods. Design evaluation guideline needs to be highly considered while performing NFC academic research. Most of the papers (of 25 research papers) use more descriptive (e.g. scenarios, use cases to demonstrate its utility) or analytical (e.g. architecture analysis) methods while developing an applications or service, rather than performing experimental or testing methods. Design evaluations are performed in most papers through scenarios or use cases, instead controlled experiments or simulations will be more useful for representing the proposed artifact rigorously.

As seen in our review, nearly all of the NFC research papers provide research contributions explicitly or implicitly, due to their nature. For instance, an NFC design science paper [14] provides varying contributions in terms of security, network and communication while proposing a new NFC enabled service.

		Paper	Guideline 1: Design as an Artifact	Guideline 2: Problem Relevance	Guideline 3: Design Evaluation	Guideline 4: Research Contributions	Guideline 5: Research Rigor	Guideline 6: Design as a Search Process	Guideline 7: Communication of Research
NFC Theory and Development	NFC Overviews, Context and Foundations	[74]	An NFC test system architecture	Clearly explained in requirements section	Evaluations through analytical, experimental	Contributes due to its nature	Rigorous; applicable and generalizability	Explicitly design search	Communicates all audiences
	Policy, Ethical and Legal Issues	[8]	Context-based adaptation system	Mentioned; to reduce the distraction caused by mobile phones	Evaluated; analytical and descriptional, cases	Clearly contributes due to its nature	Rigorous; prototype implementation in office environments	Explicitly design search	Communicates mostly technical audiences
NFC Applications & Services	Reader/Writer Mode Applications	[75]	Maintenance systems with NFC	Mentioned; to improve recurring maintenance processes	not mentioned; only implications of the system	Contributes due to its nature; design artifact	Somewhat rigorous work, implementation	Somewhat search process	Communicates all audiences
	Tag Emulation Mode Applications	[79]	Apps for University environment	Mentioned the requirements but not so much satisfactory	Evaluation is done through descriptional -scenarios	Contributes due to its nature	Somewhat rigorous; applicable	Not a complete search process	Communicates technical audiences
	Peer-to-Peer Mode Applications	[65]	Hot in City application	Implicitly defined the business requirements	Descriptional and architectural, implications of the system	Contributes implicitly	Somewhat rigorous work, real implementation	Somewhat search process	Communicates mostly technical audiences
NFC Infrastructure	Tags, Antennae, Readers and NFC Chip	[7]	Guidelines for estimation of the capacity performance	Not a specific problem; only analyzed the capacity performance of the inductive coupling NFC system	Good analysis, evaluation based on theoretical background	Contributes due to its nature	Rigorous work	Not a complete search process	Communicates technical audiences
	Network and Communication	[27]	Verify Protocol	Explicitly, well defined requirements	Analysis of protocols, analytical and descriptive methods mostly	Clear Contributions	Rigorous, real implementation, applicable, performance evaluations	Highly search process	Communicates mostly technical audiences
	Security and Privacy	[25]	UICC and payment applications	Mentioned explicitly	Not explicitly mentioned, experiments evaluations	Clear contributions	Somewhat rigorous, a real project's intermediate results actually	Somewhat search process	Communicates all audiences
NFC Ecosystem	NFC Business Models and Processes	[13]	Platform management model for NFC ecosystem	Mentioned properly	Descriptively analyzed the model	Contributes actually	Somewhat rigorous, like a proposal	Not a complete search process	Communicates all audiences
	NFC Stakeholders, Structure and Culture	[50]	Approaches for adaptability of RFID-NFC	Clearly specified	Evaluated; architectural and descriptional	Somewhat contributes	Not clear	Not Clear	Communicates all audiences

Table 4: Design Science Guideline Evaluations of Representative NFC Studies

In regard to the research rigor perspective which is concerned with the construction and evaluation of design artifact, the design artifact's applicability and generalizability should be addressed. There is a clear need for rigorous NFC research papers, where design rationale and cycle should be explicit. This is needed to achieve an effective communication for exploiting research results in the appropriate communities (engineering- or management-oriented audiences).

5 Research Agenda for NFC

NFC as an emerging research area has attracted the attention of both practitioners and academics. As cited before, academic research activities on NFC area have increased significantly after the year 2006. We believe that, this study is the first academic literature review on NFC technology. With this literature review, we want to shed light on the current status of NFC research. This review identified 109 academic papers composed of studies from 2006 to 2012. The results from NFC classification scheme and from design science guideline evaluations have several important implications.

It is true that NFC technology has become a promising, challenging research area in recent years. There is a clear need for more journal publications to provide business related and rigorous research papers on NFC technology.

Among all these possible questions, we expect that calls for the following three subjects may draw considerable attention from academics and practitioners as well:

NFC Ecosystem and Business Models. The notion of ecosystem appears to be granted in both in NFC World (both academics and practitioners' point of view). Business requirements and ecosystem rational are hardly taken into account in the proposed models, which questions how, if possible at all, comparative are these proposed models? Whether commonalities and differences on the model element at the foundation level or not? Nevertheless, what needed is an explicit interrogation of what constituents (primitives) the very notion of ecosystem in the NFC context. The challenge for NFC stakeholders today is to promote and combine creativity in order to bring substantial improvements in terms of economic and social aspects. One needs to address some challenges for establishing a successful ecosystem. In this regard, possible follow-up would be: How are you going to get all those potential participants to believe that they can work together effectively and creatively? What will be the key roles? How will you let each group innovate relatively freely, but ensure that as the project proceeds all of the contributions will come together?

NFC Secure Element Analysis. NFC enabled services must assure users and service providers that the transaction takes place in a protected environment. This protection is achieved by use of a secure element (SE), which can be referred as the components in the device providing the security required to support various business models. The SE is concerned with technical issues (combination of hardware, software, interfaces and protocols) and management issues as well. Furthermore, there are various architectural options for a SE depending on its implementation options such as Embedded Hardware as non-removable SE, Stickers, Secure Micro SD cards and UICCs as removable SEs, Trusted Mobile Base as a combination of software programs on dedicated hardware. Several questions can be raised to address NFC SE issues such as: how to manage SE for

Framework Element	Some Research Opportunities
NFC Ecosystem - NFC Economics and Strategy - Business Models and Processes - Stakeholders, Structure and Culture	- Proposing an underlying value typology for NFC applications - Evaluating the impacts of NFC on business process and value-added activities - Determining generic stakeholders and meta-model describing interactions among them in an NFC ecosystem - Empirically testing ecosystem models and a comparative analysis in various industry and country settings, comparative study - Evaluating cultural factors on adopting NFC enabled applications and services - Macro and micro economic analysis of developing a specific NFC enabled application - Determining the effects of NFC use at multiple levels, including overall business, operations, individual
NFC Infrastructure - Network and Communication - Tags, Antennae, Readers and Chips - Security	- Evaluating existing NFC enabled device internal hardware, network, and communication standards, and their implementation - Proposing new architectures/standards or extensions whenever required - Evaluating proposed hardware, network, and communication models and standards for NFC - Designing and modifying security architectures /standards - Evaluating proposed security models for NFC - Examining Compatibility Issues with NFC-enabled devices and solutions - Testing performance, processing, data storage and data communication NFC applications with different infrastructures
NFC Applications, Architecture and Services - Reader/Writer Mode - Tag/Card Emulation Mode - Peer-to-Peer Mode	- Developing new applications, architecture and services - Evaluating the proposed applications and services - Identifying novel applications for each NFC mode - Evaluation of proposed artifacts in terms of their benefits, contribution analysis - Proposing NFC artifacts in the form of applications, model, and instantiation - Secure Element Alternatives
NFC Theory and Development - Context and Foundations - Policy, Legal, Privacy and Ethical issues	- Developing meta-elements for NFC policies, regulations and legal standards at the individual and organization level - Adopting appropriate accounts to identify legal, privacy, and ethical issues concerning NFC use - Empirically testing NFC adoption in different contexts (various users profiles, application characteristics) - Providing useful methods, models, guidelines for developing NFC enabled applications

Table 5: Research Agenda with respect to potential research questions.

concurrent applications? What criteria should be taken into account to assess possible SE implementation option? Who share what data in the SE for privacy, loyalty service provisioning?

User Perception on and Privacy Issues with NFC. Industry reports have been published to indicate countries adoption situation in present and upcoming years. It seems that in compare to similar technologies (e.g., RFID), the adoption lifecycle for countries is to be shorter. But, [88] states "The most surprising result of the survey was the respondents' low expectations in regards to customer acceptance". This is in clear contrast to the reports on NFC trials which generally describe participants as enthusiastic about the technology". Thus, there is still an open question concerning Is the customer ready for NFC use? Surely, various factors including appropriate ecosystem, market fragmentation, and service availability are essential for successful NFC roll out. In literature, adoption factors in IT in general and mobile technology in particular are studied [88]. One needs to investigate if and how such factors affect intention to use NFC services. Industry organizations will benefit from those studies using empirical setting to assess user behaviors on NFC use. Case studies, including [91], bring up important issues with control, consent and accountability related to NFC.

6 Conclusion

As stated in [92], it is important for behavioral and design science researchers to understand new emerging technologies such as RFID, NFC. With recent endeavors of practitioners and academics concerning the use of Near Field Communication (NFC), one can expect a bright future of NFC along with business opportunities. But, several challenges remain ahead for enhancement of Body-of-Knowledge for NFC. This study goes beyond a typical literature in that it employs Design Science Perspective to articulate BoK for NFC, examines its progress and proposes a research agenda with promising research areas and questions.

Noticeably, with the development of more and innovative NFC enabled applications, the need for standards and policies is increased. At the same time, strategy for diffusion and adoption of NFC systems and economy of NFC systems need to be considered while developing new services, which includes the costs of designing, developing, controlling and updating such systems.

The framework proposed is found to be useful to organize a number of existing studies (i.e., 202 papers in the last five years), we expect that more sub-topics should be added and updated in the framework. Since most of the studies in the BoK focus on artifact development and its instantiation, the design science research perspective serves an appropriate ground for assessing its progress. Accordingly, the three cycles of DSR have not been equally realized in the present BoK. There should be a call for those studies paying attention on especially rigor and design cycles. More specifically, as [2] stated, NFC studies ought to consider a research rationale in terms of what design processes (search heuristics) will be used to build the artifact? How are the artifact and the design processes grounded by the knowledge base? What, if any, theories support the artifact design and the design process? Furthermore, we expect more studies where design evaluation is to be explicit by using observational, experimental techniques.

Based on the organizing framework, we put forward a list of research opportunities. One can see that every framework element has a potential to investigate its research topics further. As the review shows that application, architecture element has been a center of attention so far, we expect that more emphasis would be given on NFC ecosystem, underlying theory and its adaption. In this regard, we suggest that secure element analysis, NFS ecosystem and business model, user perceptions on NFC are worth to invest as specific research areas from the academics and practitioners points of views.

Acknowledgement

The authors are grateful to the Associate Editor Maria Ganzha and the reviewer's valuable comments that improved the manuscript. The insights provided by Vedat Coskun and Kerem Ok on the earlier versions of the manuscript were fruitful to establish the basis of this research.

References

[1] E.W.T. Ngai, K. K. L. Moon, F. J. Riggins, C. Y. Yi, "RFID research: An academic literature review (1995–2005) and future research directions", International Journal of Production Economics 112, pp. 510–520, 2008.

[2] E.W.T. Ngai, A. Gunasekaran, "A review for mobile commerce research and applications", Decision Support Systems 43, pp. 3 – 15, 2007.

[3] E.W.T. Ngai, F.K.T. Wat, "A literature review and classification of electronic commerce research", Information & Management 39, pp. 415–429, 2002.

[4] Y. Wang, J. Li, P. Liu, F. Yang, "Electronic Commerce Research Review: Classification and Analysis", 2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCom.

[5] A. Urbaczewski, L. M. Jessup, B. Wheeler, "Electronic Commerce Research: A Taxonomy and Synthesis", Journal of Organizational Computing and Electronic Commerce, 12: 4, pp. 263 — 305, 2002.

[6] B. Benyó, "Business Process Analysis of NFC-based Services", in IEEE 7th International Conference on Computational Cybernetics, Palma de Mallorca, Spain, 2009.

[7] H. C. Jing, Y. E. Wang, "Capacity Performance of an Inductively Coupled Near Field Communication System", in Antennas and Propagation Society International Symposium, San Diego, CA, 2008

[8] S. Krishnamurthy, D. Chakraborty, S. Jindal, S. Mittal, “Context-Based Adaptation of Mobile Phones Using Near-Field Communication” in Third Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, San Jose, CA, 2006.

[9] P. Schoo, M. Paolucci, “Do you talk to each poster? Security and Privacy for Interactions with Web Service by means of Contact Free Tag Readings”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[10] J. Bravo, R. Hervás, G. Chavira, S. W. Nava, V. Villarreal, “From Implicit to Touching Interaction: RFID and NFC Approaches”, in Conference on Human System Interactions, Krakow, 2008

[11] Sixto Ortiz Jr., “Is Near-Field Communication Close to Success?”, Computer, Volume 39, Number 3, pp. 18-20, Mar. 2006.

[12] G. Madlmayr, O. Dillinger, J. Langer, J. Scharinger, “Management of Multiple Cards in NFC-Devices”, in Proceedings of the 8th IFIP WG 8.8/11.2 International Conference on Smart Card Research and Advanced Applications, London, UK, 2008.

[13] G. Madlmayr, J. Langer, J. Scharinger, “Managing an NFC Ecosystem”, in 7th International Conference on Mobile Business, Barcelona, 2008

[14] S. Dominikus, M. Aigner, "mCoupons: An Application for Near Field Communication (NFC)," in 21st International Conference on Advanced Information Networking and Applications Workshops, Niagara Falls, Ontario, Canada, 2007.

[15] S. Karpischek, F. Michahelles, F. Resatsch, E. Fleisch, "Mobile Sales Assistant - An NFC-Based Product Information System for Retailers," in First International Workshop on Near Field Communication, Hagenberg, 2009.

[16] J. Ondrus, Y. Pigneur, “Near Field Communication: An Assessment for Future Payment Systems”, Information Systems and E-Business Management, Volume 7, Number 3, pp. 347-361, June 2009.

[17] J. Morak, D. Hayn, P. Kastner, M. Drobics, G. Schreier, “Near Field Communication Technology As The Key For Data Acquisition In Clinical Research”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[18] K. S. Kadambi, J. Li, A. H. Karp, “Near-Field Communication-Based Secure Mobile Payment Service”, in International Conference on E-commerce, Taipei, Taiwan, 2009.

[19] J. Fischer, “NFC in Cell Phones: The New Paradigm For An Interactive World”, IEEE Communications Magazine, Volume 47, Issue 6, pp. 22-28, June 2009.

[20] S. L. Ghìron, S. Sposato, C. M. Medaglia, A. Moroni, "NFC Ticketing: A Prototype and Usability Test of an NFC-Based Virtual Ticketing Application”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[21] A. Fressancourt, C. Hérault, E. Ptak, "NFCSocial: Social Networking in Mobility through IMS and NFC”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[22] F. Michahelles, F. Thiesse, A. Schmidt, J. R. Williams, "Pervasive RFID and Near Field Communication Technology," IEEE Pervasive Computing, Volume 6, Number 3, pp. 94-96, c3, July-Sept. 2007.

[23] V. Alimi, M. Pasquet, "Post-Distribution Provisioning and Personalization of a Payment Application on a UICC-Based Secure Element," in International Conference on Availability, Reliability and Security, Fukuoka, Japan, 2009.

[24] R. G. Mair, “Protocol-Independent Detection of Passive Transponders for Near-Field Communication Systems”, IEEE Transactions on Instrumentation and Measurement, Volume 59, Number 4, April 2010.

[25] N. Kefalakis, N. Leontiadis, J. Soldatos, K. Gama, D. Donsez, “Supply Chain Management and NFC Picking Demonstrations using the AspireRfid Middleware Platform”, in Proceedings of the ACM/IFIP/USENIX Middleware Conference Companion, Leuven, Belgium, 2008.

[26] G. Kálmán, J. Noll, “SIM as Secure Key Storage in Communication Networks”, in Proceedings of the Third International Conference on Wireless and Mobile Communications, Guadeloupe, 2007.

[27] J. Woo, A. Bhargav-Spantzel, A. C. Squicciarini, E. Bertino, "Verification of Receipts from M-commerce Transactions on NFC Cellular Phones," in 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008.

[28] S. Grunberger, J. Langer, “Analysis and test results of tunneling IP over NFCIP-1”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[29] Y. Anokwa, G. Borriello, T. Pering, R. Want, “A User Interaction Model for NFC Enabled Applications”, in Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops, White Plains, New York, USA, 2007.

[30] I. Sánchez, M. Cortés , J. Riekki , “Controlling Multimedia Players using NFC Enabled Mobile Phones”, in Proceedings of the 6th International Conference on Mobile and Ubiquitous Multimedia, Oulu, Finland, 2007.

[31] I. Cappiello, S. Puglia, A. Vitaletti, "Design and Initial Evaluation of a Ubiquitous Touch-Based Remote Grocery Shopping Process”, in First International Workshop on Near Field Communication, Hagenberg, 2009.

[32] J. Morak, V. Schwetz, D. Hayn, F. Fruhwald, G. Schreier, “Electronic Data Capture Platform for Clinical Research based on Mobile Phones and Near Field Communication Technology”, in 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, 2008.

[33] R. Hardy, E. Rukzio, M. Wagner, M. Paolucci, "Exploring Expressive NFC-based Mobile Phone Interaction with Large Dynamic Displays", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[34] G. M. Miraz, I. L. Ruiz, M. Á. Gómez-Nieto, "How NFC can be used for the Compliance of European Higher Education Area Guidelines in European Universities", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[35] E. Siira, T. Tuikka, V. Tormanen, "Location-based Mobile Wiki using NFC Tag Infrastructure", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[36] C.Y. Leong, K. C. Ong, K. K. Tan, O.P. Gan, "Near Field Communication and Bluetooth Bridge System for Mobile Commerce", in IEEE International Conference on Industrial Informatics, Singapore, 2006.

[37] B. Benyó, Member, IEEE, A. Vilmos, K. Kovacs, L. Kutor, "NFC Applications and Business Model of the Ecosystem", in 16th IST Mobile and Wireless Communications Summit, Budapest, 2007.

[38] G. Madlmayr, J. Langer, C. Kantner, J. Scharinger, "NFC Devices: Security and Privacy", in The Third International Conference on Availability, Reliability and Security, Barcelona, 2008.

[39] J. Ylinen, M. Koskela, L. Iso-Anttila, P. Loula, "Near Field Communication Network Services", in Third International Conference on Digital Society, Cancun, 2009.

[40] L. Francis, G. Hancke, K. Mayes, K. Markantonakis, "Potential Misuse of NFC Enabled Mobile Phones with Embedded Security Elements as Contactless Attack Platforms", in International Conference for Internet Technology and Secured Transactions, London, 2009.

[41] H. Mika, H. Mikko, Y. Arto, "Practical Implementations of Passive And Semi-Passive NFC Enabled Sensors", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[42] M. Reveilhac, M. Pasquet, "Promising Secure Element Alternatives for NFC Technology", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[43] X. Yu-ning, "Research on NFC and SIMpass Based Application", in International Conference on Management and Service Science, Wuhan, 2009.

[44] G. Madlmayr, J. Langer, C. Kantner, J. Scharinger, I. Schaumüller-Bichl , "Risk Analysis of Over-the-Air Transactions in an NFC Ecosystem", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[45] M. Pasquet, J. Reynaud, C. Rosenberger, "Secure Payment With NFC Mobile Phones In The Smart Touch Project", in International Symposium on Collaborative Technologies and Systems, Irvine, CA, 2008.

[46] G. Yang, Z. Huang, L. Wan, "The Development of RFID Module in NFC Phone", in 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, Hong Kong, 2009.

[47] B. Benyó, A. Vilmos, G. Fördős, B. Sódor, L. Kovács , "The StoLPan View of the NFC Ecosystem", in Proceedings of the Conference on Wireless Telecommunications Symposium, Prague, 2009.

[48] A. Marcus , G. Davidzony, D. Law , N. Verma , R. Fletcher , A. Khanz, L. Sarmenta, "Using NFC-enabled Mobile Phones for Public Health in Developing Countries", in First International Workshop on Near Field Communication, Hagenberg, 2009.

[49] V. Kostakos, E. O'Neill, "NFC on Mobile Phones: Issues, Lessons and Future Research", in Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops, White Plains, New York, USA, 2007.

[50] J. Bravo, R. Hervás, R. Gallego, G. Casero, M. Vergara, T. Carmona, C. Fuentes, S.W. Nava, G. Chavira, V. Villarreal, "Enabling NFC Technology to Support Activities in an Alzheimer's Day Center", in Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments, Athens, Greece, 2008.

[51] G. Chavira, S. W. Nava, R. Hervás, V. Villarreal, J. Bravo, S. Martín, M. Castro, "Services through NFC technology in AmI Environment", in Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, Linz, Austria, 2008.

[52] M. Roland, H. Witschnig, E. Merlin, C. Saminger, "Automatic Impedance Matching For 13.56 Mhz NFC Antennas", in The 6th International Symposium on Communication Systems, Networks and Digital Signal Processing, Graz, 2008.

[53] Y. L. Sylvester, D. Blaauw, "Near-Field Communication using Phase-Locking and Pulse Signaling for Millimeter-Scale Systems", in Proceedings of The IEEE Custom Integrated Circuits Conference, San Jose, CA, 2009.

[54] D. Remedios, L. Sousa, M. Barata, L. Osorio, "NFC Technologies in Mobile Phones and Emerging Applications", in IFIP International Federation for Information Processing, Volume 220, Information Technology for Balanced Manufactiuing Systems, ed. Shen, W., (Boston: Springer), pp. 425-434, 2006.

[55] Y. Chang, C. Chang, Y. Hung, C. Tsai, "NCASH: NFC Phone-Enabled Personalized Context Awareness Smart-Home Environment", Cybernetics and Systems, Volume 41, Issue 2, pp. 123 – 145, February 2010.

[56] M. Isomursu, "Tags and The City", PsychNology Journal, Volume 6, Number 2, pp. 131-156, 2008

[57] J. Neefs, F. Schrooyen, J. Doggen, K. Renckens, "Paper Ticketing vs. Electronic Ticketing Based on Off-Line System 'Tapango'", in Second

International Workshop on Near Field Communication, Monaco, 2010.
[58] P. C. Garrido, G. M. Miraz, I. L. Ruiz, M. Á. Gómez-Nieto, "A Model for the Development of NFC Context-Awareness Applications on Internet of Things", in Second International Workshop on Near Field Communication, Monaco, 2010.
[59] R. Steffen, J. Preißinger, T. Schöllermann, A. Müller, I. Schnabel, "Near Field Communication (NFC) in an Automotive Environment", in Second International Workshop on Near Field Communication, Monaco, 2010/
[60] H. Aziza, "NFC Technology in Mobile Phone Next-Generation Services", in Second International Workshop on Near Field Communication, Monaco, 2010.
[61] M. Vergara, P. Díaz-Hellín, J. Fontecha, R. Hervás, C. Sánchez-Barba, C. Fuentes, J. Bravo, "Mobile Prescription: An NFC-Based Proposal for AAL", in Second International Workshop on Near Field Communication, Monaco, 2010.
[62] Z. Lou, "NFC Enabled Smart Postal System", in Second International Workshop on Near Field Communication, Monaco, 2010.
[63] H. Franssila, "User Experiences and Acceptance Scenarios of NFC Applications in Security Service Field Work", in Second International Workshop on Near Field Communication, Monaco, 2010.
[64] B. Benyó, B. Sódor, G. Fördos, L. Kovács, A. Vilmos, "A Generalized Approach for NFC Application Development", in Second International Workshop on Near Field Communication, Monaco, 2010.
[65] E. Siira, V. Törmänen, "The Impact of NFC on Multimodal Social Media Application", in Second International Workshop on Near Field Communication, Monaco, 2010.
[66] F. Köbler, P. Koene, H. Krcmar, M. Altmann, J. M. Leimeister, "LocaTag - An NFC-Based System Enhancing Instant Messaging Tools with Real-Time User Location", in Second International Workshop on Near Field Communication, Monaco, 2010.
[67] S. Cecil, G. Schmid, K. Lamedschwandner, J. Morak, G. Schreier, A. Oberleitner, M. Bammer, "Numerical Assessment of Specific Absorption Rate in the Human Body Caused by NFC Devices", in Second International Workshop on Near Field Communication, Monaco, 2010.
[68] M. Roland, J. Langer, "Digital Signature Records for the NFC Data Exchange Format", in Second International Workshop on Near Field Communication, Monaco, 2010.
[69] M. Gebhart, R. Szoncso, "Optimizing Design of Smaller Antennas for Proximity Transponders", in Second International Workshop on Near Field Communication, Monaco, 2010
[70] W. Chen, G.P. Hancke, K.E. Mayes, Y. Lien, J. H. Chiu, "NFC Mobile Transactions and Authentication Based on GSM Network", in Second International Workshop on Near Field Communication, Monaco, 2010.
[71] J. Cho, J. Kim, S.Kim, "An NFC Transceiver with Dual Antenna Structure to Support RF-Powered Transponder Mode", IEICE Transactions on Communications, Volume E92-B No.1 pp. 310-313, 2009.
[72] [73] M. Massoth, T. Bingel, "Performance of Different Mobile Payment Service Concepts Compared With a NFC-Based Solution", in Proceedings of the Fourth International Conference on Internet and Web Applications and Services, Venice/Mestre, Italy, 2009.
[73] J. Morak, A. Kollmann, G. Schreier, "Feasibility and Usability of a Home Monitoring Concept based on Mobile Phones and Near Field Communication (NFC) Technology", in Proceedings of The 12th World Congress On Health (Medical) Informatics, 2007.
[74] J. Langer, C. Saminger, S. Grünberger, "A Comprehensive Concept and System For Measurement and Testing NFC Devices", in EUROCON, St.-Petersburg, 2009.
[75] S. Karpischek, F. Michahelles, A. Bereuter, E. Fleisch, "A Maintenance System Based on Near Field Communication", in Third International Conference on Next Generation Mobile Applications, Services and Technologies, Cardiff, Wales, UK, 2009.
[76] G. Madlmayr, "A Mobile Trusted Computing Architecture for A Near Field Communication Ecosystem", in Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, Linz, Austria, 2008.
[77] F. Kneißl, R. Röttger, U. Sandner, J. M. Leimeister, H. Krcmar, "All-I-Touch as Combination of NFC and Lifestyle", in First International Workshop on Near Field Communication, Hagenberg, 2009.
[78] E. Strömmer, J. Kaartinen, J. Pärkkä, A. Ylisaukko-oja, I. Korhonen, "Application of Near Field Communication for Health Monitoring in Daily Life", in Proceedings of the 28th IEEE Engineering in Medicine and Biology Science Annual International Conference, 2006.
[79] G. M. Miraz, I. L. Ruiz, M. Á. Gómez-Nieto, "University of Things: Applications of Near Field Communication Technology in University Environments", The Journal of E-working, Volume 3, Issue 1, pp. 52-64, 2009.
[80] Raghu Das, "NFC-enabled phones and contactless smartcards 2008–2018", Card Technology Today, Volume 20, Issues 7-8, July-August 2008, pp. 11-13.
[81] Mandl, T., Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance, Informatica 32 pp. 27–38, 2008.
[82] Rosemann, M. and Vessey, I., "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly*, (32: 1), 2008.

[83] Benbasat I and Zmud R , The identity crisis within the IS discipline: defining and communicating the discipline's core properties. MIS Quarterly 27(2), 183–194, 2003.

[84] Webster, J., and Watson, R. T. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," MIS Quarterly (26:2), 2002, pp. xiii-xxiii.

[85] Ok K., Aydin M. N., Coskun V., Ozdenizci B., Current Benefits and Future Directions of NFC Services *IEEE International Conference on Education and Management Technology*, Cairo, Egypt, November 2-4, pp. 334-338, 2010.

[86] Weber, Sven, "Design Science Research: Paradigm or Approach?" (2010). *AMCIS 2010 Proceedings.* Paper 214. http://aisel.aisnet.org/amcis2010/214

[87] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. And Cayirci, E., A survey on sensor networks. IEEE Communications Magazine 40, 8 (August), 102–114, 2002.

[88] Wiechert, T., Thiesse, F., Schaller, A., and Fleisch, E., NFC based Service Innovation in Retail: An explorative Study. In *Proc. ECIS'09*, Verona, Italy, 2009

[89] Kim C., Mirusmonov M., Lee I., An empirical examination of factors influencing the intention to use mobile payment, *Computers in Human Behavior*, Volume 26, Issue 3, May 2010, pp. 310-322.

[90] Ondrus, J. (2011). Mobile Payments Market: Towards Another Clash of the Titans?, Tenth International Conference on Mobile Business, June 2011, Italy

[91] Liebenau, Jonathan and Elaluf-Calderwood, Silvia and Hosein, Gus and Kärrberg, Patrik (2011) Near field communications: privacy, regulation & business models. Retrieved from *http://www2.lse.ac.uk/management/research/initiatives/nokia-near-field-communications-and-privacy-study/home.aspx*

[92] A. R. Hevner, S. T. March, J. Park, S. Ram, "Design Science in Information Systems Research", MIS Quarterly Vol. 28 No. 1, pp. 75-105/March 2004.

APPENDIX

Paper	Guideline 1: Design as an Artifact	Guideline 2: Problem Relevance	Guideline 3: Design Evaluation	Guideline 4: Research Contributions	Guideline 5: Research Rigor	Guideline 6: Design as a Search Process	Guideline 7: Communication of Research
[6]	NFC ecosystem and business analysis	Mentioned; to combine the business process approach with their significant technology developments	Only implications of the ecosystem are mentioned	Clearly contributes due to its nature	Not clear	Not clear	Communicates all audiences
[9]	Security and privacy requirements of an NFC based application	Clearly specified, mentioned	Not explicitly done, case study	Contributes due to its nature	Somewhat rigorous; needs more technical evaluations	Not a complete search process	Communicates mostly technical audiences
[12]	Secure element controller apprach	Clearly mentioned; states the problem	Evaluated through cases, analytical	Clearly contributes due to its nature	Rigorous work	Explicitly design search	Communicates technical audiences
[14]	M-coupons and protocols	Explicitly mentioned the motivation for m-coupon, business needs	Well evaluated through analytical methods	Contributes explicitly	Rigorous but need more performance evaluations	Search Process	Communicates mostly technical audiences
[15]	Mobile Sales Assistant (MSA)	Mentioned	Evaluated the system descriptively	Contributes implicitly	Somewhat rigorous, prototype implementation but not enough	Not a complete search process	Communicates all audiences
[17]	Electronic data capture (EDC) system	Clearly mentioned; to design and develop an additional path for clinical data acquisition	Clear system evaluation, in terms of usability and feasibility, observational	Contributes due to its nature	Somewhat rigorous; needs more technical evaluations	Not a complete search process	Communicates mostly technical audiences
[18]	Secure mobile payment solution	Clearly mentioned need for secure transactions	Evaluation, architectural analysis etc.	Explicitly contributes	Rigorous, real implementation, prototype	Explicitly design search	Communicates all audiences
[20]	NFC-based Virtual Ticketing application	Explicitly mentioned, its ease of use and to its higher security level	Evaluates through architectural and usability analysis (testing), statistical analysis	Contributes explicitly	Rigorous, real implementation, prototype	Not a complete search process	Communicates all audiences

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

☐ Send Informatica free of charge

☐ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than nineteen years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: .....

Title and Profession (optional): .....

.....

Home Address and Telephone (optional): .....

.....

Office Address and Telephone (optional): .....

.....

E-mail Address (optional): .....

Signature and Date: .....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

Ajith Abraham, Siby Abraham, Renato Accornero, Raheel Ahmad, Cutting Alfredo, Hameed Al-Qaheri, Gonzalo Alvarez, Wolfram Amme, Nicolas Anciaux, Rajan Arora, Costin Badica, Zoltán Balogh, Andrea Baruzzo, Borut Batagelj, Norman Beaulieu, Paolo Bellavista, Steven Bishop, Marko Bohanec, Zbigniew Bonikowski, Borko Bosković, Marco Botta, Pavel Brazdil, Johan Brichau, Andrej Brodnik, Ivan Bruha, Maurice Bruynooghe, Wray Buntine, Dumitru Dan Burdescu, Yunlong Cai, Juan Carlos Cano, Tianyu Cao, Norman Carver, Marc Cavazza, Jianwen Chen, L.M. Cheng, Chou Cheng-Fu, Girija Chetty, G. Chiola, Yu-Chiun Chiou, Ivan Chorbev, Shauvik Roy Choudhary, Sherman S.M. Chow, Lawrence Chung, Mojca Ciglarič, Jean-Noël Colin, Vittorio Cortellessa, Jinsong Cui, Alfredo Cuzzocrea, Darko Čerepnalkoski, Gunetti Daniele, Grégoire Danoy, Manoranjan Dash, Paul Debevec, Fathi Debili, Carl James Debono, Joze Dedic, Abdelkader Dekdouk, Bart Demoen, Sareewan Dendamrongvit, Tingquan Deng, Anna Derezinska, Gaël Dias, Ivica Dimitrovski, Jana Dittmann, Simon Dobrišek, Quansheng Dou, Jeroen Doumen, Erik Dovgan, Branko Dragovich, Dejan Drajic, Jozo Dujmovic, Umut Riza Ertürk, CHEN Fei, Ling Feng, YiXiong Feng, Bogdan Filipič, Iztok Fister, Andres Flores, Vladimir Fomichov, Stefano Forli, Massimo Franceschet, Alberto Freitas, Jessica Fridrich, Scott Friedman, Chong Fu, Gabriel Fung, David Galindo, Andrea Gambarara, Matjaž Gams, Maria Ganzha, Juan Garbajosa, Rosella Gennari, David S. Goodsell, Jaydeep Gore, Miha Grčar, Daniel Grosse, Zhi-Hong Guan, Donatella Gubiani, Bidyut Gupta, Marjan Gusev, Zhu Haiping, Kathryn Hempstalk, Gareth Howells, Juha Hyvärinen, Dino Ienco, Natarajan Jaisankar, Domagoj Jakobovic, Imad Jawhar, Yue Jia, Ivan Jureta, Dani Juričić, Zdravko Kačič, Slobodan Kalajdziski, Yannis Kalantidis, Boštjan Kaluža, Dimitris Kanellopoulos, Rishi Kapoor, Andreas Kassler, Daniel S. Katz, Samee U. Khan, Mustafa Khattak, Elham Sahebkar Khorasani, Ivan Kitanovski, Tomaž Klobučar, Ján Kollár, Peter Korošec, Valery Korzhik, Agnes Koschmider, Jure Kovač, Andrej Krajnc, Miroslav Kubat, Matjaz Kukar, Anthony Kulis, Chi-Sung Laih, Niels Landwehr, Andreas Lang, Mohamed Layouni, Gregor Leban, Alex Lee, Yung-Chuan Lee, John Leggett, Aleš Leonardis, Guohui Li, Guo-Zheng Li, Jen Li, Xiang Li, Xue Li, Yinsheng Li, Yuanping Li, Shiguo Lian, Lejian Liao, Ja-Chen Lin, Huan Liu, Jun Liu, Xin Liu, Suzana Loskovska, Zhiguo Lu, Hongen Lu, Mitja Luštrek, Inga V. Lyustig, Luiza de Macedo, Matt Mahoney, Domen Marinčič, Dirk Marwede, Maja Matijasevic, Andrew C. McPherson, Andrew McPherson, Zuqiang Meng, France Mihelič, Nasro Min-Allah, Vojislav Misic, Vojislav Mišić, Mihai L. Mocanu, Angelo Montanari, Jesper Mosegaard, Martin Možina, Marta Mrak, Yi Mu, Josef Mula, Phivos Mylonas, Marco Di Natale, Pavol Navrat, Nadia Nedjah, R. Nejabati, Wilfred Ng, Zhicheng Ni, Fred Niederman, Omar Nouali, Franc Novak, Petteri Nurmi, Denis Obrul, Barbara Oliboni, Matjaž Pančur, Wei Pang, Gregor Papa, Marcin Paprzycki, Marek Paralič, Byung-Kwon Park, Torben Bach Pedersen, Gert Schmeltz Pedersen, Zhiyong Peng, Ruggero G. Pensa, Dana Petcu, Marko Petkovšek, Rok Piltaver, Vid Podpečan, Macario Polo, Victor Pomponiu, Elvira Popescu, Božidar Potočnik, S. R. M. Prasanna, Kresimir Pripuzic, Gabriele Puppis, HaiFeng Qian, Lin Qiao, Jean-Jacques Quisquater, Vladislav Rajkovič, Dejan Rakovic, Jean Ramaekers, Jan Ramon, Robert Ravnik, Wilfried Reimche, Blagoj Ristevski, Juan Antonio Rodriguez-Aguilar, Pankaj Rohatgi, Wilhelm Rossak, Eng. Sattar Sadkhan, Sattar B. Sadkhan, Khalid Saeed, Motoshi Saeki, Evangelos Sakkopoulos, M. H. Samadzadeh, MariaLuisa Sapino, Piervito Scaglioso, Walter Schempp, Barabara Koroušić Seljak, Mehrdad Senobari, Subramaniam Shamala, Zhongzhi Shi, LIAN Shiguo, Heung-Yeung Shum, Tian Song, Andrea Soppera, Alessandro Sorniotti, Liana Stanescu, Martin Steinebach, Damjan Strnad, Xinghua Sun, Marko Robnik Šikonja, Jurij Šilc, Igor Škrjanc, Hotaka Takizawa, Carolyn Talcott, Camillo J. Taylor, Drago Torkar, Christos Tranoris, Denis Trček, Katarina Trojacanec, Mike Tschierschke, Filip De Turck, Aleš Ude, Wim Vanhoof, Alessia Visconti, Vuk Vojisavljevic, Petar Vračar, Valentino Vranić, Chih-Hung Wang, Huaqing Wang, Hao Wang, Hui Wang, YunHong Wang, Anita Wasilewska, Sigrid Wenzel, Woldemar Wolynski, Jennifer Wong, Allan Wong, Stefan Wrobel, Konrad Wrona, Bin Wu, Xindong Wu, Li Xiang, Yan Xiang, Di Xiao, Fei Xie, Yuandong Yang, Chen Yong-Sheng, Jane Jia You, Ge Yu, Borut Zalik, Aleš Zamuda, Mansour Zand, Zheng Zhao, Dong Zheng, Jinhua Zheng, Albrecht Zimmermann, Blaž Zupan, Meng Zuqiang

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2013 (Volume 37) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Janez Perš)
Slovenian Artificial Intelligence Society (Dunja Mladenić)
Cognitive Science Society (Urban Kordeš)
Slovenian Society of Mathematicians, Physicists and Astronomers (Andrej Likar)
Automatic Control Society of Slovenia (Sašo Blažič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Vojteh Leskovšek)
ACM Slovenia (Andrej Brodnik)

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 37 Number 2 June 2013 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **37** (2013) Number 2, pp. 113–219