

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan (Eds)

Computer Science & Information Technology

Fifth International Conference on Computer Science, Engineering and
Applications (CCSEA-2015)
Dubai, UAE, January 23 ~ 24 - 2015



AIRCC

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

Sundarapandian Vaidyanathan,
R & D Centre,
Vel Tech University, India
E-mail: sundarvtu@gmail.com

ISSN: 2231 - 5403

ISBN: 978-1-921987-26-7

DOI : 10.5121/csit.2015.50201 - 10.5121/csit.2015.50218

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

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

Fifth International Conference on Computer Science, Engineering and Applications (CCSEA-2015) was held in Dubai, UAE, during January 23 ~ 24, 2015. Third International Conference on Data Mining & Knowledge Management Process (DKMP 2015), International Conference on Artificial Intelligence and Applications (AIFU-2015) and Fourth International Conference on Software Engineering and Applications (SEA-2015) were collocated with the CCSEA-2015. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCSEA-2015, DKMP-2015, AIFU-2015, SEA-2015 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CCSEA-2015, DKMP-2015, AIFU-2015, SEA-2015 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCSEA-2015, DKMP-2015, AIFU-2015, SEA-2015.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan

Organization

General Chair

Sundarapandian Vaidyanathan

Vel Tech University, India

Program Committee Members

Abd El-Aziz Ahmed
Abdelhamid Mansor
Abduladhim Ashtaiwi
Adam Przybylek
Ahmad Tayyar
Ahmed Abdul-Monem
Ali Chaabani
Ali Hussein
Alireza Afshari
Ankit Chaudhary
Anthony Amankwah
Anwasha Khasnobish,
Aram Aerakelyan
Ashraf A.Shahin
Assem A.hamied Mousa
Baghdad Atmani
Benyettou Abdelkader
Bo Zhao
Cecil Cantos
Chaabani Ali
Chaitan Chaitan
Chin-Chih Chang
Dac-Nhuong Le
Dusan Krokavec
Esmail Nojavani
Eva Esther Shalin Ebenezer
Faiyaz Ahamad
Farhad Nadi
Fatih Korkmaz
Geetha Ramani
Isa Maleki
Israa SH.Tawfic
Israashaker Alani
Issam Haamdi
Iyad Alazzam
Izzat Alsmadi
Jalel Akaichi
Javed Mohammed

Cairo University, Egypt
University of Khartoum, Sudan
University of Tripoli, Libya
Gdansk University of Technology, Poland
Isra University, Jordan
Taibah University, Saudi Arabia
National School of Engineering, Tunisia
Alexandria University, Egypt
Islamic Azad University, Iran
MUM University, USA
University of Ghana, Ghana
Jadavpur University, India
Yerevan State University, Armenia
Cairo university, Egypt
Cairo University, Egypt
University of Oran, Algeria
University USTO-MB, Algeria
Samsung Research America, United States
Enverga University Foundation, Philippines
National School of Engineering, Tunisia
New York Institute of Technology, USA
Chung Hua University, Taiwan
Haiphong University, Vietnam
Technical University of Kosice, Slovakia
University of Isfahan, Iran
Pentecost University College, Ghana
Integral University, India
Universiti Sains Malaysia, Malaysia
Cankiri Karatekin University, Turkey
Anna University, India
Islamic Azad University, Iran
GaziAntep University, Turkey
Gaziantep University, Turkey
Universite de Sfax, Tunisia
Yarmouk University, Jordan
Boise State University, USA
University of Tunis, Tunisia
New York Institute of Technology, USA

Kashif Mahmood	Telenor Research, Norway
Ke-Lin Du	Xonlink Inc, China
Kwan Hee Han	Gyeongsang National University, South Korea
Kinjal Roy	Indian Institutes of Technology, India
Laudson Souza	Integrated Faculties of Patos (FIP), Brazil
Laura Felice	Universidad Nacional del Centro, Argentina
Maleki I	Islamic Azad University, Iran
Manoj Franklin	University of Maryland, USA
Mansoul abdelhak	Universiy of Skikda, Algeria
Manuel S	Enverga University Foundation, Philippines
Maria Cecilia G. Cantos	Enverga University, Philippines
Meshrif Altruist	Aljouf University, Saudi Arabia
Mohamed Hashem Abd El-Aziz Ahmed	Ain Shams University, Egypt
Mohamed Issa	Zagazig University, Egypt
Mohamed Sahbi Bellamine	University of Carthage, Tunisia
Mohammed A. AlGhamdi	Umm AlQura University, Saudi Arabia
Mucahit Altintas	Istanbul Technical University, Turkey
Nagaratna P Hegde	Vasavi College of Engineering, India
Narges Shafieian	SADAD Informatics Corporation, Iran
Natarajan Meghanathan	Jackson State University, USA
Neetesh Saxena	SUNY Korea & SBU USA, South Korea
Nisheeth Joshi	Banasthali University, India
Noureddine Bouhmala	Buskerud and Vestfold University, Norway
Octavio Jose salcedo Parra	District University of Bogota, Colombia
Omar S.Soliman	Cairo University, Egypt
Parastou Shahsamandi	Islamic Azad University, Iran
Prabhat Mahanti	University of New Brunswick, Canada
Pulkit Vohra	University of Warwick, UK
Raed I. Hamed	University of Anbar, Iraq
Rahali Bouchra	University of Tlemcen Algeria, Algeria
Rahul Johari	Guru Gobind Singh University, India
Rajput H	Indian Institute of Technology, India
Rasi	Shiraz University of Technology, Iran
Reza Ebrahimi Atani	University of Guilan, Iran
Saad Darwish	Alexandria University, Egypt
Saad Mohamed Saad Darwish	University of Alexandria, Egypt
Saravanan Nagenthram	MIMOS Berhad, Malaysia
Sejdi Sejdiu	AAB University, Kosovo
Soumen Kanrar	Vehere Interactive, India
Stanley H Mneney	University of KwaZulu-Natal, South Africa
Sudip Kumar Sahana	Birla Institute of Technology, Mesra
Vijayalakshmi Saravanan	Vellore Institute of Technology university, India
Vishakha Tiwari	Dayalbagh Educational Institute, India
Wajeb Gharibi	Jazan University, KSA
Yao-Nan Lien	National Chengchi University, Taiwan
Yasser Rostamiyan	Islamic Azad University, Iran
Yingchi Mao	Hohai University, China

Technically Sponsored by

Networks & Communications Community (NCC)



Computer Science & Information Technology Community (CSITC)



Digital Signal & Image Processing Community (DSIPC)



Organized By



Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Fifth International Conference on Computer Science, Engineering and Applications (CCSEA-2015)

A Framework for Plagiarism Detection in Arabic Documents.....	01 - 09
<i>Imtiaz Hussain Khan, Muazzam Ahmed Siddiqui, Kamal Mansoor Jambi and Abobakr Ahmed Bagais</i>	
A Web Content Analytics Architecture for Malicious JavaScript Detection....	11 - 19
<i>JongHun Jung, Chae-tae Im, Soojin Yoon, hcbae</i>	
Semantic Extraction of Arabic Multiword Expressions.....	21 - 31
<i>Samah Meghawry, Abeer Elkorany, Akram Salah and Tarek Elghazaly</i>	
Analysis of Computational Complexity for HT-Based Fingerprint Alignment Algorithms on Java Card Environment.....	33 - 41
<i>Cynthia S. Mlambo, Meshack B. Shabalala and Fulufhelo V. Nelwamondo</i>	
Multiple User Interfaces and Crossplatform User Experience : Theoretical Foundations.....	43 - 57
<i>Khalid Majrashi, Margaret Hamilton and Alexandra L. Uitdenbogerd</i>	
Quality Assessment for Online IRIS Images.....	59 - 71
<i>Sisanda Makinana, Tendani Malumedzha and Fulufhelo V Nelwamondo</i>	
Application of Rhetorical Relations Between Sentences to Cluster-Based Text Summarization.....	73 - 92
<i>N. Adilah Hanin Zahri, Fumiyo Fukumoto, Matsyoshi Suguru and Ong Bi Lynn</i>	
An Empirical Evaluation of Cryptool in Teaching Computer Security.....	93 - 100
<i>Mabroka Maeref and Fatma Algali</i>	
Enterprise Data Protection : Meeting Requirements with Efficient and Cost Effective Methods.....	101 - 110
<i>Khaled Aldossari</i>	
E-Education with Facebook - A Social Network Service.....	111 - 121
<i>Mohammad Derawi</i>	
A New Hybrid Metric for Verifying Parallel Corpora of Arabic English.....	123 - 139
<i>Saad Alkahtani, Wei Liu and William J. Teahan</i>	

Intra-Cluster Routing with Backup Path in Sensor Networks..... 141 - 154
Turki Abdullah, Hyeoncheol Zin, Mary Wu and ChongGun Kim

Recognizing Named Entities in Turkish Tweets..... 155 - 162
Beyza Eken and A. Cüneyd Tantug

Third International Conference on Data Mining & Knowledge Management Process (DKMP 2015)

**Developing a Framework for Prediction of Human Performance Capability
Using Ensemble Techniques.....** 163 - 173
Gaurav Singh Thakur and Anubhav Gupta

**Knowledge Management in Higher Education : Applicability of LKMC
Model in Saudi Universities.....** 175 - 181
Farzana Shafique

International Conference on Artificial Intelligence and Applications (AIFU-2015)

**An Approximate Possibilistic Graphical Model for Computing Optimistic
Qualitative Decision.....** 183 - 196
BOUTOUHAMI Khaoula and KHELLAF Faiza

Real Time Clustering of Time Series Using Triangular Potentials..... 197 - 212
Aldo Pacchiano and Oliver J. Williams

Fourth International Conference on Software Engineering and Applications (SEA-2015)

**A Novel Approach Based on Topic Modeling for Clone Group
Mapping.....** 213 - 222
Ruixia Zhang, Liping Zhang, Huan Wang and Zhuo Chen

A FRAMEWORK FOR PLAGIARISM DETECTION IN ARABIC DOCUMENTS

Imtiaz Hussain Khan¹, Muazzam Ahmed Siddiqui², Kamal Mansoor
Jambi¹ and Abobakr Ahmed Bagais¹

¹Department of Computer Science, Faculty of Computing and Information
Technology, King Abdulaziz University, Saudi Arabia

²Department of Information Systems, Faculty of Computing and Information
Technology, King Abdulaziz University, Saudi Arabia

ihkhan@kau.edu.sa, maasiddiqui@kau.edu.sa, kjambi@kau.edu.sa,
abobakr.a.2012@gmail.com

ABSTRACT

We are developing a web-based plagiarism detection system to detect plagiarism in written Arabic documents. This paper describes the proposed framework of our plagiarism detection system. The proposed plagiarism detection framework comprises of two main components, one global and the other local. The global component is heuristics-based, in which a potentially plagiarized given document is used to construct a set of representative queries by using different best performing heuristics. These queries are then submitted to Google via Google's search API to retrieve candidate source documents from the Web. The local component carries out detailed similarity computations by combining different similarity computation techniques to check which parts of the given document are plagiarised and from which source documents retrieved from the Web. Since this is an ongoing research project, the quality of overall system is not evaluated yet.

KEYWORDS

Plagiarism Detection, Arabic NLP, Similarity Computation, Query Generation, Document Retrieval.

1. INTRODUCTION

Plagiarism is becoming a notorious problem in academic community. It occurs when someone uses the work of another person without proper acknowledgement to the original source. The plagiarism problem poses serious threats to academic integrity and with the advent of the Web, manual detection of plagiarism has become almost impossible. Over past two decades, automatic plagiarism detection has received significant attention in developing small- to large-scale plagiarism detection systems as a possible countermeasure. Given a text document, the task of a plagiarism detection system is to find if the document is copied partially or fully from other documents from the Web or any other repository of documents. It has been observed that plagiarists use different means to hide plagiarism so that a plagiarism detection system cannot catch plagiarism cases. In an interesting paper [1], Alzahrani and colleagues report different types of plagiarism, including verbatim/exact copy, near copy and modified copy. Whereas verbatim copy can easily be detected by a plagiarism detection system, modified copies pose real challenge

to find their original source because in such cases a plagiarist often makes heavy revisions in the original text by making use of structural and semantic changes.

Two approaches have commonly been used in developing such systems: extrinsic or external approach and intrinsic approach. The extrinsic plagiarism detection uses different techniques to find similarities among a suspicious document and a reference collection. In this approach, usually a document is represented as an n -dimensional vector where n is the number of terms or some derived features from the document. A number of measures are available to compute the similarity between vectors including Euclidean distance, Minkowski distance, Mahalanobis distance, Cosine similarity, Simple Matching Coefficient, and Jaccard similarity. This approach effectively detects verbatim or near copy cases, however, with the heavily modified copies the performance of an extrinsic-based plagiarism detection system is greatly reduced. On the other hand, in intrinsic plagiarism detection, the suspicious document is analyzed using different techniques in isolation, without taking a reference collection into account [2-3]. Assuming that a good-enough writing style analysis is available, this approach can effectively detect heavy-revision plagiarism cases or even plagiarism cases from a different language (multi-lingual plagiarism).

The research in automatic plagiarism so far has mostly been confined to English, paying little attention to other languages like Arabic. Research in automatic plagiarism detection for the Arabic language is much demanding and timely. This is because Arabic is the fourth most widely spoken language in the world, and most Arab countries, including Kingdom of Saudi Arabia, have adopted the use of e-learning systems in their educational institutions. In an e-learning environment, where students generally have an access to the World Wide Web, the problem of plagiarism can be very threatening. This calls for the development of state-of-the-art tools to automatically detect plagiarism in Arabic documents.

In this paper, we describe an ongoing plagiarism detection project which intends to develop an online plagiarism detection system for Arabic documents. The proposed plagiarism detection framework comprises of two main components, one global and the other local. The global component is heuristics-based, in which a potentially plagiarized given document is used to construct a set of representative queries. These queries are then submitted to Google via Google API to retrieve candidate source documents from the Web. Next, the local component carries out detailed similarity computations to detect if the given document was plagiarized from the documents retrieved from the Web or not.

Rest of this paper is organised as follows. In Section 2, related work is discussed. The scope of the proposed project and our approach is described in Section 3. The proposed plagiarism detection framework is outlined in Section 4 followed by discussion in Section 5. Finally, Section 6 concludes the paper.

2. BACKGROUND AND RELATED WORK

A considerable amount of research has focused on automatic plagiarism detection. Here we review some interesting literature on automatic plagiarism detection in natural language text; an in-depth discussion can be found in [4]. Various approaches have been proposed in past two decades to automatically find plagiarism in written documents. Earlier approaches are mainly based on fingerprinting, keyword matching (or term occurrence) and style analysis [5-10]. Brin et al. [5] developed COPS, a system designed to detect plagiarism in research articles using fingerprinting mechanism. Their system works in two phases. In a first phase, they eliminate the most common sentences, and then in a second phase the remaining text is compared to detect plagiarism. A notable limitation of their system is that it is based on exact copy of sentences and

therefore cannot deal with paraphrases, for example. Building on COPS, Shivakumar and Garcia-Molina [6] developed SCAM system for detecting identical documents, based on word level analysis. In SCAM, the original documents are registered to a dedicated server; an attempt to register plagiarised documents can be detected by comparing the latter with the already stored documents. This system works reasonably well for documents with high degree of overlap; however, its performance degrades significantly when there are small overlaps. Si et al. [7] developed CHECK, a plagiarism detection system for documents written in the same domain, for example Physics. Their system works incrementally: first they compare a set of primary keywords in the suspected and source documents, followed by a more fine-grained comparisons only if there was similarity at the top level. This is the kind of approach we aim to adopt in our research, but we also aim to build a plagiarism detection system in a domain independent way. In [8], Broder used document fingerprints to detect the overall similarity between suspected and source documents. He chose the smallest k-gram hashes from the entire document which permits detection of overall similarity between documents for duplicate detection, but not smaller overlaps between documents. Monostori and colleagues [9] built MatchDetectReveal system, which uses algorithms for exact string comparison. They represent the suspected document as a suffix tree data structure, without any loss of information, and then compare this document with other documents represented as strings of texts. The accuracy of their approach is good enough, but this is very time consuming and also requires a lot of space. In [10], Khmelev and Teahan, instead of using suffix trees, adopted the idea of suffix arrays to reduce the memory problem found in suffix trees. However, both Monostori et al. and Khmelev and Teahan do not take paraphrases into account.

Recently, some researchers have proposed to use natural language processing techniques to plagiarism detection [11-15]. Runeson et al. [11] proposed shallow NLP approaches to detect duplicate documents. They used tokenisation, stemming, and stop-word removal. Although the techniques used were simple, the authors reported promising results. Leung and Chan [12] put forward some proposals to apply advanced NLP techniques, including syntactic and semantic analysis to improve automatic plagiarism detection. They proposed to use WordNet to find the synonyms of the keywords used in the document under scrutiny, and compare these synonyms with the documents in the database. If it is suspected that the document under scrutiny contains some contents from the database, the sentences of the document would be further analysed for detailed comparison. In another study [13], Ceska and Fox applied pre-processing techniques to improve automatic plagiarism detection. They used simple heuristics, including numbers' replacement by dummy symbols, removing punctuations, and lemmatisation. Their results suggest a significant impact of applying NLP to plagiarism detection. In yet another study [14], Chong and colleagues applied various NLP techniques, varying from shallow techniques (e.g. simple string matching) to more advanced techniques (e.g. structure analysis of text). They used similarity metrics, including tri-gram similarity and longest common subsequence to measure the similarity scores between suspected and original documents. These similarity scores were then used to train a model which, rather than binary classification, classifies the documents under scrutiny into four levels: exact copy, light revisions, heavy revisions, and no plagiarism. They report promising results.

Very recently, Arabic NLP community has shown interest in developing plagiarism detection systems for Arabic language [16-19]. In [16], Alzahrani and Salim reported on an Arabic plagiarism detection system which combines the fuzzy similarity model and semantic similarity model derived from a lexical database. First, they retrieve a list of candidate documents for each suspicious document using shingling and Jaccard coefficient, and then they make sentence-wise detailed comparison between the suspicious and associated candidate documents using the fuzzy similarity model. Their preliminary results indicate that fuzzy semantic-based similarity model can be used to detect plagiarism in Arabic documents. In another study, Bensalem and colleagues [17] have developed a system which uses various stylistic features to account for intrinsic

plagiarism. The system was evaluated on a small corpus, so it is difficult to quantify its effectiveness. In yet another study, Menai [18] used a top-down approach, whereby in a first step a global similarity is measured between a suspicious document and candidate documents. In a second step, a detailed analysis is done at paragraph- and sentence-level.

3. SCOPE AND APPROACH

It is important to mention at the outset what is the scope of our work. We are interested in developing a web-based plagiarism detection system which can detect plagiarism cases in written Arabic documents. The scope of this project is limited to Arabic natural language text and we do not take plagiarism in a programming language code into account. Also we do not consider multi-lingual plagiarism. Rather, we address mono-lingual plagiarism in relatively smaller domain, student assignments and small-scale research papers (whose length is less than 50 pages). Moreover, we assume that the input suspicious document is plain text or a word document only. We do not consider other file formats like .pdf or .eps nor other modalities for example images. These assumptions will help us evaluate our system in a more informed and systematic way.

Our approach is hybrid, that is we incorporate both intrinsic and extrinsic techniques in the single framework. The former is mainly used in this project to generate queries to retrieve candidate documents, whereas the latter is used to thoroughly compute similarity between potential plagiarised and source documents. We consider the problem of plagiarism detection as falling under the general problem of finding similarity among documents.

4. THE PLAGIARISM DETECTION FRAMEWORK

The proposed plagiarism detection framework comprises of two main components, one global and the other local. The global component is heuristics-based, in which a potentially plagiarized given document is used to construct a set of representative queries. These queries are then submitted to Google via Google API to retrieve candidate source documents from the Web. Next, the local component carries out detailed similarity computations to detect if the given document was plagiarized from the documents retrieved from the Web or not. The plagiarism detection framework is depicted in Figure 1.

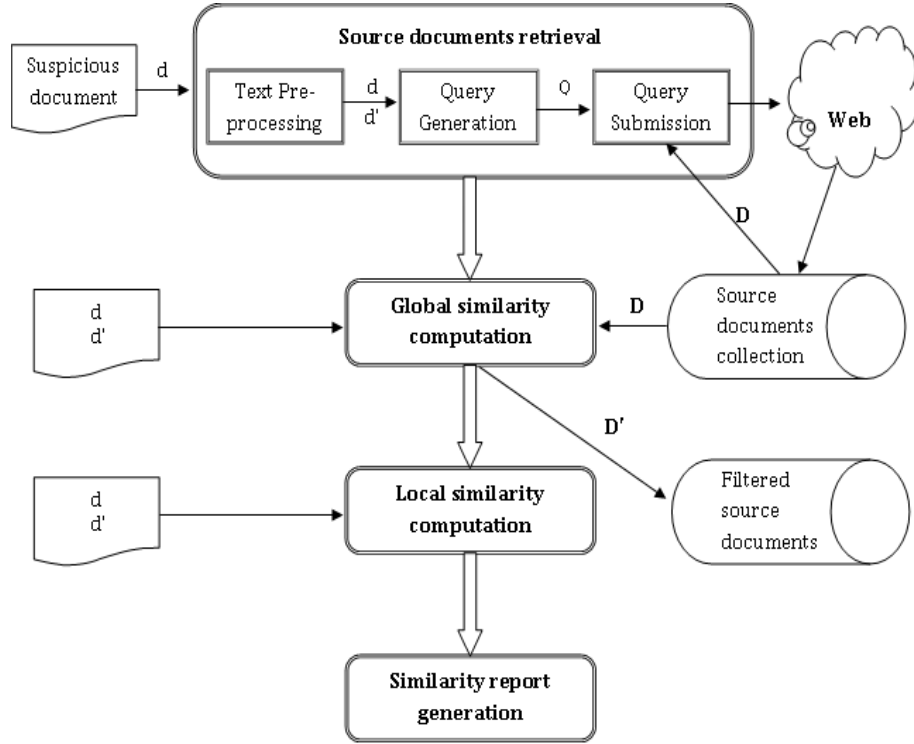


Figure 1: Plagiarism detection framework

In what follows, each component of the proposed framework is discussed in turn. (It is important to mention here that source document retrieval (4.1) is fully implemented and thoroughly evaluated; global and local similarity components (4.2, 4.3) are partially implemented and tested.)

4.1. Source Document Retrieval

We developed an information retrieval system which attempts to retrieve source documents from the Web against a given suspicious document. The system takes the suspicious document d as an input and goes through the following steps to find the potential source documents.

- a) **Text pre-processing step:** The system starts by pre-processing the input suspicious document d . First of all, the document d is converted into a standard UTF-8 file format. Next, it is tokenized into words using a custom-built Java tokeniser. The resulting words are then converted to their base form using Khoja stemmer [20]. Then, the document is segmented into sentences which allows line-by-line processing in the subsequent phases. Finally, the stopwords (functional words which are common across documents, for example في meaning in) and other punctuation marks are removed to generate a pre-processed document d' .
- b) **Query generation step:** Different query generation heuristics are used to generate a set of queries Q from the given suspicious document d . The query generation module takes the document d , the pre-processed document d' and the query generation heuristic as input and returns a set of queries Q as output. We developed different document retrieval heuristics, including key-phrase based heuristic, variance in readability score across sentences and first sentence in every paragraph of the document. These heuristics were

thoroughly evaluated in terms of precision, recall and f-measure on a sizeable corpus [21] and selected the top three best performing heuristics. The evaluation results also showed that a combination of those three select heuristics was significantly better than the each individual heuristic that is why we combine them for the source document retrieval. In is instructive to briefly describe here one heuristic, namely key-phrases based heuristic (for details, see [18]). This heuristic takes the pre-processed document d . We sampled a set of top N ($N = 5$ in this study) distinct keywords, based on the frequency of each word in the entire document. Then, for each keyword we constructed a phrase (henceforth key phrase) by taking two preceding and two succeeding words, at its first appearance in the original document (i.e., without preprocessing). If the keyword appeared at the beginning (or end) of a sentence, four preceding (or four succeeding words) words were used to construct the key phrase. An example, key phrase is "جيل من الطلاب قادر على" ("A generation of students capable of"), in which keyword is underlined. It is important to mention here that we developed different heuristics and thoroughly evaluated their performance in terms of precision, recall and f-measure on the corpus [21] developed as part of our project.

- c) **Query submission step:** Queries Q are submitted (one at a time) to the Web via Google's search API to retrieve source documents. Google's search API attempts to find relevant documents from the Web and returns the results including URL of the source document. Subsequently, these URLs are extracted from the returned results and the respective documents (at the URL) are downloaded and saved locally. The query submission procedure works as follows. The first query is submitted to the Web and top 10 matching documents are downloaded, maintaining a set D of documents. Subsequently, a query is only submitted to the Web if its *extension*, denoted as $[[q]]$, does not contain a document in the local document collection D . Extension of a query, $[[q]]$, is a set of documents which contains q . Our query submission approach is similar in spirit to Haggag and El-Beltagy [22], but we compute $[[q]]$ in a different way. In Haggag and El-Beltagy's case, a document $d \in D$ is in the $[[q]]$ set if 60% or above tokens in q are also present in d . They do not take position of those tokens into account though. We compute the extension of a query $[[q]]$ by using Ferret tri-gram model [23]. Accordingly, a document $d \in D$ is in the $[[q]]$ set if a sequence of three query words appear in d . It is important to remember here that we use 5-words long queries, generated in the previous step (see above).

4.2. Global Similarity Computation

After downloading the source documents from the Web in a local repository, the next step is the detailed similarity analysis to find which parts of the suspicious document are plagiarised from which documents in D . However, before carrying out this task, the source document collection D needs some necessary pre-processing. This is because the documents in D may contain some unnecessary HTML tags, which need to be cleaned up to extract the actual text. We implemented an HTML clean-up module which does the necessary clean up. Moreover, the source documents are converted into one single file format UTF-8, which is also the format of the given suspicious document.

Before computing the detailed similarity between suspicious document and documents in D , it is important to incorporate some filtration process to discard some documents from D which may have very little similarity with the suspicious document. This is important to avoid some unnecessary computation, which may degrade the overall efficiency of the system. However, this step should provide a reasonable balance between computational cost and accuracy of the system. That is, only some unwanted documents from D should be filtered out with minimum computational cost. To achieve such a balance, we employed a simple document-level similarity

heuristic which computes the similarity between the suspicious document d and a source document s as follows (equation 1).

$$sim(d,s) = \left(\frac{|d \cap s|}{\min(|d|, |s|)} \right) \quad \text{-----} \quad (1)$$

We discard the document s from D (resulting a new document collection D' , cf. Figure 1) if the similarity score sim is less than 0.2: experts suggest that around 20% similarity between two documents may not be considered as plagiarism. A preliminary investigation of our own corpus reveals that this similarity threshold (i.e. 0.2) is reasonable.

4.3. Local Similarity Computation

As mentioned earlier, we use both extrinsic and intrinsic plagiarism detection techniques to compute similarity between two documents. The detailed similarity computation module combines different similarity measures, including Euclidean distance, Minkowski distance, Mahalanobis distance, Cosine similarity, Simple Matching Coefficient, and Jaccard similarity, to find one final similarity score. The similarity between two documents (d and s) will be computed across two dimensions, precision and recall. Recall will indicate how much of d matches s , and precision will indicate the level of similarity, e.g. exact or near copy.

The local similarity module will also be spotting which sentence (or phrase of at least 5 consecutive words) is plagiarised from which source document on the Web. Such a pairing will be shown in the similarity report generated in the next step.

4.4. Similarity Report Generation

Finally, a similarity report for the suspicious document d will be generated, where the plagiarized parts of d will be highlighted with different colors indicating the source as shown in iThenticate and other well known plagiarism detection systems like Turnitin.

5. DISCUSSION

Building automatic plagiarism detection systems has gain much popularity and attention over past 20 years. Different plagiarism detection systems have been developed, however, the challenge still remains how to effectively identify the plagiarism cases. The challenge is even worse for Arabic language because of its complex and morphologically rich nature. In this paper, we proposed a plagiarism detection framework for Arabic. This research raised some interesting questions some of them were unexpected:

- Performance of our system is partially dependant on the accuracy of Google search results. This is because in a first step, we retrieve potential source documents from the Web using Google's search API. We believe that with the improvement of Google search techniques, particularly use of synonymy and other related techniques accuracy of our system would increase significantly. This is important, because if potential source documents are not retrieved in this initial stage, accuracy of subsequent stages would degrade accordingly.
- Google has limitation on maximum number of submissions per day: 100 queries per free-subscription account. Also, it places limits on query length. Moreover, Google results contain HTML tags in the returned documents, so HTML cleanup is necessary to extract the actual text.
- Global similarity computation may exclude some potential source documents. Care must be taken in selecting an appropriate threshold value. A preliminary value 0.2 seems

reasonable but more and thorough experimentation is needed to adjust this threshold value.

- One important aspect of Arabic writing came to fore during the corpus analysis. In Arabic documents, sentence length vary pretty unpredictably: we found sentences of length 3 words only or as maximum as 250 words. This may greatly affect intrinsic techniques to plagiarism detection which are mainly based on readability score.

6. CONCLUSION

This paper described the proposed plagiarism detection framework for Arabic documents. The proposed plagiarism detection framework comprises of two main components, one global and the other local. The global component is heuristics-based, in which a potentially plagiarized given document is used to construct a set of representative queries by using different best performing heuristics. These queries are then submitted to Google via Google's search API to retrieve candidate source documents from the Web. Next, the local component carries out detailed similarity computations to detect if the given document was plagiarized from the documents retrieved from the Web or not. The global component is thoroughly evaluated, whereas the local component is partially implemented so far.

In future, we intend to integrate the different components of the system to build one final web-based plagiarism detection system. We will be thoroughly investigating the performance of different similarity measures before incorporating them in the final similarity computation model. The implemented system would then be thoroughly evaluated using our own corpus [21] before deploying.

ACKNOWLEDGEMENTS

This work was supported by a King Abdulaziz City of Science and Technology (KACST) funding (Grant No. 11-INF-1520-03). We thank KACST for their financial support.

REFERENCES

- [1] Alzahrani, S.M., Salim, N.& Abraham, A.(2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews*, 42(2), pp. 133-149.
- [2] Eissen, M., Stein, B. & Kulig, M.(2007). Plagiarism detection without reference collections. In *Proceedings of the advances in data analysis*, pp. 359–366.
- [3] Benno, S., Moshe, K. & Efstathios, S.(2007). Plagiarism analysis, authorship identification, and near-duplicate detection. In *Proceedings of the ACM SIGIR Forum PAN07*, pp 68–71, New York.
- [4] Clough, P. (2003). Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service*, (February edition).
- [5] Brin, S., Davis, J., & Garcia-Molina, H.(1995). Copy detection mechanisms for digital documents. In *proceedings of the ACM SIGMOD annual conference*.
- [6] Shivakumar, N., & Garcia-Molina, H.(1996). Building a scalable and accurate copy detection mechanism. *Proceedings of the first ACM international conference on digital libraries*.
- [7] Si, Leong, H.V., & Lau, R.W.(97). CHECK: A document plagiarism detection system. In *Proceedings of ACM symposium for applied computing*, pp. 70-77.
- [8] Broder, A.Z. (1997). On the resemblance and containment of documents. In *compression and complexity of sequences* , pp. 21-29.
- [9] Monostori, K., Zaslavsky, A., & Schmidt, H. (2000). MatchDetectReveal: Finding overlapping and similar digital documents. In *proceedings of information resources management association international conference*, pp. 955-957.

- [10] Khmelev, D., & Teahan, W. (2003). A repetition based measure for verification of text collections and for text categorization. In Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, pp. 104-110.
- [11] Runeson, P., Alexandersson, M., & Nyholm, O. (2007). Detection of duplicate defect reports using natural language processing. In proceedings of 29th international conference on software engineering, pp. 499-510.
- [12] Leung, C.-H., & Chan, Y.-Y. (2007). A natural language processing approach to automatic plagiarism detection. In proceedings of the 8th ACM SIGITE conference on information technology education, (pp. 213-218).
- [13] Androutsopoulos, I., & Malakasiotis, P.(2009). A Survey of paraphrasing and textual entailment methods. Technical report, Athens University of Economics and Business, Greece.
- [14] Ceska, Z., & Fox, C.(2009). The influence of text pre-processing on plagiarism detection. In recent advances in natural language processing, RANLP'09 .
- [15] Chong, M., Specia, L., & Mitkov, R. (2010). Using natural language processing for automatic detection of plagiarism. In proceedings of 4th international plagiarism conference.
- [16] Alzahrani, S.M. & Salim, N. (2009) Fuzzy semantic-based string similarity for extrinsic plagiarism detection. In Proceedings of the 2nd international conference on the applications of digital information and Web technologies., London, UK.
- [17] Bensalem, I.Rosso, P. & Chikhi, S. (2012). Intrinsic plagiarism detection in Arabic text: preliminary experiments. In Proceedings of the 2nd Spanish conference on information retrieval, Spain.
- [18] Menai, M.(2012) Detection of plagiarism in Arabic documents. International journal of information technology and computer science (IJITCS), 4(10).
- [19] Khan, I.H.,Siddiqui, M. Jambi, K. M., Imran, M & Bagais, A. A. (2014). Query optimization in Arabic plagiarism detection: an empirical study. To appear in International Journal of Intelligent Systems and Applications.
- [20] Khoja, S.(1999). Stemming Arabic Text. Online available: <http://zeus.cs.pacificu.edu/shereen/research.htm>.
- [21] Siddiqui, M.A., Elhag, S.,Khan, I.H., & Jambi, K. M. Building an Arabic plagiarism detection corpus. To appear in language resources and engineering.
- [22] Haggag, O. & El-Beltagy, S. (2013). Plagiarism candidate retrieval using selective query formulation and discriminative query scoring. In proceedings of PAN, CLEF.
- [23] Ferret (2009). Online available at University of Hertfordshire: <http://homepages.feis.herts.ac.uk/~pdgroup/>.

AUTHORS

Imtiaz Hussain Khan is an assistant professor in Department of Computer Science at King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. He received his MS in Computer Science from the University of Essex UK in 2005 and PhD in Natural Language Generation from the University of Aberdeen UK in 2010. His areas of research are Natural Language Processing and Evolutionary Computation.

Muazzam Ahmed Siddiqui is an assistant professor at the Faculty of Computing and Information Technology, King Abdulaziz University. He received his BE in electrical engineering from NED University of Engineering and Technology, Pakistan, and MS in computer science and PhD in modeling and simulation from University of Central Florida. His research interests include text mining, information extraction, data mining and machine learning.

Kamal Mansoor Jambi is a professor in Department of Computer Science at King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. He received his Masters (Computer Science) degree from Michigan State University, USA in 1986. He earned his PhD (Artificial Intelligence, OCR) degree from the Illinois Institute of Technology, Chicago, USA in 1991. His areas of research are Natural Language Processing, Speech Recognition, OCR and image processing.

Abobakr Ahmed Bagais received his BSc degree in Computer Science from King Abdulaziz University, Saudi Arabia. He is currently pursuing M.E. in Network optimization from King Abdul-Aziz University. His areas of interest include Arabic natural language processing, bioinformatics and optimization network.

INTENTIONAL BLANK

A WEB CONTENT ANALYTICS ARCHITECTURE FOR MALICIOUS JAVASCRIPT DETECTION

JongHun Jung, Chae-tae Im, Soojin Yoon, hcbae

Internet Incidents Response Architecture Team
Korea Internet & Security Agency

{jjh2640, chtim, sjyoon, hcbae}@kisa.or.kr

ABSTRACT

Recent web-based cyber attacks are evolving into a new form of attacks such as private information theft and DDoS attack exploiting JavaScript within a web page. These attacks can be made just by accessing a web site without distribution of malicious codes and infection. Script-based cyber attacks are hard to detect with traditional security equipments such as Firewall and IPS because they inject malicious scripts in a response message for a normal web request. Furthermore, they are hard to trace because attacks such as DDoS can be made just by visiting a web page. Due to these reasons, it is expected that they could result in direct damages and great ripple effects. To cope with these issues, in this article, a proposal is made for techniques that are used to detect malicious scripts through real-time web content analysis and to automatically generate detection signatures for malicious JavaScript.

KEYWORDS

Script-based Cyber Attacks; Forward-Proxy Server; Malicious Java Script API; Deep Content Inspection; API Call Trace.

1. INTRODUCTION

Recent introduction of Ajax and HTML5 technologies has enabled dynamic representation of web content, providing compatibility between a client and a server in web environment. However, the efforts to deal with new security vulnerabilities in these technologies, such as the awareness, countermeasure technology development, and standardization are still insufficient. In particular, web-based attacks using malicious scripts can bypass traditional security equipments, such as IDS, IPS and Web Firewall, because, unlike conventional malicious code attacks, they do not download an executable file directly, but they still can be made by combining normal built-in APIs in JavaScript. And also, it is getting harder to detect these attacks as they employ traffic encryption and script obfuscation. The Figure 1 illustrates how a DDoS attack can be made with JavaScript just by accessing a web page. Furthermore, due to the accelerated introduction of HTML5, it is expected that cyber attacks exploiting vulnerabilities of new tags and APIs will grow rapidly.

In this article, a proposal is made for techniques that are used to detect malicious scripts through collection of HTTP Web traffics and static/dynamic analysis, and to generate a detection signature automatically. Chapter 2 shows trends in related studies. Chapter 3 describes techniques

that are used to collect and analyze web content for detection of malicious JavaScript. Chapter 4 describes more compact techniques that are used to generate a detection signature automatically with less false positive rate. Finally, Chapter 4 concludes the article.

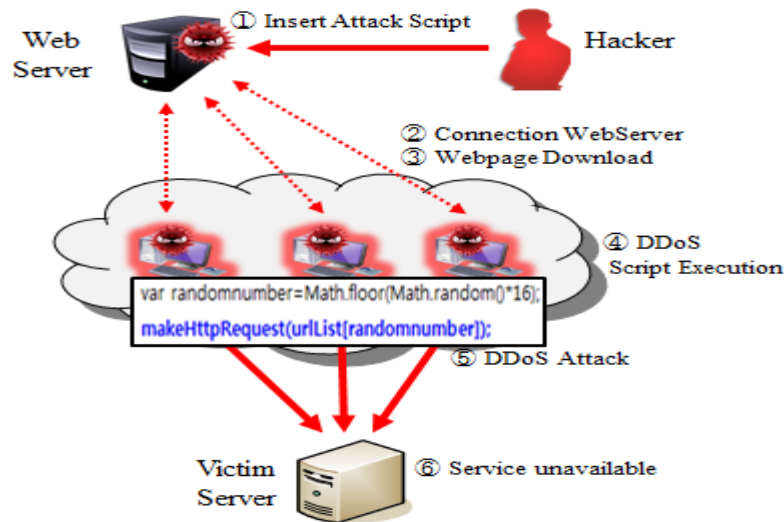


Figure 1. A DDoS attack using JavaScript code

2. RELATED WORK

2.1. WebShiled

Using a modified browser that consists of DOM API, HTML/CSS Parser and JavaScript Engine only, parse web content in proxy, turn it into the form of DOM Structure, and store it. Send the DOM Structure in the string format to the client. Send a script to the client only if it turns out that the script is safe after running it in the modified browser. However, prevention of exploitation of new vulnerabilities in HTML5 is insufficient.

2.2. A signature for Malware Detection

The method of Automatic generation of a signature for malware or worm can be divided into 5 categories: vulnerability-based, content-based, content-shifting, semantic-aware and honeypot-based. Among these, the content-based is the one that is proposed in this article.

In the content-based method, a signature target set is determined based on traffic and the same malicious behavior, and then a signature is generated based on the content.

A content-based signature [1] can be divided into Longest Common Substring, Longest Common Subsequence, Conjunction Signature and Bayes Signature. For the Longest Common Substring and Longest Common Subsequence, one retrieves the longest common substring and longest common subsequence respectively from the target set. For the Conjunction Signature, one uses a set of strings that appear in all targets, as a signature. For Bayes Signature, one checks whether a string in a sample appears in the malicious, and then determines whether the sample is malicious or not based on the percentage of malicious strings.

2.3 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH is an algorithm for hierarchical clustering for a large database. BIRCH allows addition of a new value in a clustered tree as a new entity is added, eliminating the need of re-clustering.

BIRCH creates a CF (Clustering Feature) tree that has distance information for all leaves under a single node. As a new entity is added, it searches for the closest node. It adds the entity to the cluster of the closest node if the distance is same or shorter than threshold, or creates a new cluster and adds the entity to it if the distance is same or longer than threshold.

3. WEB CONTENT ANALYSIS TECHNOLOGY

For real-time detection of malicious JavaScript, one collects HTTP traffics by configuring a proxy server, and parses a HTML document and crawls a link to external resource in order to generate content for analysis. One performs static analysis, such as pattern-matching of web content, and dynamic analysis, such as checking whether obfuscated or not and the HTML5 tag percentage, to determine if the content is malicious. If a malicious script is found, remove the function that actually causes malicious behaviors before sending the script to the client. The Figure 2 shows the proposed system architecture that can be used to detect malicious scripts at network level.

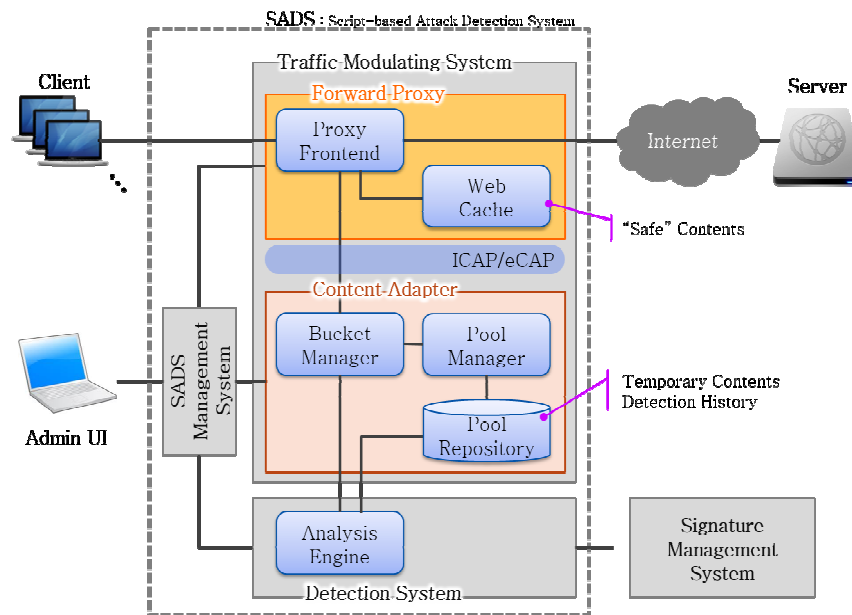


Figure2. System Architecture for Real-time Detection of Malicious Scripts

It consists of modules: i) Forward-Proxy, ii) Web Content Generation Module, iii) Analytics Engine (Static/Dynamic Analysis), and iv) System Control Module.

3.1 Forward-Proxy and Content Adapter

For collection of web traffics, Squid-Proxy Server is configured in the in-line form between clients and Web Server, where all HTTP Request and Response packets are collected and Internet Content Adaptation Protocol (ICAP) is used to pass the received HTTP traffics to Web Content Control Module. Then, Web Content Control Module extracts the external JavaScript link data

contained in the document, using HTML Parser received from Proxy Server, and collects resources for the link with a separate crawler to generate web content for analysis.

3.2 Web Content Analysis

The term ‘web content’ refers to the entire document that includes both a HTML document and external resources. As shown in the Figure 3, web content goes through the fast static analysis process that performs pattern-matching based on Yara-RuleSet[2]. However, because some sources, such as those obfuscated, require additional analysis, they go through the dynamic analysis process that uses Rihno Browser Engine to run the script and extract call trace data for detection.

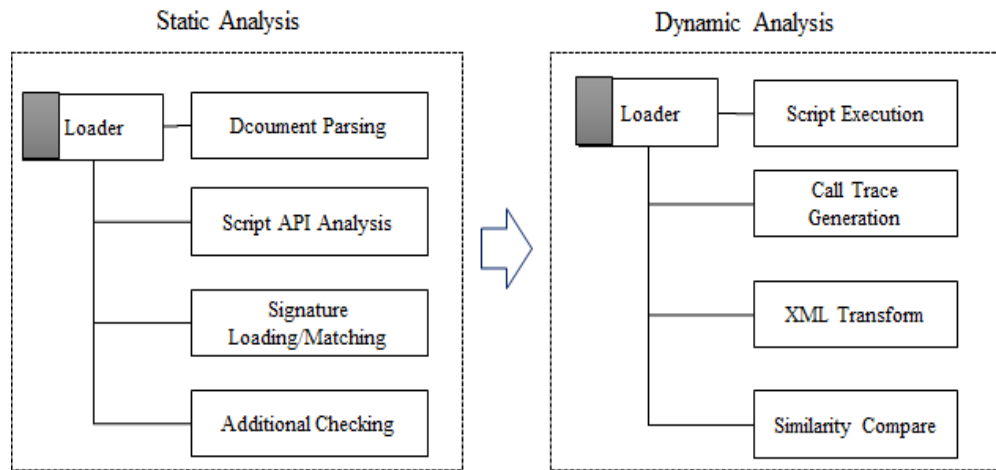


Figure3. Configuration of Web Content Analytics Engine

An input data set is in the format of JSON that consists of a HTML document, external JavaScript, and meta data (IP, port, protocol, domain, etc.). First, extract the primary key token to classify the type of the malicious behaviour. Table 1 shows summary of basic keywords contained in each malicious behaviour

Table 1. Examples of Basic Keywords for Each Malicious Behaviour

Malicious Type		The Keyword
DoS Attack	HashDoS	setInterval, open, send, ActiveXObject, XMLHttpRequest, XMLHttpRequest
	XML HttpObject DoS	
Scan Attack	Network Scan	open, ActiveXObject, XMLHttpRequest, XMLHttpRequest, Date, readyState
	Port Scan	
Geolocation		coords, getCurrentPosition
Web Socket		parse, eval, WebSocket, JSON, send
Web Worker DDoS		postMessage, Worker, XMLHttpRequest, open, send

Look up the signature for a malicious behaviour and then perform signature-matching check to determine whether it is malicious or not.

Additionally, score the JavaScript obfuscation and the percentage of HTML5 new tag usage in the entire document, and then perform dynamic analysis if the score is the same or above the predetermined level. JavaScript obfuscation check is performed because most of malicious JavaScript codes are obfuscated, and it is hard to determine whether it is malicious just by doing signature-matching during static analysis. The Figure 4 illustrates process of the JavaScript obfuscation check[3]. As these are main characteristics of the obfuscated JavaScript, if a special character in a JavaScript string is frequently used, if there is a string with abnormal length, or if the entropy score of characters in the JavaScript is low, score them and if the total score is the same or above the cutoff, consider it obfuscated and perform dynamic analysis additionally.

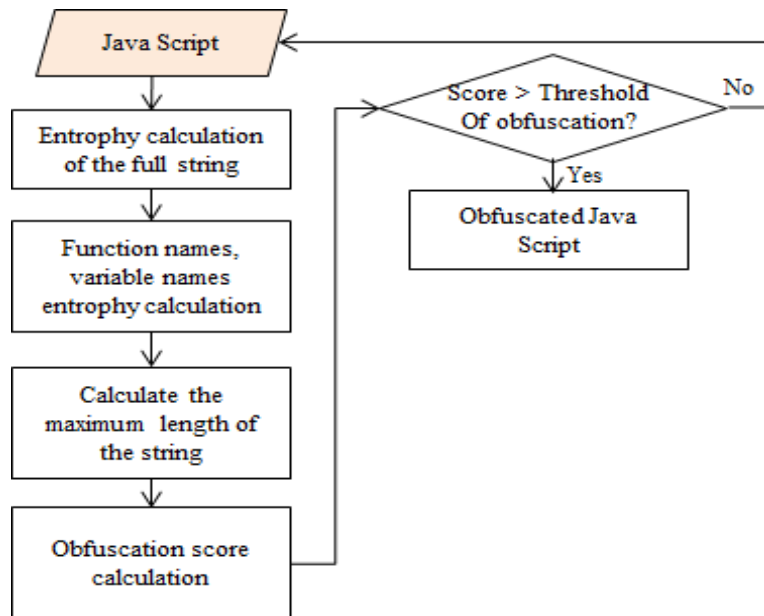


Figure4. JavaScript Obfuscation Analysis Process

And also, the usage of HTML5 tags is checked to detect malicious scripts such as jacking or cross-site scripts exploiting new tags of HTML5 (Canvas, Audio, Video). It has been arranged to perform dynamic analysis if a weight for each HTML5 tag is applied and the score is the same or above the predetermined level.

During dynamic analysis in real world situation, a malicious JavaScript is executed using open source-based Rhino JavaScript Engine with a built-in sandbox, and JavaScript API Call Trace data is extracted and stored in XML data format.

The Figure 5 shows the Function Call Trace data for Port Scan malicious JavaScript, converted to XML format.

```

1 <root>
2 <document.write>
3 <P1><div id="comments_threads"><Comments.</div></P1>
4 <Loc>Sample1:14398</Loc>
5 </document.write>
6 <setInterval>
7 <P1>100</P1>
8 <P2>function startRequest() {
9   createXMLHttpRequest();
10  xmlHttp.onreadystatechange = handleStateChange;
11  xmlHttp.open("GET", settingUrl, false);
12  xmlHttp.send();}</P2>
13 <Loc>Sample1:15232</Loc>
14 </setInterval>
15 <XMLHttpRequest.open>
16 <P1>GET</P1>
17 <P2>http://192.168.159.133</P2>
18 <P3>false</P3>
19 <Loc>Sample1:15622</Loc>
20 </XMLHttpRequest.open>
21 <XMLHttpRequest.send>
22 <Loc>Sample1:15665</Loc>
23 </XMLHttpRequest.send>
24 </root>

```

<setInterval>

<XMLHttpRequest.open>

<XMLHttpRequest.send>

Figure 5. Trace Data of a Port Scan Malicious Script

In this article, SimHash Algorithm[4] is proposed for comparison of JavaScript Function Call Trace similarities. SimHash utilizes Local Sensitive Hashing (LSH) for similarity comparison, and LSH maximizes conflicts between similar items rather than avoiding them. That is, the algorithm generates similar results for similar items. Using this function, regardless of the input value size, generate FingerPrint in an array in bit form just like the outcome of a normal hash function, and then use the hamming distance to measure the similarity.

4. TECHNIQUE OF GENERATING A SIGNATURE DEDICATED FOR DETECTION

In this article, the malicious script, malicious type, obfuscation status, meta data and other data received from the analytics engine are used for automatic generation of signature for malicious JavaScript. It is proposed that a detection signature can be automatically generated by clustering with a malicious script from the registered malicious script pool, generating the combined signature, and refining the signature. Figure 6 illustrates the process of signature generation.

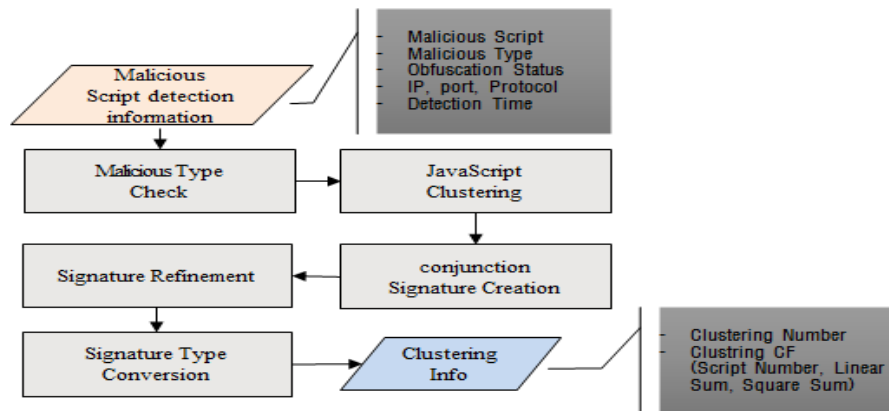


Figure6. Automatic Generation of the Detection Signature for a Malicious Script

4.1 Malicious Script Clustering

In this article, it is proposed to use the script clustering technique for automatic generation of the detection signature for a malicious script. The goal of clustering is to streamline the signature itself and improve the false positive rate by grouping malicious scripts showing similar behaviors, and thus preventing extraction of unnecessary tokens. For each token of malicious JavaScript, calculate the Term Frequency-Inverse Document Frequency value and vectorize it. The TF-IDF[5] weight is a statistical figure that is used to evaluate the importance of a certain term in a document, and it can be calculated as the product of Term Frequency and Inverse Document frequency.

The Term Frequency simply indicates how often a term appears in the document, and the Inverse Document Frequency provides general importance of the term.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

- $n_{i,j}$ indicates the number of times that Term t_i appears in document d_j .

$$idf_i = \log \frac{|D|}{|d_i : t_i \in d_j|} \quad (2)$$

- $|D|$ indicates Total Document Numbers
- $|d_i : t_i \in d_j|$ indicates number of documents in which term t_i appears

$$tfidf_{ij} = tf_{ij} * idf_i \quad (4)$$

- TF-IDF weight is calculated by multiplying the TF and IDF.

Using the vector created with TF-IDF, perform hierarchical clustering in the Complete-linkage Cluster method. By improving BIRCH Algorithm for hierarchical clustering, quantify the vector distance and meta data (time similarity, IP, port and protocol), and then take their sum as the similarity score to determine whether malicious script clustering can be done.

Figure 7 shows the clustering process through modified distance calculation. For clustering purpose, the score of significant meta data similarity is applied to distance measurements between basic vectors in order to form a clustering tree.

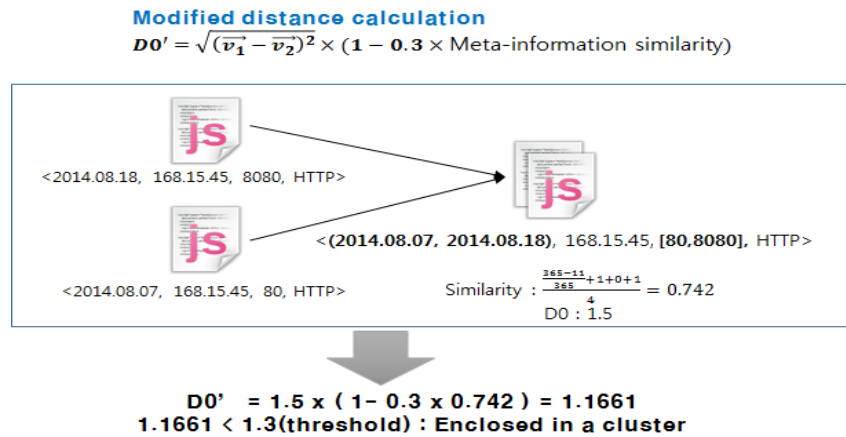


Figure7. Modified Distance Calculation Formula and Meta Similarity Application

4.2 Generating a Conjunction Signature

Extract a common token from a malicious script file within the allowed distance in a cluster to generate a conjunction signature. Convert a token in the same form, such as IP, to a regular expression before processing. The Table 2 shows the combined signature generated with Port Scan JavaScript.

Table 2. Examples of Conjunction Signature for Port Scan Detection

```
output, targetIP, endtime, starttime, ate, appendChild, break, wordWrap, createElement,
onRequest, ActiveXObject, majorPort, (?:25[0-5]|2[0-4][0-9]|01?[0-9][0-9]?)\.\. {3
}(?:25[0-5]|2[0-4][0-9]|01?[0-9][0-9]?)XMLHttpRequest, style, open, innerHTML, Array,
XMLHTTP, true, onreadystatechange, restime, Microsoft, Close, send, scanRes,
getElementById
```

4.3 Refining the Signature and Verifying the False Positive Rate

Verifying performance based on the detection signature generated shows that the number of unnecessary tokens or the false positive rate grows depending on the number of malicious script samples. Accordingly, an additional refinement of the signature is carried out by comparing with a token extracted from a normal web document, and eliminating the one that is duplicate or less than a certain length (3 characters).

The Table 3 shows the number of tokens and the false positive rate after the signature is refined. In this specific example, the signature has been compared against 28 malicious JavaScript codes and 300 JavaScript codes collected randomly for false positive verification. Group 3 shows the results after signature refinement. It can be seen that the false positive rate and the length of the signature generated (the number of tokens) have been significantly improved.

Table 3. The Results of Signature Refinement – False Positives

Section	Group 1	Group 2	Group 3
Average Detection Rate	100%	100%	100%
Average False Positive Rate	0%	8.8%	0%
Average Number of Tokens	146.7	89.5	115.9
The Number of Groupings	4 JavaScript Codes	9 JavaScript Codes	-
Refinement	×	×	○

5. CONCLUSION

In this article, a proposal has been made for techniques that are used to detect malicious JavaScript and to automatically generate detection signatures. While it shows good results if the signatures generated using the proposed techniques are employed to detect malicious scripts and measure the latency time, it requires additional experiments on a larger pool of samples and higher volume of traffics. Furthermore, to deal with security vulnerabilities of new APIs in HTML5, it is planned to expand the scope of the proposed dynamic analysis and conduct related

studies on detection of malicious behaviors by monitoring behaviors caused by JavaScript running,

ACKNOWLEDGMENT

This work was supported by the ICT R&D Program of MSIP/IITP. [14-912-06-002, The Development of Script-based Cyber Attack Protection Technology]

REFERENCES

- [1] Z. Li, M. Sanghi, Y. Chen, M. Y. Kao, and B. Chavez, "Hamsa: Fast signature generation for zero-day polymorphic worms with provable attack resilience.", IEEE Symposium on Security and Privacy, May 2006.
- [2] YARA Documentation, <http://yara.readthedocs.org/en/latest/index.html>
- [3] Xu, Wei, Fangfang Zhang, and Sencun Zhu. "The power of obfuscation techniques in malicious JavaScript code: A measurement study." Malicious and Unwanted Software (MALWARE), 2012 7th International Conference on. IEEE, 2012.
- [4] Charikar, Moses S. "Similarity estimation techniques from rounding algorithms." Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM, 2002.
- [5] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, Vol.28, No.1, 1972, pp.11-21.

INTENTIONAL BLANK

SEMANTIC EXTRACTION OF ARABIC MULTIWORD EXPRESSIONS

Samah Meghawry^{1,*}, Abeer Elkorany², Akram Salah², and
Tarek Elghazaly¹

¹Institute of statistical studies and research, computer science, Cairo University
samah1984@gmail.com, tarek.elghazaly@cu.edu.eg

²Faculty of Computers and information, computer science, Cairo University
a.korani@fci-cu.edu.eg, a.salah@fci-cu.edu.eg

ABSTRACT

A considerable interest has been given to Multiword Expression (MWEs) identification and treatment. The identification of MWEs affects the quality of results of different tasks heavily used in natural language processing (NLP) such as parsing and generation. Different approaches for MWEs identification have been applied such as statistical methods which employed as an inexpensive and language independent way of finding co-occurrence patterns. Another approach relies on linguistic methods for identification, which employ information such as part of speech (POS) filters and lexical alignment between languages is also used and produced more targeted candidate lists. This paper presents a framework for extracting Arabic MWEs (nominal or verbal MWEs) for bi-gram using hybrid approach. The proposed approach starts with applying statistical method and then utilizes linguistic rules in order to enhance the results by extracting only patterns that match relevant language rule. The proposed hybrid approach outperforms other traditional approaches.

KEYWORDS

Multiword expressions (MWEs), Statistical Measures, Part of speech tagging (POS), Nominal MWEs, verbal MWEs.

1. INTRODUCTION

Recent research on Multiword Expressions (MWEs) has devoted considerable attention to their identification. One of the problems that these works address is that MWEs can be defined as combinations of words that have idiosyncrasies in their lexical, syntactic, semantic, pragmatic or statistical properties. There is no uniform definition of MWEs. The definition of MWEs given by Sag is “any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts” [12]. In other words, Multiword expressions are groups of words which, taken together, can have unpredictable semantics. MWE is an important task in many applications such as automatic translation [1], ontology engineering and information retrieval [2]. There are two main approaches for extracting MWEs. The statistical approach that uses a set of standard statistical association measures based on frequency and co-occurrence such as T-score [3], log likelihood ratio (LLR) [4], FLR [5] and Mutual Information (MI3) [6] in order

to estimate the degree of association between its words. The second approach makes use of the rules of the language such as morphological, syntactic or semantic information implemented in language-specific rules. Alignment-based MWE extraction method, which lends itself to linguistic approach, looks for the sequences of source words that are frequently joined together during the alignment despite the number of target words involved. These MWE candidates may then be automatically validated, and the noisy non-MWE cases among them removed

However, each of those approaches suffers from great limitation [7], for example, statistical approaches “are unable to deal with low-frequency of MWEs”. On the other hand, linguistic approaches are “language dependent and not flexible enough to cope with complex structures of MWEs”. In order to overcome these weaknesses, a hybrid approach that combines statistical calculus and linguistic information is used. This paper proposes a framework for extracting Arabic Multiword Expressions from unannotated corpus using hybrid model that rely on frequency counts, statistical measures, and linguistic rules in order to create a refined list of candidates MWE. During the first phase of the proposed approach, lexical association measures based on the frequency distribution and co-occurrence patterns is applied in order to extract the first candidate set of MWE. Next, linguistics rules that utilize POS-tagger are applied to exclude specific patterns that match the relevant POS patterns according to Arabic grammar rules. In order to validate the effectiveness of the proposed model, three different Arabic corpuses were used during our experiments. Our experiments confirmed that the proposed approach outperform previous methods. This paper is organized as follows; Section2 presents different approaches applied for extracting MWEs for various languages. In section3 the proposed hybrid framework for Arabic MWE is illustrated. Results of experiment applied using different Arabic corpus are discussed in section4. Finally, section5 concludes the presented work and demonstrate potential future works.

2. RELATED WORK

A considerable amount of research has focused on the identification and extraction of MWEs. Given the heterogeneity of MWEs, different approaches were devised. Unfortunately, unlike in English, there is no capital letters in Arabic to distinguish the compound names and the geographical compound names. Statistical approaches have mostly been applied to bigrams and trigrams, and it becomes more problematic to extract MWEs of more than three words. Pecina evaluates 82 lexical association measures for the ranking of collocation candidates and concludes that it is not possible to select a single best universal measure, and that different measures give different results for different tasks depending on data, language, and the types of MWE that the task is focused on [14]. Similarly, Ramisch investigate the hypothesis that MWEs can be detected solely by looking at the distinct statistical properties of their individual words and conclude that the association measures can only detect trends and preferences in the co-occurrences of words [13]. The linguistic methods are based on linguistic information such as, morphological, syntactic and/or semantic information to generate the types of words. Traboulsi used the local grammar approach to extract person names from Arabic counterparts [11]. Harris defines a local grammar as a way of describing syntactic restrictions of certain subsets of sentences, which are closed under some or all of the operations in the language. Frozen expressions may be considered as a subset of sentences that have such syntactic restrictions. One can in fact observe restricted distributions over a number of words. Consider for example: Director of (company + thesis + conscience + *chocolate) (financial + stock + E) market The 20 March (next + 2006 + *bombastic) [10].

Hybrid approaches that combine the statistical approaches with the linguistic rules can cover a large part of the problem of MWEs identification and extraction [9]. Boulaknadel developed a multi-word term (MWT) extraction tool for Arabic. She adopted the standard approach that combined grammatical patterns and statistical score. First, she defined the linguistic specification of MWTs for Arabic language. Then, she developed a term extraction program and evaluated several statistical measures in order to filter the extracted term-like units for keeping the most representative of domain specific corpus [7]. Hybrid approaches may also combines the alignment technique with statistical approach like Helna [16] that proposed an approach for the identification of MWEs in a multilingual context, as a by-product of a word alignment process, that not only deals with the identification of possible MWE candidates, but also associates some multiword expressions with semantics.

3. HYBRID MODEL FOR ARABIC MWE EXTRACTION

The proposed model aims to extracts multi-word expressions from Arabic specialized corpora by combining statistical methods with linguistic rules. The standard approach to MWE identification is n-gram classification. However, our model is limited to multi-words composed of two elements (bigrams). This section discusses three different phases of the proposed model- the preprocessing phase, statistical phase and linguistic phase.

3.1 Preprocessing phase

Text preprocessing is the basic stage needed for MWE. Its main objective is, in one hand to remove all the unnecessary particles and mistyping words and in another hand to transform document contents to a suitable form which can be used easily by different algorithm. Thus during the preprocessing phase, we start by splitting the corpus to set of words, cleaning the corpus from delimiters and symbols, storing each two consecutive words in the corpus into database.

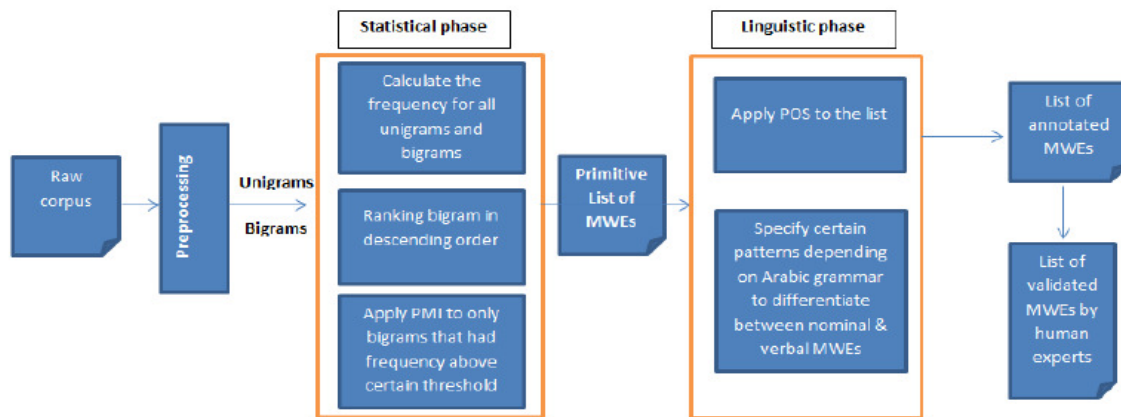


Fig.1. Architecture of the proposed hybrid framework for Extracting MWEs

3.2 Statistical phase

Association measures are inexpensive and language-independent means for discovering recurrent patterns, or habitual collocates. Association measures are defined by Pecina[14] as mathematical formulas that determine the strength of the association, or degree of connectedness, between two or more words based on their occurrences and co-occurrences in a text. The higher the connectedness between words, the better the chance they form a collocation. One of widely applied method is Point-wise Mutual Information (PMI) [9] that compares the co-occurrence probability of words given their joint distribution and given their individual (marginal) distributions under the assumption of independence. For two-word expressions, it is defined as:

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x, *)p(*, y)}$$

Where $p(x, y)$ is the maximum likelihood (ML) estimation of the joint probability (N is the corpus size):

$$p(x, y) = \frac{f(x, y)}{N}$$

And $p(x, *)$, $P(*, y)$ are estimations of marginal probabilities computed in the following manner:

$$p(x, *) = \frac{f(x, *)}{N} = \frac{\sum_y f(x, y)}{N}$$

And analogically for $P(*, y)$.

The following steps were applied during phase1 of the proposed model

1. Calculate the frequency of all the unigrams and bigrams in the corpus.
2. Calculate the PMI to all bigrams that have a frequency above certain threshold
3. Bigrams are ranked in descending order.

Here in this stage we have a list of MWEs with its PMI sorted in descending order.

3.3 Linguistics filtering of Arabic MWE

Extracting MWEs using statistical approach depends on the idea of occurrences and co-occurrences of two words would lead to generate patterns that may not be MWEs such as " ذِيَالٍ َ ُضِيُو " or " انْحكى حوَل ". Those bigrams repeated many times in the same corpus but are not considered a MWE. Thus, it is important to utilize linguistic rules to identify the correct MWEs

from the ranked list of MWEs generated by the previous statistical phase. These linguistic rules are illustrated in this subsection.

3.3.1 Selected Linguistic rules.

In order to be considered as a multi-word expression, a sequence of words should fulfill syntactic and semantic conditions. In fact, we can distinguish many types of MWEs [15] such as:

- Idioms (e.g. وَرَّانَعَهِي)
- Phrasal verbs (e.g. عَهِي دَّيَعَر)
- Verbs with particles (e.g. ع يَعْفُو)
- Compound nouns (e.g. زَاوَالَا جَزِيذَج)
- Collocations (e.g. يَعْزُوف مَّإِع)

Furthermore, a compound noun belongs to one of the following categories:

- Annexation compound noun (الاضافي انرزكية): an expression composed of an indefinite noun and one of the following elements:
 - A possessive pronoun (e.g. طياردَّ : his car),
 - Any simple or compound definite noun (e.g. عَهِي طيارج : the car of Ali),
 - An indefinite adjective compound noun (e.g. يُّ غ رجم طيارج : the car of a rich man).

The first component is called صَّافَا ن (first term of annexation) while the second is called صَّافَا ن (second term of annexation). The definiteness of the compound noun is equal to the definiteness of the second component.

- Adjective compound noun (انوصفي انرزكية): an expression composed of a noun (either simple or compound) which is called "عُوخ ي" (The modified word) and an adjective (ان عُد) having the same definiteness (e.g. يُّ غ رجم : a rich man). The gender of the two elements must be agreed.
- Substitution compound noun (انثذل انرزكية): an expression composed of a demonstrative pronoun and a definite noun (e.g. انظيارج دَّ, this car). Such expression is always definite.
- Prepositional compound noun: two nouns linked by a preposition (e.g. انحهاء ي وَع : a kind of sweet).
- Conjunctive compound noun: two nouns linked by a conjunction (e.g. وانفأرا نَقْطَر : the cat and the mouse).
- Compound nouns linked by composite relations: two or more linkers (prepositions and/or conjunctions) are used to link two nouns (e.g. حُ ط نحواني زَار الاطر : To persist for about one year).

Since the proposed framework is applied only for bigram, only linguistic rules for adjective compound noun and substitution compound noun are applied as shown in figure2.

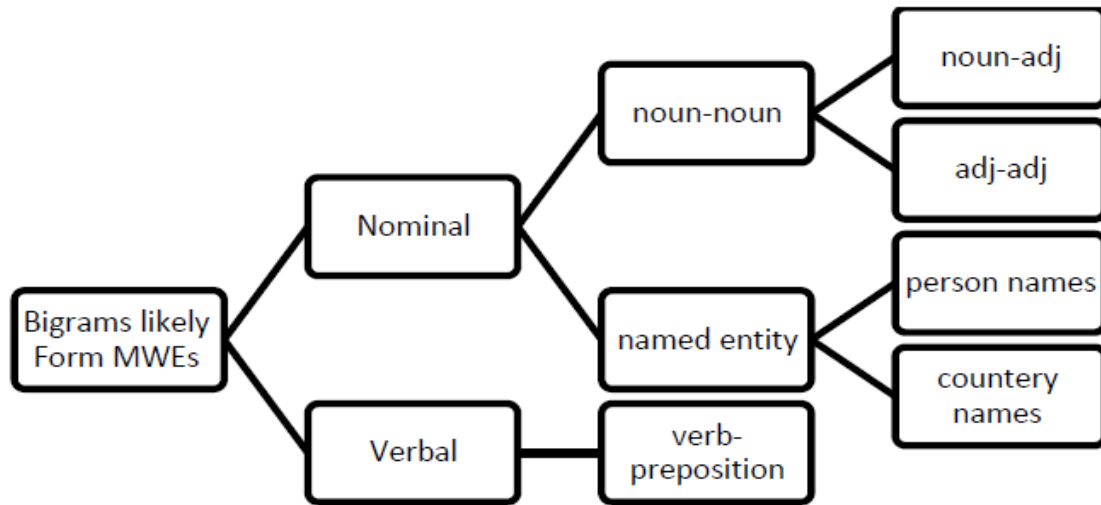


Fig. 2 Sample of used linguistic patterns

Furthermore, we also applied some linguistic rules of verbs such as verbs with particle that represents verb followed by preposition like "إني أدى", "في شارك", "في فشم", or "عهي يضي".

3.3.2 Filtering Identified pattern

As explained above, the list of ranked bigrams is applied to part of speech tagger (POS) in order to identify the type of the words (noun, verb, preposition or etc.). This framework uses the Stanford POS tagger -a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. Part-of-speech tags are assigned to each single word according to its role in the sentence. Traditional grammar classifies words based on eight parts of speech: the verb (VB), the noun (NN), the pronoun (PR+DT), the adjective (JJ), the adverb (RB), the preposition (IN), the conjunction (CC), and the interjection (UH)- <http://www.clips.ua.ac.be/pages/pattern>. Next, the linguistic rules illustrated in figure2 are applied to those tagged pattern to extract more meaningful pattern. . It is significant to mention that the main objective of applying linguistic rules after using statistical approach is to limit the scope of the MWE identification process where the experiment yields many bigrams that had a high frequency generated in the statistical phase list like "نعدّي". This bigram had a high frequency but did not represent actual MWEs so it was filtered in the linguistic phase according to the patterns specified in fig.2.

4. EXPERIMENT

Three different corpus were used in our experiment. The first one, archives from Omani newspaper Alwatan of the year 2004 [8]- <https://sites.google.com/site/mouradabbas9/corpora>. The size of the extracted corpus is about 10 millions terms which correspond to 9000 articles, distributed over six topics, in this case: Culture, religion, economy, local news, international news and sports. The second corpus is the Arabic Newswire Part 1 This publication contains the Arabic Newswire a Corpus, Linguistic Data Consortium (LDC) catalog number LDC2001T55 and ISBN

1-58563-190-6. The Arabic Newswire Corpus is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER-style SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from May 13, 1994 to December 20, 2000. There are 209 Mb of compressed data (869 Mb uncompressed) with approximately 383,872 documents containing 76 million tokens over approximately 666,094 unique words. The third one is Named Entity Corpus from Arabic Language Technology Center "ALTEC" <https://sites.google.com/site/mouradabbas9/corpora>.

4.1 Experiment setup

The following pre-processing steps have been applied for the corpus:

- Cleaning the corpus from punctuations and symbols.
- Splitting it to set of unigrams and bigrams.
- Storing all unigrams and bigrams into database.

4.2 Results of Experiment

The first experiment was applied in order to identify the value of threshold that should be used during phase1 (statistical phase). Thus, we change the frequency used in the statistical phase from 20,30,40 and 50 respectively in order to study the effect of changing the threshold on the accuracy of the result. As shown in figure 3, with decreasing the frequency during statistical phase, the number of candidate MWE increases. As explained earlier, statistical phase did not consider any linguistic features, it only depend on the degree of connectedness between two or more word. Accordingly, increasing the number of obtained MWE from phase1 would lead to provide more set of candidate MWE to be used during linguistic phase and avoid missing any candidate MWE from corpus. However, linguistic phase plays a significant role in enhancing the final results as the number of final MWE dramatically decreased to almost half in all cases as shown in figure3.

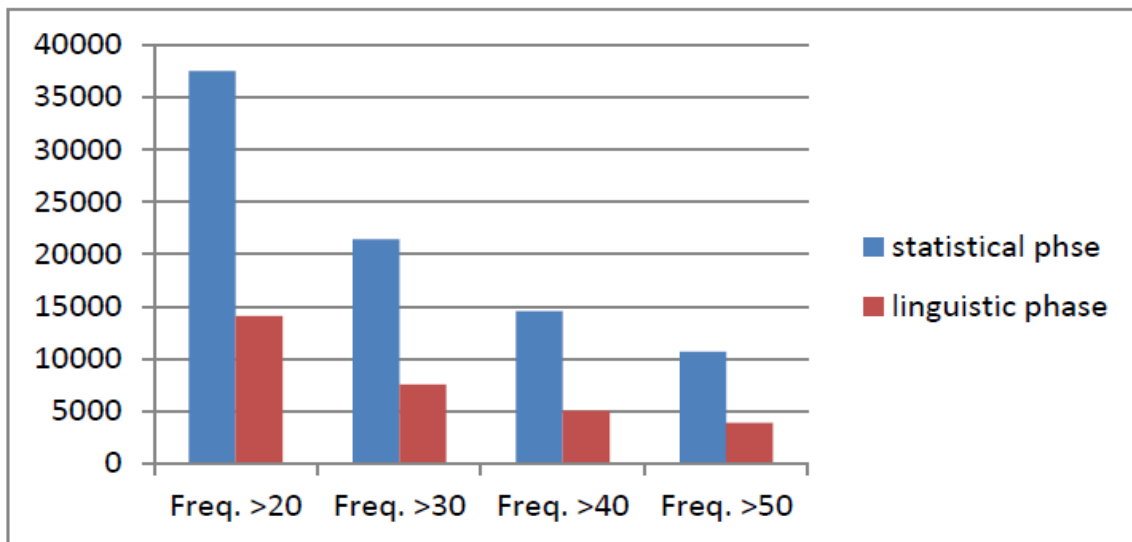


Fig.3 The effect of frequency change on number of MWEs in each phase

The aim of second experiment is to identify the number of candidate MWE after applying each phase and compare the results with the proposed framework by Attia [9]. Therefore, in this experiment during the statistical phase we set the frequency to 50 to be able to compare it with Attia (although this is not matching with the result obtained in the first experiments that recommend setting the frequency to 20). The results summarized in table1 shows that, the number of detected MWE after applying linguistic phase decreased to one-third and the final result outperformed the results obtained when applying statistical phase. According to table1, our proposed framework generate more final MWEs due to applying more linguistic rule (such as those related to verbs) that those proposed by Attia[9] which decrease the possibility of omitting significant MWEs patterns that are not of type (noun-noun, noun- adjective).

	Our corpus	Attia's corpus
Total number of bi-grams	98,070,263	875,920,195
After grouping distinct bigrams	3,588,041	134,411,475
After applying PMI to bigrams with freq. >50	10,704	1,497,214
Selecting only patterns that Attia used	3714	217,630
Ratio between the number of MWEs generated from linguistic phase to statistical phase	35%	15%
Selecting our pattern using POS	3,831	

Table 1. Comparison between the number of generated MWE using proposed model and Attia

Next, ground truth is used to identify the correct set of final list of MWEs. Therefore, we present the final list generated from proposed model as well as the list generated when applying Attia model to domain expert to validate the correctness of identified MWEs. Human experts have annotated the list obtained from both models in order to compute the precision of them as shown in table 2.

First word	Second word	MWE(1)/NON-MWEs (0)
DTNN الامم/	DTJJ المتحدة/	1
DTNNS الولايات/	DTJJ المتحدة/	1
NNP فرانس/	NNP برس/	1
NN وزير/	DTN الخارجية/	1
NN مجلس/	DTNN الامن/	1
DTNN اليوم/	DTNN الخميس/	0
DTNN اليوم/	DTNN الأخير/	1
NN اطلاق/	DTNN النار/	1
VBD طلع/	IN على/	1
VBD كان/	IN في/	0
NNP لاطلاق/	DTNN النار/	0
NNP اسحق/	NNP رابين/	1
NNP لوس/	NNP انجليس/	1
NNP لحزب/	NNP الله/	1

Table2 sample of the human expert annotation to the final list of MWEs

Finally, precision is calculated in order to compare the accuracy of our proposed with Attia. According to figure4, the value of precision increase to 67% when applying of the whole model compared to Attia (about 34 %). It is significant to mention that applying both nominal and verbal linguistic phase rule during increase the value of precision by 3%. This indicates that the verbal MWEs represent a smaller number of MWEs in comparison with nominal MWEs in Arabic language.

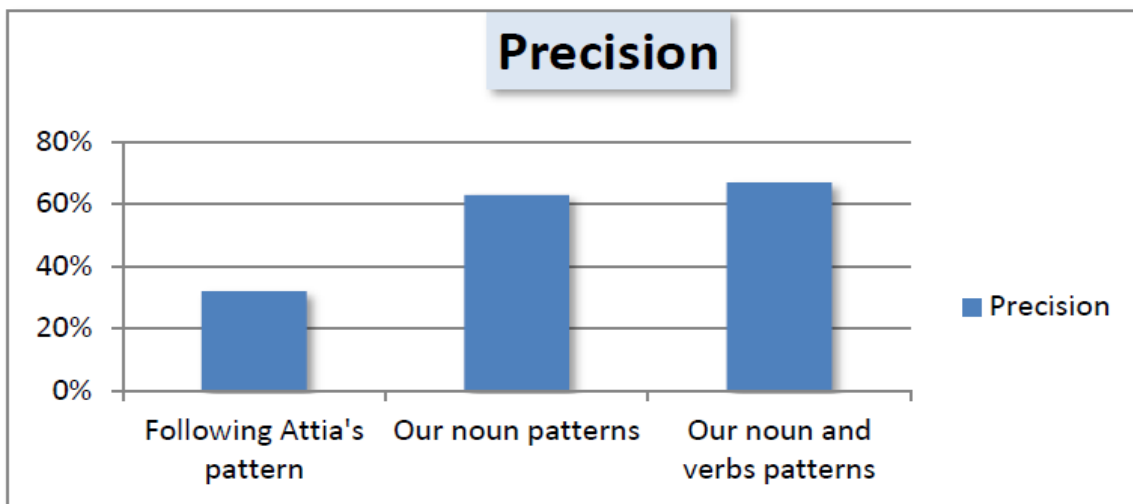


Figure4: Comparison between results of precision when applying Attia and our proposed framework.

5. CONCLUSION

The process of extracting MWEs is a very complicated task to be solved by one single solution. In this paper we develop a framework for extracting Arabic MWEs using hybrid approach that combine the statistical approach with the linguistic rules and the results obtained validated by human experts and the precision differed according to the threshold determined in statistical phase. We find that the more the threshold that set in the statistical phase is low the more we get greater number of MWEs, the statistical approach measures the connectedness of each two consecutive words in the corpus regardless these two words are MWEs or not so the linguistic approach increases the accuracy of the generated MWEs list from the statistical phase by filtering undetermined patterns, after applying our experiment into different data sources we find that the ratio between nominal MWEs and verbal MWEs in the list generated from phase1 and phase2 represents 97:3 respectively.

REFERENCES

- [1] O. Kraif, (2003) "Repérage de traduction et commutation interlingue :Intérêt et méthodes", Traitement Automatique des Langues Naturelles TALN 2003, Batz-sur-Mer, France, June 11-14, 2003.
- [2] V. Malaisé, (2005) "Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels", doctoral thesis, University of Paris 7 – Denis Diderot, 2005.
- [3] K.W. Church, W. Gale, P. Hanks, and D. Hindle, (1991) "Using statistics in lexical analysis". In *Lexical Acquisition, Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, Michigan, USA: Zernik Uri ed., London, Lawrence Erlbaum Associates, 1991, pp.115-164.
- [4] T. Dunning, (1994) "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, vol. 19(1), pp. 61-74, 1994.
- [5] H. Nakagawa, T. Mori, and H. Yumoto, (2003) "Term Extraction Based on Occurrence and Concatenation Frequency", *Journal of Natural Language Processing*, vol. 10 (1), pp.27-45, 2003.
- [6] B. Daille, (1994) "Approchemixte pour l'extraction de terminologie : statistiquelexicale et filtreslinguistiques", doctoral thesis, University of Paris 7, 1994.
- [7] S. Boulaknadel, B. Daille and D. Aboutajdine, (2008) "A multi-word term extraction program for Arabic language", the 6th international Conference on Language Resources and Evaluation LREC 2008, Marrakech, Morocco, 28-30 May 2008, pp. 1485-1488.
- [8] Abbas, M., Smaili, K., & Berkani, D. (2010) "Tr-classifier and knn evaluation for topic identification tasks", *The International Journal on Information and Communication Technologies (IJICT)*, 3(3), 65-74.
- [9] Attia, M., Antonio Toral, Lamia Tounsi, PavelPecina and Josef van Genabith,(2010) "Automatic Extraction of Arabic Multiword Expressions", In: *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pp: 18–26, Beijing, China. 2010.
- [10] Z. Harris, (1991) "Theory of Language and Information: A Mathematical Approach", Oxford & New York: Clarendon Press, 1991.
- [11] Traboulsi, H,(2009) "Arabic named entity extraction: A local grammar-based approach", In: *Proceedings of the International Multiconference on Computer Science and Information Technology*, vol. 4, pp. 139–143 (2009) .
- [12] Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger, (2002) "Multiword Expressions: A Pain in the Neck for NLP" In the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), volume 2276 of *Lecture Notes in Computer Science*, pp. 1.15, London, UK. Springer-Verlag.

- [13] Ramisch, Carlos, Paulo Schreiner, Marco Idiart and Aline Villavicencio, (2008), "An Evaluation of Methods for the Extraction of Multiword Expressions", In the Workshop on Multiword Expressions, the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 50-53. Marrakech, Morocco.
- [14] Pecina, Pavel, (2010) "Lexical association measures and collocation extraction", In Language Resources and Evaluation (2010), 44:137-158.
- [15] Bounhas, I. and Y. Slimani, (2009) "A hybrid approach for Arabic multi-word term extraction", Proceeding of the International Conference on NLP-KE 2009, Department of Computer Science, University of Tunis, Sept. 24-27, Tunis, Tunisia, pp: 1-8. DOI: 10.1109/NLPKE.2009.5313728.
- [16] Helena M. Caseli, Carlos Ramisch, Maria G. V. Nunes, and Aline Villavicencio, (2009) "Alignment-based extraction of multiword expressions", Language resources and evaluation 44 (1-2), 59-77.

INTENTIONAL BLANK

ANALYSIS OF COMPUTATIONAL COMPLEXITY FOR HT-BASED FINGERPRINT ALIGNMENT ALGORITHMS ON JAVA CARD ENVIRONMENT

Cynthia S. Mlambo¹, Meshack B. Shabalala¹,
Fulufhelo V. Nelwamondo^{1,2}

¹ Council for Scientific and Industrial Research, Pretoria, South Africa,
smlambo@csir.co.za, mshabalala@csir.co.za

² Department of Engineering, University of Johannesburg
fnelwamondo@csir.co.za

ABSTRACT

In this paper, implementations of three Hough Transform based fingerprint alignment algorithms are analyzed with respect to time complexity on Java Card environment. Three algorithms are: Local Match Based Approach (LMBA), Discretized Rotation Based Approach (DRBA), and All Possible to Match Based Approach (APMBA). The aim of this paper is to present the complexity and implementations of existing work of one of the mostly used method of fingerprint alignment, in order that the complexity can be simplified or find the best algorithm with efficient complexity and implementation that can be easily implemented on Java Card environment for match on card. Efficiency involves the accuracy of the implementation, time taken to perform fingerprint alignment, memory required by the implementation and instruction operations required and used.

KEYWORDS

Fingerprint Alignment, Java Card, Hough Transform, Smart Cards, Time Complexity.

1. INTRODUCTION

The Java Card Environment have a limited instruction sets, unlike other languages. The challenge is that Smart Card applications are increasing in the market as one of the mostly used technologies. Currently what is happening in the industry is that most of the applications are shifting from computer based and large applications into small portable applications that can function on smart Cards. The basic recent application is the use of identification and verification of an individual using the Smart Card; this involves fingerprint based recognition systems.

Fingerprint alignment is a process of superimposing two different features of fingerprints that are captured at different instances [1]. This process is important in identifying or verifying if two fingerprint features captured at different instances are from the same finger. This is because there is always rotation and translation of a finger during the process of capturing fingerprint features [1]-[3]. One of the mostly used methods of performing fingerprint alignment is based on Hough Transform. The Hough Transform (HT) based methods accumulate votes for the most occurring rotation ($\Delta\theta$) and translation (Δx , Δy) between two sets of fingerprint features that are to be

matched. Since Java Card applications are one of the intelligent technologies that are increasingly used these days [4]-[5]. However, The Java Card Environment has limited computing resources, such as; instruction sets, memory space and power sources [6]. There is a need of analysis existing applications on how is their performance on Java Card technologies. The challenge is that when implementing algorithms on Java Card, it needs to be modified so that it can meet the specifications of the Java Card environment [6]. Therefore, in this paper, four implementations of Hough Transform based Fingerprint Alignment algorithms are analysed on Java Card. These algorithms are: Local Match Based Approach (LMBA), Discretized Rotation Based Approach (DRBA), and All Possible to Match Based Approach (APMBA).

The aim of this paper is to present the complexity and implementations of existing work of one of the mostly used method of fingerprint alignment. In addition, two questions are answered in this research, stated as: How much changes or modifications are required? What are the effects of those modifications on time complexity, memory and performance of the algorithm? So that the complexity can be simplified or find the best algorithm with efficient complexity and implementation that can be easily implemented on smart cards. Efficiency involves the accuracy of the implementation, time taken to perform fingerprint alignment, memory required by the implementation and instruction operations required and or performed.

Fingerprint features used in this paper are minutiae points because they require less memory as they are represented as points [7]. Minutiae points are where ridges in minutiae points ends and split, alternatively called ridge ending and ridge bifurcation, respectively. Each minutia is presented in three coordinate, x-coordinate, y-coordinate and the orientation of the ridge.

This paper is organised as follows, firstly in Section II is a brief description of the HT based fingerprint alignment algorithms. Section III is the comparison of studied algorithms. Finally, section IV is the conclusion.

2. DEFINING PROBLEM STATEMENT

In this section the challenge on implementing HT-based fingerprint alignment algorithms is explained from the specifications of the Java Card environment [8]. In general definition, Java Card technology enables programs written in the Java programming language to run on smart cards.

The Java instruction sets are too large to fit on a resource constrained device such as a smart card [9]. That is why the Java Card environment consist limited instruction sets and card commands, which affect programming style and implementation of Java or low-level programming algorithms. Features that are common in Java language and Java Card are:

- Small primitive data types (Boolean, byte, short, int),
- One-dimensional arrays,
- Packages, classes, interfaces,
- Object-oriented features, (inheritance, virtual methods, overloading, dynamic object creation, access scope, binding rules),
- Exceptions [8].

However, the Java Card does not support some features [10]. As a result, the implementation of the algorithms that requires the following features needs to be changed. Unsupported features are:

- Large primitive data types (long, double, float),
- Characters and strings,
- Multidimensional arrays,
- Dynamic class loading,

- Security manager,
- Threads,
- Object cloning,
- Garbage collection,
- Object serialization [8].

Having discussed these Java Card features, it is common for most HT-based algorithms to require for example, multidimensional arrays [11] – [15]. Therefore, the research presented in analyses each algorithm based on how many does it require.

3. HOUGH TRANSFORM-BASED METHODS

In this section all three analysed algorithms are explained in details with their functionality on how they determine alignment parameters (AP) for translation and rotation. Two sets of fingerprints are taken as inputs to the algorithms, input (I) and template (T) set. Where I is a set of minutiae points from the input minutiae and T is a set of minutiae points from the database which was captured and stored.

3.1. All Possible Matching Based Approach

The first stage in this approach is performed by considering all points as possible matches [16], [17]. Therefore, for each minutiae point in input I and template T , all points are paired and alignment parameters are computed. For each set of computed alignment parameters, a vote is added to the accumulator array. In the last stage, after all minutiae points are paired, the most highly voted alignment parameters are determined and deemed as the best parameters for alignment of I and T [16]. This algorithm is explained in Algorithm 1.

The motive on this approach is that since the aim of alignment process is to estimate corresponding pairs. Minutiae points are not checked if are corresponding but all points are considered, with the idea of the HT, state: All corresponding points between set I and T will accumulate similar transformation into accumulator array A [17].

Pros:

- Simplicity and ease of implementation
- The use of one-dimensional array makes easy implementation
- General complexity of $O(n*m)$

Cons:

- Very inefficient for large number of minutiae points when performing voting process.

Algorithm 1: An Alignment Algorithm for Improved HT-Based Approach

Input: Sets of minutiae points from I and T , and size of an image $[MaxSizeX, MaxSizeY]$ and maximum resolution $MaxSize\theta$.

Output: Set of $(\Delta_x, \Delta_y, \Delta_\theta)$.

Initialize Hough space into three 3D array

$ASize = \sqrt{NumberOfT * NumberOfI}$

$BinSizeX = MaxSizeY / ASize$

$BinSizeY = MaxSizeX / ASize$

$BinSize\theta = MaxSize\theta / ASize$

Discretize values for $(\Delta_x, \Delta_y, \Delta_\theta)$ by initializing

$A_x[ASize], A_y[ASize], A_\theta[ASize]$

for each minutiae m_t in T do

for each minutiae m_i in I do

 Compute direction difference θ^+ between (θ_i) and (θ_t)

$\theta^+ = \min(|\theta_i - \theta_t|, 360 - |\theta_i - \theta_t|)$

 Compute translation parameters

$$\begin{bmatrix} \Delta_x^+ \\ \Delta_y^+ \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \begin{bmatrix} \cos \theta^+ & -\sin \theta^+ \\ \sin \theta^+ & \cos \theta^+ \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

 To increase accuracy and to overcome distortions induced during finger image acquisition, the exact values of $\Delta_x^+, \Delta_y^+, \theta^+$ are added into A , and the average value is computed

 The implementation in Java allows Increment

$(\Delta_x, \Delta_y, \Delta_\theta)$ in the accumulator array by adding evidence into accumulator A of the corresponding bin

$A_x[\Delta_x^+] = (A_x[\Delta_x^+]_N + 1) \cdot (A_x[\Delta_x^+]_{\Delta_x} + \Delta_x^+) / 2$

$A_y[\Delta_y^+] = (A_y[\Delta_y^+]_N + 1) \cdot (A_y[\Delta_y^+]_{\Delta_y} + \Delta_y^+) / 2$

$A_\theta[\theta^+] = (A_\theta[\theta^+]_N + 1) \cdot (A_\theta[\theta^+]_{\theta^+} + \theta^+) / 2$

Find indexes of maximum A , which are the most voted alignment parameters $[\Delta_x, \Delta_y, \Delta_\theta] = \max(A)$

3.2. Discretized Rotation Based Approach

In the DRBA approach, it is common to consider all given points from I and T as possible corresponding points [11]. In addition, by checking if the direction difference of minutiae orientation is less than a defined threshold [12]. The second stage is to estimate AP from estimated corresponding pairs. The rotation angle is taken from the discretized data and used to compute AP . Translation parameters are computed using the affine transformation with the rotation angle from discretization data. At the third stage the accumulator array A is required to store all possible AP . The bin size is used to specify the step size in A and it is used when voting for the nearest bins of the current estimated AP . During the voting procedure it is general to cast the votes on the nearest bins, and the bin sizes are experimentally defined by considering different values from too small to large amounts [18]. The number of votes is accumulated by adding a vote for each computed parameters, shown in equation (1).

$$A[\Delta_x^+, \Delta_y^+, \theta^+] = A[\Delta_x^+, \Delta_y^+, \theta^+] + 1 \quad (1)$$

It is common in both approaches to define the accumulator array as a 3D array for rotation angles, and translations along the x and the y axis. The last step is to find the best alignment set, which can be one set or N sets of indexes of A with the largest votes. The implementation is shown in Algorithm 2.

Algorithm 2: An Alignment Algorithm for Discretized Rotation Based Approach

Input: Sets of minutiae points from I and T .
Output: Set of $(\Delta_x, \Delta_y, \Delta_\theta)$.
Set direction tolerance θ_0 , and initialize Hough space into 3D array
Discretize values for $(\Delta_x, \Delta_y, \Delta_\theta)$ by initializing $A[\Delta_x, \Delta_y, \Delta_\theta]$

```

for each minutiae  $m_t$  in  $T$  do
  for each minutiae  $m_i$  in  $I$  do
    for each rotation angle  $\theta^+$  in  $\Delta_\theta$  do
      Compute direction difference  $dd$  between  $(\theta_i + \theta^+)$ 
      and  $(\theta_t)$ 
       $dd = \min(|(\theta_i + \theta^+) - \theta_t|, 360 - |(\theta_i + \theta^+) - \theta_t|)$ 
      if  $dd < \theta_0$  then
        Compute translation parameters
        
$$\begin{bmatrix} \Delta_x^+ \\ \Delta_y^+ \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \begin{bmatrix} \cos \theta^+ & -\sin \theta^+ \\ \sin \theta^+ & \cos \theta^+ \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

        To increase accuracy and to overcome distortions
        induced during finger image acquisition,
         $\Delta_x^+, \Delta_y^+, \theta^+$  are quantized to the nearest bins
        for  $\Delta_\theta = (\theta^+ - Err_\theta) \rightarrow (\theta^+ + Err_\theta)$  do
          for  $\Delta_x = (\Delta_x^+ - Err_x) \rightarrow (\Delta_x^+ + Err_x)$  do
            for  $\Delta_y = (\Delta_y^+ - Err_y) \rightarrow (\Delta_y^+ + Err_y)$  do
              Increment  $(\Delta_x, \Delta_y, \Delta_\theta)$  in the
              accumulator array by adding evidence
              into accumulator  $A$ 
               $A[\Delta_x^+, \Delta_y^+, \theta^+] =$ 
               $A[\Delta_x^+, \Delta_y^+, \theta^+] + 1$ 
    Find indexes of maximum  $A$ , which are the most voted alignment
    parameters  $[\Delta_x, \Delta_y, \Delta_\theta] = \max(A)$ 

```

Pros:

- Overcomes distortion when voting for nearest bins in all sets of alignment parameters.

Cons:

- General complexity of $O(m*n*\log(m*n))$.
- Very inefficient for large number of minutiae points.
- Performs lots of operations.

3.3. Local Match Based Approach

In the LMBA approach, the first stage in this case is performed by using some methods to determine matching points, for example: by finding pair of points with similar Euclidean distance from their locations; or by first determining corresponding triangles between I and T [14] [15]; or by using similar triangles from Delaunay triangulation [15], and then, estimate matching points from corresponding triangles. The second stage is performed by using the affine transformation with the computed rotation angle to compute AP . In the third stage it is common to define different bins of the accumulator array, e.g. starting from a large size of bins to the small size of bins to find the finer results of AP . The number of votes is accumulated by adding a number of aligned points determined after aligning points using each set of parameters, as shown in equation (2).

$$A[\Delta_x^+, \Delta_y^+, \theta^+] = A[\Delta_x^+, \Delta_y^+, \theta^+] + N \quad (2)$$

At the end of this approach, a set of AP with the highest number of votes is deemed as the one that represent the best transformation of tested sets of minutiae points [19]. The implementation and explanation of this algorithm is in Algorithm 3 and Algorithm 4.

Algorithm 3: An Alignment Algorithm for Local Match Based Approach

Input: Sets of minutiae points from I and T .
Output: Set of $(\Delta_x, \Delta_y, \Delta_\theta)$.
Define $\theta_{length}, X_{length}, Y_{length}$ and number of matching levels e.g. $levels = 3$
Define required thresholds, l_0 for lengths of the triangles, and a_0 for the largest angle. Define possible alignment parameters for $(\Delta_x, \Delta_y, \Delta_\theta)$
Determine Delaunay triangles DT_I and DT_T for I and T , respectively, and compute $invar_features[]$
Compute alignment parameters as follows.
if $levels! = 0$ **and** $alignment_score < set_tolerance$ **then**
 Set the unit size for $A[\Delta_x, \Delta_y, \Delta_\theta]$
 $\Delta_{\theta_size} = \theta_{length} / 2^{levels-1}$
 $\Delta_{x_size} = X_{length} / 2^{levels-1}$
 $\Delta_{y_size} = Y_{length} / 2^{levels-1}$
 Initialize $A[\Delta_{x_size}, \Delta_{y_size}, \Delta_{\theta_size}]$
 Set the bin size of rotation and translation
 $\Delta_{binsize} = 2^{levels-1}$
 Set $(\theta_{length}, X_{length}, Y_{length}) = (\Delta_{x_size}, \Delta_{y_size}, \Delta_{\theta_size})$
 for each triangle in DT_T **do**
 for each triangle in DT_I **do**
 Determine a triangle in DT_I that correspond with a current triangle from DT_T by For each corresponding triangle, compute alignment parameters using Algorithm 4
 Find index of maximum A which is the most voted alignment parameters
 $[\Delta_x, \Delta_y, \Delta_\theta] = max(A)$

Pros:

- Alignment results are accurate.

Cons:

- Performs lots of computations.
- General complexity of $O(m^2 * n^2)$.
- Requires lots of multidimensional arrays.
- Very inefficient for large number of minutiae points.

4. COMPARATIVE ANALYSIS

Table I summarizes the performance and implementation of presented algorithms. Time complexity was calculated from the implementation of each algorithm. The LMBA shows a time complexity with high number of operations which leads to a low performance. This is because there are lots of computations involved when determining corresponding minutiae points from triangles. Another challenge is the memory required to process corresponding triangles which result in that the LMBA required more use of arrays and functions that are not supported in Java Card. In addition, in the LMBA, alignment process is performed for each computed set of alignment parameters. Therefore, computation time and operations are required to perform alignment.

The computational complexity of the DRBA is caused by involving the discretized rotations because this process requires repetitions of testing if the orientation differences for each pair of minutiae points are within the threshold. The AMPBA requires operations when computing the

average values of alignment parameters. However, the operations require most of supported instructions by the Java Card environment.

In Figure 1, the complexity in terms of instructions executed and operations by each algorithm is presented with respect to number of minutiae points or fingerprint features that need to be aligned.

Table I. Computational complexity of algorithms

Parameter	DRBA	LMBA	AMPBA
Time Complexity	$O(m*n*\log(m*n))$	$O(m^2*n^2)$	$O(m*n)$
Space Complexity	$O(n)$	$O(n)$	$O(1)$
Use of Arrays	60%	90%	40%
Unsupported Functions	30%	60%	30%

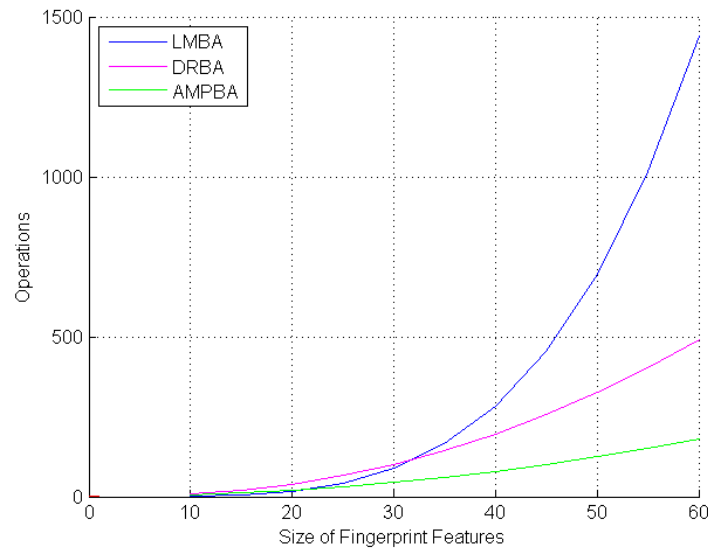


Figure 1. Operations performed with respect to fingerprint feature

5. CONCLUSIONS

From the above analysis it can be said that: the APMBA required less computational complexity compared to other algorithms although it increases as the number of fingerprint features increases but the time taken is less than that of the LMBA and DRBA. In addition, the implementation of the APMBA with one dimension arrays gives it advantages of simplified implementation. The LMBA requires more operations when the number of minutiae points increases, as a result is the slowest algorithm.

The future work is to study the implementation of promising matching algorithms on Java match on card to identify the one with least complexity in terms of time, memory and accuracy.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Department of Science and Technology, for funding this research.

REFERENCES

- [1] H. Pompei and D. A. Russell, (2012), Advances in fingerprint analysis. *Angewandte Chemie International Edition* 51, (15):3524–3531.
- [2] V. Krithika and V. S. Kumar, (2011), Fingerprint identification: A brief literary review.
- [3] J. Bringer H. Chabanne T. Chouta J. Danger M. Favre B. Mael, Y. Bocktaels and T. Graba, (2013), Studying potential side channel leakages on an embedded biometric comparison system. *Database* 4(5(7)).
- [4] CardLogix Corporation, Smart Card Standards, (2010), [http:// www.smartcardbasics.com/smart-card-standards.html](http://www.smartcardbasics.com/smart-card-standards.html), (Last visited 08/08/14).
- [5] C. S. Mlambo, F.V. Nelwamondo, M.E. Mathekga, (2014), Comparison of effective Hough Transform-based fingerprint alignment approaches, *International Symposium on Biometrics and Security Technologies*, IEEE.(in press)
- [6] ORACLE, “Java Card Technology Documentation” <http://docs.oracle.com/javame/javacard/javacard.html>, 2012. (Last visited 20/11/14).
- [7] Precise Biometrics, (2013), “Match on Card”, <http://www.matchoncard.com/what-is-moc/smart-cards-and-fingerprint-recognition/>, (Last visited 09/10/2014).
- [8] ORACLE Inc. (2010), Java Card™ 3 Platform, Application Programming Notes.
- [9] Infineon Ltd, (2014) “National ID”, <http://www.infineon.com/cms/en/product/smart-card-ic>, 2014, (Last accessed 18/10/14).
- [10] CardLogix Corporation, (2010) “Smart Card Standards”, [http:// www.smartcardbasics.com/smart-card-standards.html](http://www.smartcardbasics.com/smart-card-standards.html), 2010, (Last visited 08/10/14).
- [11] A. Paulino, J. Feng and A. Jain, (2013), Latent Fingerprint Matching Using Descriptor-Based Hough Transform, *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 31-45.
- [12] R. Zhou, D. Zhong, and J. Han, (2013), Fingerprint Identification Using SIFT-Based Minutia Descriptors and Improved All Descriptor-Pair Matching, *Sensors*, ISSN: 1424-8220.
- [13] F. Chen, X. Huang, and J. Zhou, (2013), Hierarchical Minutiae Matching for fingerprint and Palm print Identification, *IEEE Transactions on Image Processing: a publication of the IEEE Signal Processing Society*, vol. 22, no. 12, pp. 4964-497.
- [14] G. Bebis, T. Deaconu, and M. Georgiopoulos. (1999), Fingerprint identification using Delaunay triangulation. *Information Intelligence and Systems*, 1999. Proceedings. 1999 International Conference on. IEEE, pp. 452–459.
- [15] P. R. Mendes, A. C. Junior, and D. Menotti , (2010), A Complete System for Fingerprint Authentication using Delaunay Triangulation, *Reconhecimento de Padroes, DECOM-UFOP*, pp. 1-7.
- [16] C.S. Mlambo, F.V. Nelwamondo, and M.E. Mathekga, (2014), “An improved Hough transform-based fingerprint alignment approach”, *International Image Processing, Applications and Systems Conference, IPAS'14, IEEE*, (Accepted.).
- [17] C.S. Mlambo, M. Shabalala, M.E. Mathekga, and F.V. Nelwamondo, (2014), Application of Hough transform-based fingerprint alignment on match on smart cards. *International Conference on Cyber Warfare and Security ICCWS, (ICCWS-2015)*. Accepted.
- [18] T. Uz, G. Bebis, A. Erol and S. Prabhakar, (2009), Minutiae-based Template Synthesis and Matching for Fingerprint Authentication, *Computer Vision and Image Understanding*, vol. 113(9), pp. 979-992.
- [19] A. Gheibi and A. Mohades, (2013), Stable Geometric Fingerprint Matching, *IET Computer Vision Journal*.
- [20] V. Gupta and R. Singh, (2012), Image processing and computer vision. *Fingerprint Recognition CS676*.
- [21] A.C. Lomte, and S.B. Nikam, (2013), “Biometric fingerprint authentication by minutiae extraction using USB token system”, *International Journal Computer Technology and Applications*, Vol. 4, No. 2, pp. 187-191.

- [22] F. Benhammadi, and K. B. Beghdad, (2013), “Embedded Fingerprint Matching on Smart Card”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 27, No. 02.

AUTHORS

Cynthia S. Mlambo is currently pursuing the Masters in Electrical Engineering at the University of Johannesburg. She holds an Honours Degree in Computer Engineering from the University of KwaZulu-Natal. Her areas of interest include image processing, pattern recognition and Smart ID Cards in biometrics.

Meshack B. Shabalala is a Biometric and Smart Card Researcher at the Council for Scientific and Industrial Research (CSIR), South Africa. He holds a Honours degree in Electrical Engineering from the University of the Witwatersrand.

Fulufhelo V. Nelwamondo is a Competency Area Manager for Information Security at the Council for Scientific and Industrial Research (CSIR), South Africa. He holds a PhD in Electrical Engineering from the University of the Witwatersrand and is a visiting professor of Electrical Engineering at the University of Johannesburg

INTENTIONAL BLANK

MULTIPLE USER INTERFACES AND CROSS-PLATFORM USER EXPERIENCE: THEORETICAL FOUNDATIONS

Khalid Majrashi¹, Margaret Hamilton and Alexandra L. Uitdenbogerd²

¹School of Computer Science and Information Technology,
RMIT University, Australia Institute of Public Administration and
Ministry of Higher Education, KSA
majrashik@ipa.edu.sa

²School of Computer Science and Information Technology,
RMIT University, Australia
margaret.hamilton@rmit.edu.au, alexandra.uitdenbogerd@rmit.edu.au

ABSTRACT

Evaluating the user experience of cross-platform interactive systems has become a research issue of increasing importance. There is a lack of clear concepts and definitions for testing, evaluating or even teaching cross-platform user experience. In this paper, we review the actual meanings and interpretations of different concepts in the field of Human-Computer Interaction (HCI) relevant to cross-platform service usage. We also investigate the traditional definitions of usability and user experience before extending them to develop precise definitions for cross-platform usability and user experience. Our paper builds on existing theories to establish the theoretical foundations that can help us better conceptualise cross-platform user experience evaluation.

KEYWORDS

Cross-platform, User Experience, Usability, Multiple User Interfaces

1. INTRODUCTION

Nowadays, end-users can interact with a service and information using different types of computing platforms including traditional office desktops, smart TVs, tablets, and mobile phones. This allows users to migrate their tasks from one user interface to another across devices or platforms. For example, a user can search for a restaurant from specific service, and then switch to the service image from their mobile phone to find the restaurant contact information, and then might transit to use a tablet to write a review about the restaurant. This brings a new user experience theme in which a user interacts with Multiple User Interfaces (MUIs) to achieve goals horizontally (across platforms). This type of MUI access is different from traditional user experience involving interaction with a single user interface (vertical interaction) [1]. There are new aspects in MUI interaction, including, switching a process from one user interface to another, migrating knowledge and tasks from one user interface to another. Despite the increased use of MUIs, and the corresponding increase in the need for cross-platform user experience evaluation,

there is a lack of explanations, definitions, and discussions of concepts in the context of cross-platform user experience.

In this paper, we review and explain different concepts related to cross-platform service, its characteristics, as well as relevant concepts in HCI. We follow this review by presenting the definitions of traditional forms of usability and user experience, and exploring the differences between them. This is to eliminate possible confusion between the two terms before defining them in the context of cross-platform interaction. Then, we provide comprehensive definitions that explain the concepts for both usability and user experience in the context of cross-platform service.

2. CROSS-PLATFORM SERVICE

In this section, we provide an overview of cross-platform service, including, definitions of the terms service and cross-platform usage, approaches for connecting a service, and configuration of cross-platform services.

2.1. What is a Service?

A service refers to software and hardware in which one or more services can be used to support a business's needs and functions. There are two primary types of services: atomic and composite [2, 3]. An atomic service is a self-contained function that does not require the use of other services. A composite service is an assembly of atomic or other composite services that may require the use of another service contained within the same composite service.

2.2. What is a Cross-Platform?

The term cross-platform can be used to characterise different entities in computer science [4]. For example, hardware devices, such as computer monitors, can be described as cross-platform as they can work with any operating system. Similarly, programming languages, such as C, can be described as cross-platform as they can be used to write software for use on any operating system. In addition, the term can be used to refer to software that can operate on more than one platform. For the purpose of this paper, we use the term cross-platform to refer to a service that can be accessed and used on two or more computing platforms.

2.3. Connection of Services

Web services provide the technologies for connecting services together. For cross-platform services, a web service can be defined as a system, which can be designed to support interoperable application-to-application communication over a network [5]. Interoperability can refer to both syntactic interoperability, and semantic interoperability [6, 7]. Syntactic interoperability depends on specified data formats and communication protocols to ensure communication and data exchange between heterogeneous software applications. With syntactic interoperability, there is no guarantee of consistent interpretations of exchanged data from one application to another. Semantic interoperability refers to the ability of various services across platforms to interpret the exchanged information meaningfully and accurately. There are multiple technologies of Web services for connection services, including the use of SOAP, WSDL, UDDI, REST, XML, and JSON, which are explained briefly as follows [8]:

1. Simple Object Access Protocol (SOAP) is a protocol for enabling communication between applications.

2. Web Service Description Languages (WSDL) is used to define web service interfaces, data and message types, and protocol mapping.
3. Universal, Description, Discovery, and Integration (UDDI) is a web service registry and discovery mechanism, used for sorting business information, and retrieving pointers to web service interface.
4. Extendable Markup Language (XML) provides a language for defining data and processing it.
5. Representational State Transfer (REST) is an alternative to SOAP that is developed on a set of principles describing how networked resources are defined and addressed.
6. JSON (JavaScript Object Notation) is an alternative to XML that uses name/value pairs instead of tags as used in XML.

2.4. Configuration of Cross-Platform Services

A cross-platform service aims to provide pervasive and synergistic support for human activities in different contexts of use. Feiner [9] presented the concept of hybrid user interfaces in which multiple heterogeneous displays and interaction devices are used synergistically to benefit from the features for each of them. Services across devices can be configured based on different user and/ or business needs, considering different device constraints and capabilities. Configuration of a cross-platform service refers to the manner in which devices are organised and the service is delivered across these devices [10]. Before discussing device organisation and service delivery, we need to clarify the concept of synergistic specificity, which is associated with different methods of configuration.

2.4.1. Synergistic Specificity

Systems across multiple platforms can reach their complete planned potential advantages when their components are used synergistically. Synergistic specificity is a term introduced by Schilling [11] to describe “the degree to which a system achieves greater functionality by its components being specific to one another” within a specific configuration. Systems with high synergistic specificity may be able to support functionality and user experiences better than segmental systems. These days many systems have core functionalities across platforms that rely on optimal coordination between their components to work with each other. These systems can lose their intended performance or become completely paralysed if their cross-platform components are used in isolation [12]. An example of a system with a high degree of synergistic specificity is a fitness system, whereby a system in a wearable device collects data (e.g., heart rate), and a web service visualises data in a meaningful way.

2.4.2. Device Organisation

In most situations of multi-device service, data and functions are distributed across devices and cannot be completely sourced from a single device. This is due to two main reasons. Firstly, technical constraints of mobile devices prevent accessing the full advantages of a large amount of data and complex functions. Secondly, device-unique capabilities can allow only the access of some functions from a specific device. For example, non-mobile devices may not have mobile device capabilities such as geo-location services, accelerometer, camera, gyroscope, and video recording. Denis and Karsenty [13] outlined three degrees of device redundancy representing levels of data and functions availability across devices.

The first level is redundant, where all the interactive systems across devices can allow access to the same data and functions. In this redundancy level, multi-device service can be classified into two types. Responsive redundant service refers to the multi-device service with the same data and

functions adapted to varying screen sizes, resolutions, aspect ratios and orientations, see Figure 1. Independent redundant service refers to multi-device service with the same data and functions with different appearance or structure in each device, see Figure 1.

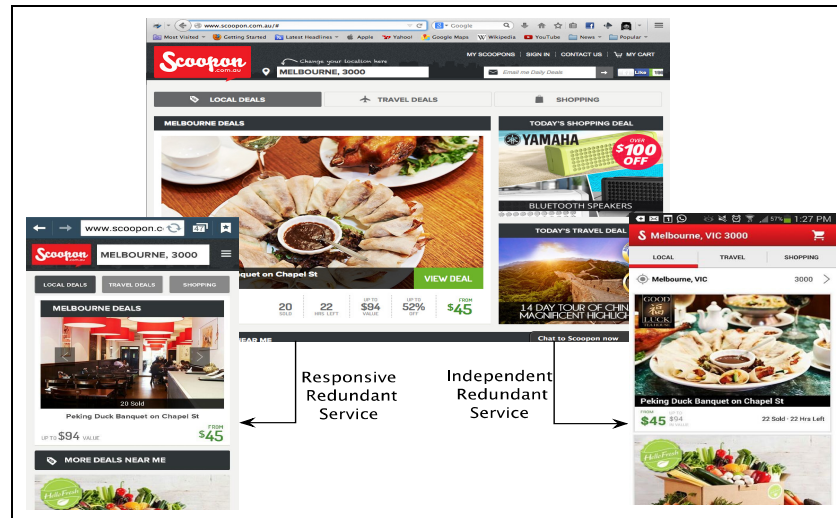


Figure 1: Responsive redundant service versus independent redundant service, for Scoopon service (www.scoopon.com)

The second level of device redundancy is exclusive, where each interactive system in each device gives access to different data and functions. This level of redundancy has the lowest degree of synergistic specificity. An example of this type of redundancy can be found with the Samsung WatchOn multi-device system (www.samsung.com/us/watchon/), which is composed of an interactive TV system and a native mobile app that is used as a remote control for the television service. From the mobile application, users can choose programs directly from their mobile devices to watch on the high-quality display screen of the television. They can also share favourite TV shows with friends.

The third level of device redundancy is complementary, whereby the interactive systems in all devices have a zone of shared data and functions, but one or more of the devices offer access to data or functions that are inaccessible on the other device(s). An example of this redundancy level can be found in Evernote multi-device service (www.evernote.com). The service allows the user to write notes of all types that can then be accessed from different devices. The interactive systems of Evernote across all devices have a shared zone of functions. However, some functions can only be found with mobile devices, such as taking a photo using a device camera to include it in user notes.

As the number of interactive cross-platform systems increases, there is greater probability of more varied configuration of device redundancies of data and functions. Figure 2 shows different degrees of device redundancy of a multi-device service across three devices.

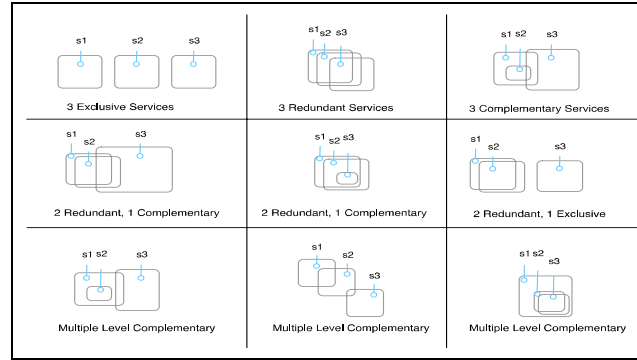


Figure 2: Degrees of device redundancy of a multi-device service across three devices

2.4.3. Service Delivery

Service delivery can be defined as the technique of delivering services to multiple devices. There are three main types of service delivery. The first type is multichannel service delivery. The concept of multichannel service delivery was coined in marketing research to describe all multiple routes (including on and off-line channels) by which customers and business interact with each other [14]. In pervasive computing and HCI, multichannel service delivery refers to the concept of functionality and content being channelled through multiple devices. The aim of this type of service delivery is to provide the practical means for anytime and anywhere customisation of a service for changing user needs and business requirements, and support access to functionality and information from multiple channels [15, 16]. Multichannel service often requires redundant or complementary devices in which core functionalities are supported in different devices. Figure 3 illustrates the conceptual view of a multichannel service delivery in which functionality and content are being channelled through multiple devices into system images.

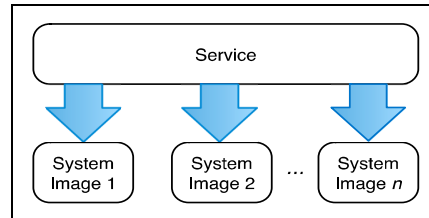


Figure 3: The conceptual view of multichannel service delivery

The second type of service delivery is cross-media (also referred to as transmedia). The term cross-media has been developed in the context of communications research traversing two computer science fields, namely, pervasive computing and human-computer interaction [17]. The term is used in communications research to describe a communication format in which the storyline invites the receiver (user) of media to cross over from one medium to the next in an attempt to achieve a goal or receive a full story [18]. For example, a user finishes watching a TV show and then follows a URL provided at the end of the show to further explore the show. The cross-media concept requires a range of devices including TVs, mobiles, PCs and so on to distribute the content and spread a story across different platforms [19]. With cross-media communication, the systems across devices are designed to be experienced fragmentarily, see Figure 4. Thus, cross-media services are highly synergistic in that users can only achieve a goal if they use the full package of systems as no single system can provide the full package of the content, see example in [20]. In contrast, multichannel services are usually characterised with less

synergistic specificity as users can achieve a goal using any number of channels in tandem or in an isolated manner. In comparison with multichannel services, cross-media services tend to employ exclusive device redundancy, and sometimes complementary device redundancy.

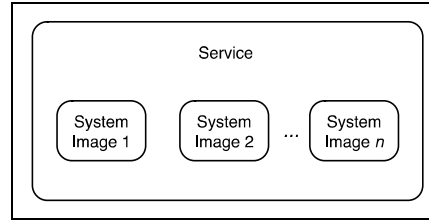


Figure 4: The conceptual view of cross-media service delivery

The third type of service delivery occurs in a cross-channel format where functionality and content is channelled through multiple devices but not in full mode like in multichannel service delivery, see Figure 5. In comparison with cross-media services, users with cross-channel services can achieve a goal within an individual channel without having recourse to other channels. However, in contrast to multichannel services, users cannot interact with all functions and content from a single channel in cross-channel services, which means that there will be at least one central service that includes all content and functions. Cross-channel services often employ complementary device redundancy and have a medium level of synergistic specificity between cross-media and multichannel services. An example of this type of service delivery can be found in the YouTube cross-channel service (www.youtube.com), whereby users can access full content and functions when using PCs, and can access fewer functions and content when using the service through Internet TV (e.g., AppleTV: www.apple.com/au/appletv/). Figure 5 illustrates the conceptual view of cross-channel service delivery in which functionality and content are being channelled through multiple devices into system images. The level of functionality and content being channelled to system images can differ from one device to another.

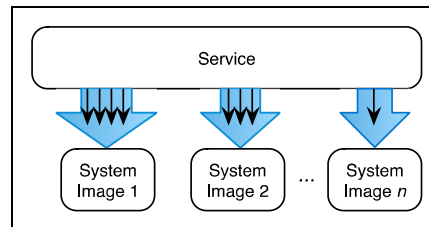


Figure 5: The conceptual view of cross-channel service delivery

3. HCI-RELATED TERMINOLOGY AND CONCEPTS

Several terms associated with cross-platform compilations have been developed in the literature. In the following section, we have reviewed some of these terms.

3.1. Distributed User Interface

There are several concepts associated with the term Distributed User Interface (DUI). One of the early concepts of DUI was migratory applications, introduced by Bharat and Cardelli [21], to describe applications that are capable of roaming on the network instead of being limited to an individual computer. The plasticity of a user interface concept is also associated with DUI, referring to the capability to adapt application interfaces to a new device with different capabilities of input (e.g., touch, stylus, or mouse) and output (e.g., screen sizes in laptop, or

mobile) [22]. Multi-device interaction technique is another concept used in the context of DUI for input redirection where input events entered by users from one device are sent to another device in the same environment [23]. An example of multi-device interaction techniques can be found in the multi-display pointers that move across multiple views [24]. As far as output technique is concerned, content redirection is the most common distribution concept relevant to the term DUI. It refers to redirecting content (e.g., graphical output) across several devices [25, 26]. DUI has also been used widely in several publications, to describe interactive systems that extend across devices (see e.g., [27-29]).

3.2. Multiple User Interface

The term Multiple User Interface (MUI) was first introduced by Seffah [30], and has subsequently gained widespread acceptance among HCI researchers (see e.g., [31, 32]). Seffah [30] used MUI to refer different views of the same information and manage the services that can be accessed by users from different computing platforms. Computing platforms in [30] refers to a combination of hardware (e.g. office desktops, laptops, mobile phones, and tablets), operating systems (e.g. iOS, Windows, Mac OS), computing capabilities and UI toolkit. MUIs can support different interaction styles across platforms, which need to take into account constraints of each device [33]. The concept of MUI is different from multi-device user interface. Multi-device user interface is concerned with whether user interface across devices are able to allow a user to interact with them with any input style [33]. This is different from the MUI concept, which is concerned with different views of the same service across platforms.

There are four main aspects of MUI [34]. Firstly, MUI allows an individual user or a group of users to interact with server-side services using different interaction styles. Secondly, MUI can be designed to support interrelated tasks that can be achieved using more than one device. Thirdly, although a user interface in each device may have its unique look and feel, MUI can display features, functions, and information that can have the same behaviour across platforms. Finally, MUI refers to various assortments of a single service, for example a user interface, for different computing devices.

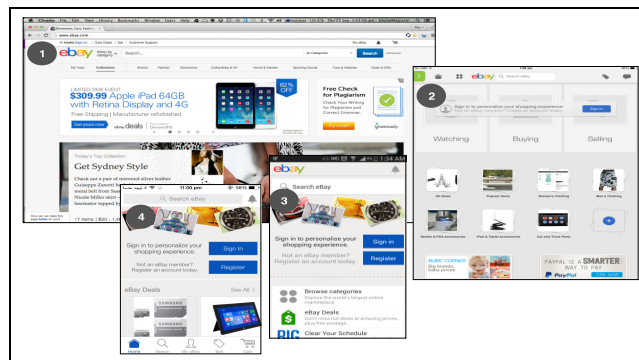


Figure 6: MUIs of eBay service across different devices and interaction styles

Figure 6 shows an MUI to the same system (www.ebay.com) from four different devices (laptop, iPad, Samsung Galaxy and iPhone). MUIs can be a combination of interaction styles [33]. The eBay system across platforms consists of four interaction styles, web based user interface, (1) in Figure 6, native iPad application, (2) in Figure 6, native android application, (3) in Figure 6, and native iPhone application, (4) in Figure 6. Hence, MUI can be a combination of any possible interaction styles that can exist across platforms, see e.g., Figure 7.

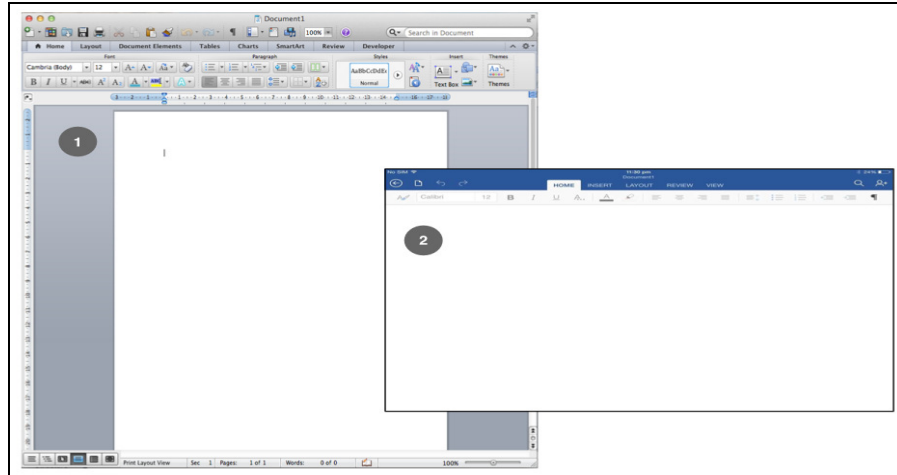


Figure 7: MUIs of Microsoft Word across two interaction styles, where (1) refers to a Graphical User Interface from Laptop running Mac OS and (2) a Native iPad Application from an iPad

There is a lack of research focus in the literature on classifying MUIs. In the following, we have attempted to categorise MUIs into three different models: on-demand model, independent model, and hybrid model. In the on-demand model, the service model can be stored in a single information repository, and delivered to the user on demand, as the user can request the service using a web browser. Figure 8 shows the on-demand model for Amazon's MUIs, which were accessed using web browsers across two devices.

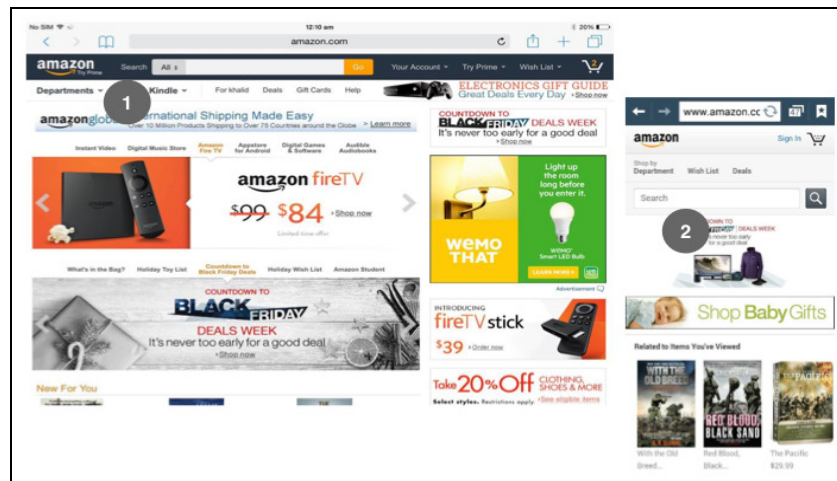


Figure 8: MUIs of Amazon service form a single repository, where (1) refers to the Amazon service accessed via web browser from iPad, and (2) refers to Amazon service accessed via web browser from Samsung Galaxy phone

In the independent model, the service model can be distributed among independent systems, while each view of the MUI can be seen as an all-inclusive user interface for each specific platform that runs it. In this type of MUI, function and information can vary from one platform to another. Figure 9 shows the independent model of Amazon's MUIs for two independent user interfaces installed on two different devices.

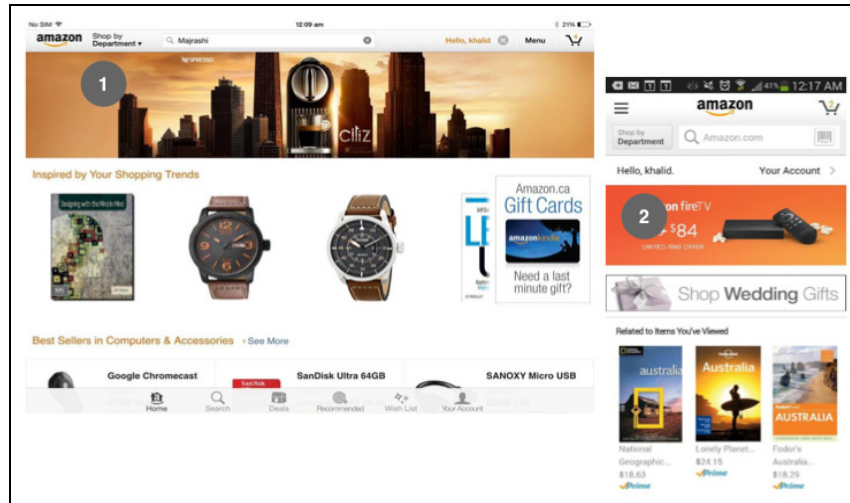


Figure 9: MUIs of Amazon service installed on different independent systems, where (1) refers to Amazon service installed on iPad, and (2) refers to Amazon service installed on Samsung Galaxy phone

In the hybrid model, the MUI can be a combination of on-demand and independent models, including services that can be accessed using web browsers and services installed on computing devices. The combination of web based application, see Figure 8, and native device applications, see Figure 9, represents the hybrid model of Amazon's MUIs.

4. TRADITIONAL USER EXPERIENCE AND USABILITY

In this section, we review traditional concepts of user experience and usability and also discuss the differences between them.

4.1. Traditional User Experience

User experience (UX) is a term used broadly by HCI practitioners and researchers to represent a variety of meanings [35]. UX is considered as an umbrella term for a range of dynamic concepts, such as traditional usability (see e.g., [36, 37]), affective, and emotional (see e.g., [38-41]), experiential (see e.g., [35, 42]), hedonic (see e.g., [43, 44]), aesthetic (see e.g., [45]), and values variables. There is also an argument that user experience goes far beyond interaction with user interfaces. For example, Jakob and Don [46] have suggested that people need to separate the association of the broad concept of user experience from the experience with regard to design of User Interface (UI). They see UI as one aspect of several forms of interactions with a service.

In the following, we present some UX definitions from the literature:

- Hassenzahl and Tractinsky [47] defined UX as “a consequence of a user’s internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.), and the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)”
- Jakob and Don [46] defined UX as “all aspects of the end-user’s interaction with the company, its services, and its products”.
- Alben [48] defined UX as “all the aspects of how people use an interactive product: the way it feels in their hands, how well they understand how it works, how they feel about it while

they're using it, how well it serves their purposes, and how well it fits into the entire context in which they are using it".

- International Organization for Standardization [49] defined UX as "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service". It is clear that all these definitions are concerned about the result of end-user interaction as a means of understanding user experience. The definition by Hassenzahl and Tractinsky [47] explicitly stated the variables that can impact the user experience of end-user interaction, whether it be the user's internal state, the system itself, or the environment where the interaction occurs.

4.2. Traditional Usability

Usability is an important attribute of software quality measured by a range of metrics and techniques to assess how easy a user interface is to use. Although usability has its academic origins in the HCI community, the term has no shared standard definition. Bevan [50] outlined that the term usability has been interpreted differentially by different people using different standards.

A few common definitions of usability are listed below:

- Shackel defined usability as "[a system's] capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and support, to fulfil a specified range of tasks, within the specified range of environmental scenarios" [51].
- IEEE defined usability as "The ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component" [52].
- ISO/IEC 9126 defined usability as "A set of attributes that bear on the effort needed for use and on the individual assessment of such use, by a stated or implied set of users" [53].
- Preece's defined usability as "a measure of the ease with which a system can be learned or used, its safety, effectiveness and efficiency, and the attitude of its users towards it" [54].
- ISO 9241-11 defined usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [55].

As it can be seen from the definitions above, usability is a combination of multiple attributes, for example, effectiveness, and efficiency. Different interpretations of usability as a term across academic and industry circles may have impacted the identification of standardised usability attributes in a consistent way over time.

In the previous sub-section, some definitions of user experience have included usability as an aspect of user experience. In the following sub-section, we illustrate differences between user experience and usability.

4.3. User Experience and Usability

In its published notes on user experience, International Organization for Standardization [49] has stated that "User experience includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use". This note incorporates usability within user experience inexplicitly, whereby "behavior and accomplishments" can include two important usability attributes; efficiency (time to execute task) and effectiveness (completion of task). Therefore, user experience can be seen an umbrella for different concepts. This judgment is supported by different definitions of user experience, such as that given by Hassenzahl and Tractinsky [47], which defined user experience

as a consequence of multiple factors including the characteristics of the designed system such as usability. The usability criteria can also be used as a metric to assess user experience [49, 56]. Based on the reviewed definitions of user experience and usability and what we have discussed on the overlapping meaning between the two terms in this section, we attempt to illustrate differences and relationships between user experience and usability, see Figure 10.

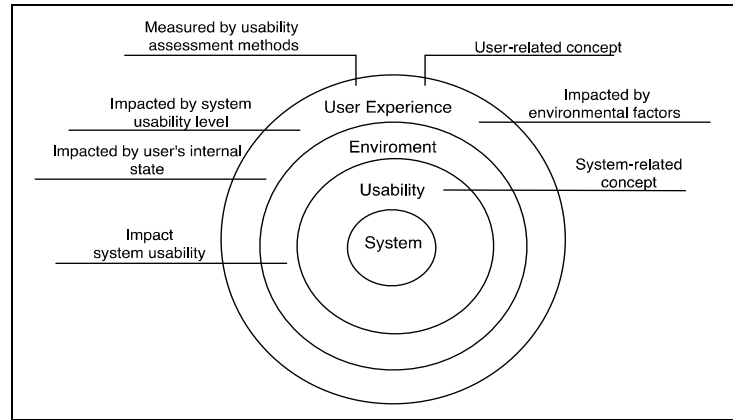


Figure 10: Differences and relationships between user experience and usability

In summary, user experience and usability can be conceptualized in different ways on basis of the following points:

- User experience is a broad term encompassing multiple factors including system usability [47].
- User experience is associated with user perception [47, 49], however, usability is more about the design of a system ISO [55].
- Usability can be impacted by environmental factors (including social and organisational factors) [51, 55]. While system usability can influence user experience negatively or positively [47], the impact of environmental factors on usability can lead to direct effect on user experience.
- Users' internal states, such as beliefs and exceptions, can impact how they use a system [57]. Thus, usability level as perceived by users can be impacted by their internal states. This can also lead to negative or positive user experience based on whether system conforms to user's mental state.
- Usability can be assessed with objective measures (e.g., time to execute task) and or subjective measures (e.g., satisfaction rate). User experience can be measured through usability assessment methods [49], based on subjective measures.

Having discussed traditional concepts of user experience and usability, we now turn to cross-platform usability, and user experience in the next two sections.

5. CROSS-PLATFORM USABILITY

These are some issue that we need to go through before defining cross-platform usability:

- Current usability metrics, for example, task execution time, focus on measuring usability of a single user interface. Thus, they need to be reconsidered for horizontal interaction across multiple user interfaces, for instance, execution time when attempting a task across platforms.

- Horizontal interaction involves using multiple user interfaces, in which every single user interface can be employed in a specific context of use. Factors that can affect usability in each context of use need to be considered when investigating usability across platforms.

After reviewing traditional definitions of usability from different standards and models, we identified characteristics of multiple user interfaces to arrive at the following definition of Cross-Platform Usability (CPU):

The extent to which a service across platforms can be used by specified users to achieve specified horizontal goals in specific or different contexts of use with acceptable level of several measurable factors including efficiency, effectiveness, and satisfaction.

6. CROSS-PLATFORM USER EXPERIENCE

There are multiple variables that can impact end-user perceptions when interacting with a service across platforms. These variables are listed below:

- The design of multiple user interfaces must be considered, particularly, given the fact that each user interface may have a unique design.
- Computing platforms used to interact with multiple user interfaces can have different characteristics. Computing platforms refers to a combination of hardware (e.g., office desktops, laptops, mobile phones, and tablets), operating systems (e.g., iOS, Windows, Mac OS), and computing capabilities. Each portion of this combination can have characteristics that can impact user experience across platforms. Some examples of these characteristics are:
 - Devices (e.g., input style, display size)
 - Operating systems (e.g., display, design, feature).
 - Computing capabilities (e.g., capabilities of processors, storage)
- Environments in which interactions with the multiple user interfaces occurs is also important.

We have adopted the traditional definition of user experience by Hassenzahl and Tractinsky [47] and modified to incorporate the variables and characteristics of cross-platform user interaction stated above. Thus, we define Cross-Platform User Experience (CPUX) as:

The consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed systems across platforms (e.g. service cohesion, composition, horizontal usability, distributed functionality, etc.), the characteristics of the computing platforms used to allow interactions with the systems (devices [display size, input style etc], operation systems [display, design, feature etc], computing capabilities [capabilities of processing, storage etc]) and the contexts (or the environments) within which the multiple interactions occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, user situation [seating, standing, driving] etc).

7. CONCLUSION

In this paper, we have provided a thorough discussion of different concepts that need to be considered in the context of cross-platform user experience. This includes concepts and practical approaches relevant to cross-platform service, and its relevant terms in the field of HCI. We have also investigated the definitions and characteristics of traditional usability and user experience. Then, we extended on these traditional concepts to develop a definition of cross-platform

usability and user experience based on characteristics of user interaction across platforms. It is hoped that the definitions and discussions in this paper have contributed in building the necessary theoretical foundations for further study on cross-platform user experience evaluation.

REFERENCES

- [1] K. Majrashi and M. Hamilton, "A Cross-Platform Usability Measurement Model," *Lecture Notes on Software Engineering*, vol. 3, 2015.
- [2] M. Bell, *Service-oriented modeling (SOA): Service analysis, design, and architecture*: John Wiley & Sons, 2008.
- [3] M. Rosen, B. Lublinsky, K. T. Smith, and M. J. Balcer, *Applied SOA: service-oriented architecture and design strategies*: John Wiley & Sons, 2008.
- [4] The Linux Information Inc, "Cross-platform Definition," 2005.
- [5] A. T. Manes, *Web Services: A Manager's Guide*: Addison-Wesley Longman Publishing Co., Inc., 2003.
- [6] Q. Yu and A. Bouguettaya, *Foundations for Efficient Web Service Selection*: Springer, 2009.
- [7] N. Ide and J. Pustejovsky, "What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology," in *Proceedings of the Second International Conference on Global Interoperability for Language Resources*. Hong Kong, China, 2010.
- [8] G. Alonso and F. Casati, "Web services and service-oriented architectures," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, p. 1147.
- [9] S. K. Feiner, "Environment management for hybrid user interfaces," *Personal Communications, IEEE*, vol. 7, pp. 50-53, 2000.
- [10] M. Wäljas, K. Segerståhl, K. Väänänen-Vainio-Mattila, and H. Oinas-Kukkonen, "Cross-platform service user experience: a field study and an initial framework," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, 2010, pp. 219-228.
- [11] M. A. Schilling, "Toward a general modular systems theory and its application to interfirm product modularity," *Academy of management review*, vol. 25, pp. 312-334, 2000.
- [12] H. A. Simon, *The architecture of complexity*: Springer, 1991.
- [13] C. Denis and L. Karsenty, "Inter-usability of multi-device systems: A conceptual framework," *Multiple user interfaces: Cross-platform applications and context-aware interfaces*, pp. 373-384, 2004.
- [14] H. Wilson, R. Street, and L. Bruce, *The multichannel challenge: integrating customer experiences for profit*: Routledge, 2008.
- [15] F. G. Kazasis, N. Moumoutzis, N. Pappas, A. Karanastasi, and S. Christodoulakis, "Designing Ubiquitous Personalized TV-Anytime Services," in *CAiSE Workshops*, 2003.
- [16] P. Fraternali, A. Bozzon, M. Brambilla, V. Croce, K. Hammervold, E. Moore, et al., "Model-driven development of personalized, multichannel interfaces for audiovisual search: the PHAROS approach," *NEM Summit*, Saint Malo, France, 2009.
- [17] C. Wiberg, K. Jegers, and J. Bodén, "Cross media interaction design," 2007.
- [18] L. V. L. Filgueiras, D. O. Correa, J. S. O. Neto, and R. P. Facis, "X-gov planning: how to apply cross media to government services," in *Digital Society, 2008 Second International Conference on the*, 2008, pp. 140-145.
- [19] J. Boumans, "Cross-media E-Content Report 8," Published in a series of E-Content Reports by ACTeN (<http://www.acten.net>), 2004.
- [20] K. Segerståhl, "Crossmedia systems constructed around human activities: a field study and implications for design," in *Human-Computer Interaction-INTERACT 2009*, ed: Springer, 2009, pp. 354-367.
- [21] K. A. Bharat and L. Cardelli, "Migratory applications," in *Proceedings of the 8th annual ACM symposium on User interface and software technology*, 1995, pp. 132-142.
- [22] D. Thevenin and J. Coutaz, "Plasticity of user interfaces: Framework and research agenda," in *Proceedings of INTERACT*, 1999, pp. 110-117.
- [23] B. Johanson, G. Hutchins, T. Winograd, and M. Stone, "PointRight: experience with flexible input redirection in interactive workspaces," in *Proceedings of the 15th annual ACM symposium on User interface software and technology*, 2002, pp. 227-234.

- [24] M. A. Nacenta, D. Aliakseyeu, S. Subramanian, and C. Gutwin, "A comparison of techniques for multi-display reaching," in Proceedings of the SIGCHI conference on Human factors in computing systems, 2005, pp. 371-380.
- [25] J. T. Biehl, W. T. Baker, B. P. Bailey, D. S. Tan, K. M. Inkpen, and M. Czerwinski, "Impromptu: a new interaction framework for supporting collaboration in multiple display environments and its field evaluation for co-located software development," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 939-948.
- [26] J. R. Wallace, R. L. Mandryk, and K. M. Inkpen, "Comparing content and input redirection in MDEs," in Proceedings of the 2008 ACM conference on Computer supported cooperative work, 2008, pp. 157-166.
- [27] K. Luyten and K. Coninx, "Distributed user interface elements to support smart interaction spaces," in Multimedia, Seventh IEEE International Symposium on, 2005, p. 8 pp.
- [28] M. Bång, A. Larsson, E. Berglund, and H. Eriksson, "Distributed user interfaces for clinical ubiquitous computing applications," in International Journal of Medical Informatics vol. 74, ed, 2005, pp. 545-551.
- [29] K. Segerståhl and H. Oinas-Kukkonen, "Distributed user experience in persuasive technology environments," in Persuasive Technology, ed: Springer, 2007, pp. 80-91.
- [30] A. a. F. Seffah, Peter, "Workshop on multiples user interfaces over the Internet: engineering and applications trends," In: HM-HCI: French/British Conference on Human Computer Interaction, Lille, France, 2001.
- [31] J. Vanderdonckt, Q. Limbourg, M. Florins, F. Oger, and B. Macq, "Synchronised, model-based design of multiple user interfaces," in Proc. 2001 Workshop on Multiple User Interfaces over the Internet, 2001.
- [32] J. McGrenere, R. M. Baecker, and K. S. Booth, "An evaluation of a multiple interface design solution for bloated software," in Proceedings of the SIGCHI conference on Human factors in computing systems, 2002, pp. 164-170.
- [33] A. Seffah and H. Javahery, Multiple user interfaces: cross-platform applications and context-aware interfaces: John Wiley & Sons, 2005.
- [34] A. Seffah, P. Forbrig, and H. Javahery, "Multi-devices "Multiple" user interfaces: development models and research opportunities," in Journal of Systems and Software vol. 73, ed, 2004, pp. 287-300.
- [35] J. Forlizzi and K. Battarbee, "Understanding experience in interactive systems," in Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques, ed, 2004, pp. 261-268.
- [36] W. Albert and T. Tullis, Measuring the user experience: collecting, analyzing, and presenting usability metrics: Newnes, 2013.
- [37] N. Bevan, "Classifying and selecting UX and usability measures," in International Workshop on Meaningful Measures: Valid Useful User Experience Measurement, 2008, pp. 13-18.
- [38] R. W. Picard, "Affective Computing for HCI," in HCI (1), 1999, pp. 829-833.
- [39] F. N. Egger, "Affective design of e-commerce user interfaces: How to maximise perceived trustworthiness," in Proc. Intl. Conf. Affective Human Factors Design, 2001, pp. 317-324.
- [40] D. A. Norman, Emotional design: Why we love (or hate) everyday things: Basic books, 2004.
- [41] H. M. Khalid and M. G. Helander, "Customer emotional needs in product design," Concurrent Engineering, vol. 14, pp. 197-206, 2006.
- [42] S. Baurley, "Interactive and experiential design in smart textile products and applications," Personal and Ubiquitous Computing, vol. 8, pp. 274-281, 2004.
- [43] P. M. Tsang and S. Tse, "A hedonic model for effective web marketing: an empirical examination," Industrial Management & Data Systems, vol. 105, pp. 1039-1052, 2005.
- [44] M. G. Helander, "Hedonomics-affective human factors design," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2002, pp. 978-982.
- [45] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," International journal of human-computer studies, vol. 60, pp. 269-298, 2004.
- [46] N. Jakob and N. Don, "The Definition of User Experience."
- [47] M. Hassenzahl and N. Tractinsky, "User experience-a research agenda," Behaviour & Information Technology, vol. 25, pp. 91-97, 2006.
- [48] L. Alben, "Quality of Experience, Interactions," 1996.
- [49] International Organization for Standardization, Ergonomics of Human-system Interaction: Part 210: Human-centred Design for Interactive Systems: ISO, 2010.

- [50] N. Bevan, "Measuring usability as quality of use," *Software Quality Journal*, vol. 4, pp. 115-130, 1995.
- [51] B. Shackel, "The concept of usability," *Visual display terminals: usability issues and health concerns*, pp. 45-87, 1984.
- [52] J. Radatz, A. Geraci, and F. Katki, "IEEE standard glossary of software engineering terminology," *IEEE Std*, vol. 610121990, p. 121990, 1990.
- [53] ISO/IEC 9126, "Information Technology, Software Product Evaluation, Quality Characteristics and Guidelines for their Use," Geneva, Switzerland: International Organization for Standardization., 1991.
- [54] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-computer interaction: Addison-Wesley Longman Ltd.*, 1994.
- [55] ISO 9241-11, "Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11: Guidance on Usability," Geneva, Switzerland: International Organization for Standardization., 1998.
- [56] K. Majrashi and M. Hamilton, *User Experience of University Websites: LAP Lambert Academic Publishing*, 2014.
- [57] N. Jakob and N. Don, "Mental Models," 2010.

AUTHORS

Khalid Majrashi was born in Jazan, Saudi Arabia, in 1987. He received the B.E degree in education of computer science from the University of Jazan, Jazan Saudi Arabia, in 2008, and the Master degree in computer science from RMIT University, Melbourne, Australia in 2012, and currently is PhD candidate (in the field of computer science) at RMIT University, Melbourne, Australia. His current research interests include software quality, user experience, usability, mobile computing, cloud computing (software as a service), and user interface models.



Margaret Hamilton, BSc (Hons), PhD, University of Wollongong, Australia, is an Associate Professor in the School of Computer Science and Information Technology at RMIT University in Melbourne Australia, where she researches in the field of human computer interaction. She works with new technologies in computer science education and mobility to investigate the human aspects and to research how computer interfaces can be adapted to be more user-friendly, and computer programs to be more relevant, realtime and personalised.



Alexandra L. Uitdenbogerd, BSc, University of Western Australia, Grad. Dip. Ed, Melbourne University, Grad Cert IT, and PhD, RMIT University, Australia, is a senior lecturer in the School of Computer Science and Information Technology at RMIT University in Melbourne Australia. Her current research interests include search engine technology, music/audio applications, information retrieval, computer-assisted language learning, pattern matching, search engines and related areas.



INTENTIONAL BLANK

QUALITY ASSESSMENT FOR ONLINE IRIS IMAGES

Sisanda Makinana, Tendani Malumedzha, and Fulufhelo V Nelwamondo

Modelling and Digital Science, CSIR, Pretoria, South Africa
smakinana@csir.co.za

ABSTRACT

Iris recognition systems have attracted much attention for their uniqueness, stability and reliability. However, performance of this system depends on quality of iris image. Therefore there is a need to select good quality images before features can be extracted. In this paper, iris quality is done by assessing the effect of standard deviation, contrast, area ratio, occlusion, blur, dilation and sharpness on iris images. A fusion method based on principal component analysis (PCA) is proposed to determine the quality score. CASIA, IID and UBIRIS databases are used to test the proposed algorithm. SVM was used to evaluate the performance of the proposed quality algorithm. . The experimental results demonstrated that the proposed algorithm yields an efficiency of over 84 % and 90 % Correct Rate and Area under the Curve respectively. The use of character component to assess quality has been found to be sufficient for quality detection.

KEYWORDS

Image quality, Iris recognition, Principal Component Analysis, Support Vector Machine.

1. INTRODUCTION

Iris recognition is an automated method of biometric identification that analyses patterns of the iris to identify an individual [1]. It is said to have high reliability in identification because each individual has unique iris patterns [2], [3]. However, due to the limited effectiveness of imaging, it is important that image of high-quality images are selected in order to ensure reliable human identification. Some advanced pre-processing algorithms can process poor quality images and produce adequate results, however they are computationally expensive and add extra burden on the recognition system time. Therefore quality determination is necessary in order to determine which algorithm to use for pre-processing. For example, if it's known that the acquired image does not meet the desired quality it can be subjected to stricter pre-processing algorithms selectively. Various quality assessment methods have been developed to ensure quality of the sample acquisition process for online systems [4]. These approaches are good for quick elimination of poor quality images and even images from which an accurate segmentation may be produced are eliminated. A more discriminative approach to quality, images can be assigned quality levels, which will provide an indication as to whether further processing can enhance them.

Generally, an iris sample is of good quality if it provides enough features for reliable identification of an individual [5]. Therefore, there is need for a standard sample quality that

stipulates the accepted quality for iris images. In this regard, ISO/IEC have developed three quality components which together defines biometric sample quality, these are; character, fidelity and utility [6]. In this paper, the focus is on character component of a biometric sample quality due to the fact that available algorithms utilises and focuses on fidelity and utility components [4], [7], [8]. This paper proposes an algorithm that assesses the quality of an iris image based on online biometric systems. Firstly, image quality parameters are estimated, i.e. contrast, sharpness, blur, dilation, area ratio, standard deviation and occlusion. Thereafter, a fusion technique based on principal component analysis is used to weight each quality parameter and obtain a quality score for each image. The remainder of this paper is organised as follows. Section II describes the overview of the proposed method. Section III provides the estimations of individual quality parameters and discusses the implementation of the proposed fusion method. Last sections provide experimental results and a conclusion.

2. ESTIMATION OF QUALITY PARAMETERS

Implementation of assessment algorithm is carried out in two steps: namely, iris segmentation and estimation and fusion of quality parameters. The subsections below details how this is done.

2.1. Iris Segmentation

To locate the iris from the acquired eye image, parameters (radius r , and the coordinates of the centre of the circle, x_0 and y_0) of detecting the centre of the iris and pupil were determined by use of integrodifferential operator discussed in [2]. This operator locates and segments the pupil and iris regions with varying coordinates. Equation 1 describes the integrodifferential operator:

$$\max(r, x_0, y_0) \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint_{r, x_0, y_0} \frac{I(x, y)}{2\pi r} ds \right| \quad (1)$$

Where $G_\sigma(r)$ is the Gaussian kernel, $I(x, y)$ is the eye image, $(x_0, y_0), r$, are the centre coordinates and radius of the circle respectively.

2.2. Estimation of quality parameters

The following are the quality parameters that are estimated for the proposed algorithm:

2.2.1 Occlusion

The occlusion measure (M_O) is the amount of iris region that is invalid due to obstruction by eyelids and eyelashes. Eyelid and eyelashes occlusion problem is a primary cause of bad quality in iris image [9]. Compared with the edge of iris texture, the edge of iris-lid and iris-lash is much sharper and usually considered to contain high pixel values [9]. To estimate the amount of occlusion at each level an occlusion is measured by calculating the ratio of total gray value of the image [9]. It is defined as:

$$T_G = A_I \times \sum_{i=1}^m \sum_{j=1}^n I \quad (2)$$

$$G_R = \frac{1}{T_G} \quad (3)$$

$$M_O = \text{mean}(G_R) \quad (4)$$

Where in (2) A_I and I is the area of the iris and the iris intensity, m and n represents the size of the image and TG is the total gray value. In (3) T_G is the ratio of the total gray value. The higher the metric value the greater is the chance for occlusion by iris lid.

2.2.2 Blur

Blur may result from many sources, but in general it occurs when the focal point is outside the depth of field of the captured object [9]. The further the object is from the focal point of the capturing device, the higher degree of blur in the output image [10]. The blur estimation is based on Crete et al [11] approach. The blur estimation procedure is illustrated in Fig. 1.

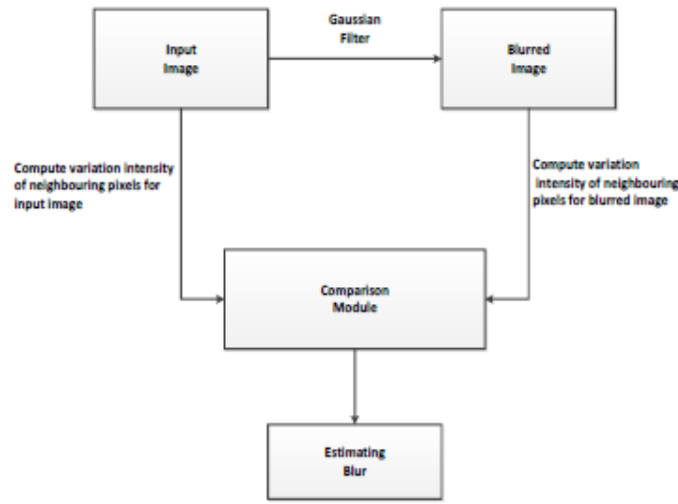


Figure 1 The framework of the blur estimation [9]

2.2.3 Area Ratio

The representation of pattern recognition should be invariant to change in the size, position and orientation of the iris image. If the subject is too close to the capturing device that may cause the captured image to be blurred. Thus, it's of utmost importance to assess the iris area ratio, which is the ratio of the iris over the image frame [9]. It is assumed that the iris is circular, therefore the area of the iris is equivalent to the area of the circle given in (4) which is then defined as [9]

$$A_I = \pi r^2 \quad (4)$$

The area of the image frame is given as:

$$A_E = H * W \quad (5)$$

Where H is the height and W is width of the image. Therefore the area ratio is derived as:

$$M_A = \frac{A_I}{A_E} \quad (6)$$

2.2.4 Contrast

The term contrast refers to the representation of colour and difference in luminance that makes a feature noticeable within an image [12]. However human vision is more sensitive to difference in colour representation than difference in luminance. According to human visual, contrast is the difference in colour and brightness of various objects within the same field of view. Contrast determines the clearness of the features within an image. High contrast means the more clearly the iris features and making easier for feature extraction. Assessing contrast is important to ensure sufficient and clear features are extracted. In measuring contrast a window of 8 x 8 pixels is defined, of which the Fast Fourier Transform (FFT) of the 2-Dimensional image is computed for each sub-window. FFT transforms the signal from time domain into a frequency domain and is defined as [13]:

$$H(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{M-1} \exp\left[-2\pi i y \frac{v}{M}\right] \exp\left[-2\pi i x \frac{u}{M}\right] h(x, y) \quad (7)$$

Where $u, v = -\frac{M}{2} \dots \frac{M}{2}$. The squared magnitude of the Fourier series coefficients which indicates power at corresponding frequencies is computed by Parseval's Theorem [14]:

$$P(u, v) = |H(u, v)|^2 = (\text{Re}(u, v) + \text{Im}(u, v))^2 \quad (8)$$

The fundamental frequency (dcpower), the total power (totpower) and the non-zero power (acpower) of the spectrum is computed which are [14]:

$$dcpower = P(0,0) = |H(0,0)|^2 \quad (9)$$

$$totpower = \sum_{u=-M}^M \sum_{v=-M}^M P(u, v) \quad (10)$$

$$acpower = totpower - dcpower \quad (11)$$

The contrast is computed as follows [14]:

$$M_c = \sqrt{\frac{acpower}{dcpower}} \quad (12)$$

2.2.4 Standard Deviation

Standard deviation method uses the standard deviation of gray-level distribution in a local area region of an iris image. The iris image is divided into N X N regions. The local standard deviation of each region is computed and added together to obtain a single standard deviation. Then, the mean of the summed standard deviation is quality score of the entire image.

$$S_k = \sqrt{\frac{1}{N^2} \sum_{x=1}^N \sum_{y=1}^N (I_{xy} - I_k)^2} \quad (13)$$

$$M_{STD} = \frac{1}{N} \sum_{k=1}^N S_k \quad (14)$$

In (13), $I_{x,y}$ is the gray level of pixel (x, y) and I_k is average gray level of the k^{th} region.

2.2.5 Sharpness

A sharp image is the one that contains fine details that determine the amount of detail an image can produce and has edges and objects appearing to be of high contrast. Images are usually affected by distortions during acquisition and processing, which may result in loss of visual quality. Therefore, image sharpness assessment is useful in such application. Sharpness generally attenuates high frequencies. Due to that factor, sharpness can be assessed by measuring high frequency of the image. Daugman [4] proposed an 8 X 8 convolution kernel to assess sharpness. Sharpness is estimated based on the gradient of the image to determine whether the image is in focus or not, because the gradient of an image is the directional change in the intensity of an image. The gradient of the image is given by:

$$\nabla G = \left(\frac{\partial G}{\partial x} \right) + \left(\frac{\partial G}{\partial y} \right) \quad (15)$$

Where $\frac{\partial G}{\partial x}$ is the gradient in the x direction and $\frac{\partial G}{\partial y}$ is the gradient in the y direction. From (15) the sharpness of the image may be calculated. The sharpness is calculated by dividing the sum of gradient amplitude by the number of elements of the gradient. The gradient amplitude is given by:

$$M_S = S \sqrt{G_x^2 + G_y^2} \quad (16)$$

Where S is the gradient amplitude, G_x and G_y in (16) are the horizontal and vertical change in intensity.

2.2.6. Dilation:

The variation in pupil dilation between the enrolment image and the image to be recognised or verified may affect the accuracy of iris recognition system [9]. The degree of dilation was measured for each iris image. The segmentation results provided the radius of the pupil and of the iris. To measure dilation, a ratio of radius of pupil and radius of iris was calculated. Since the pupil radius is always less than the radius of iris, the dilation will fall between 0 and 1 [9]. The dilation measure M_D is calculated by:

$$M_D = \frac{P_R}{I_R} \quad (17)$$

Where P_R is the pupil radius and I_R is the iris radius.

3. FUSION TECHNIQUE

A unique quality score is of value to the prediction step of iris recognition system. To obtain this quality score, a fusion technique based on Principal Component Analysis (PCA) is proposed. PCA is a widely used tool which is proficient in reducing dimensions and determining factor

relationships amongst datasets just like Factor Analysis (FA) [15]. However, FA evaluates the linear relationship between the number of variables of interest Y_1, Y_2, \dots, Y_j ; and a smaller number of unobserved factors F_1, F_2, \dots, F_k , whereas, PCA is a technique that determines the factor loadings of the dataset by calculating the eigenvectors and eigenvalues of the covariance matrix [16]. In this research PCA has been used over FA since the interest in determining the factor loading of the dataset. Factor loadings are the weights of each variable and correlations between each factor [17]. The PCA is calculated by defining the eigenvectors and eigenvalues of the covariance matrix. The covariance matrix measures the variation of the dimensions from the mean with respect to each other. Prior to applying the PCA, quality parameters need to be normalised. The quality parameters are standardized using the Z_s before obtaining the first PCA, which is:

$$Z_s = \frac{x - \mu}{\sigma} \quad (18)$$

Where μ the mean and σ is the standard deviation of the estimated measures of the entire database. Suppose n independent observation are given on X_1, X_2, \dots, X_k , where the covariance X_i and X_j is

$$\text{Cov}(X_i, X_j) = \sum i, j \quad (19)$$

For $i, j = 1, 2, \dots, k$ in (19). Then the eigenvalues and eigenvectors of the covariance matrix are calculated. W is defined to be the first principal component. It is the linear combination of the X 's with the largest variance:

$$W = a_1^T X_i \quad (20)$$

Where $i = 1, 2, \dots, k$ and a is the eigenvector of the covariance. The quality score is obtained by multiplying normalised measures of parameters with weights for each quality parameter of the image. The fusion quality index is given as:

$$Q_s = \sum_{i=1}^N Q_p W_p \quad (21)$$

Where Q_s the quality score, Q_p is the estimated quality parameter and W_p is the amount of influence each parameter has on the quality score. The scores represent the global quality score of the iris segmented images.

4. QUALITY SCORE NORMALIZATION

Prior to fusion of the parameters to form a quality score, some parameters need to be normalized between [0, 1]. Sharpness, Area Ratio, dilation and blur are already in the desired score range. Occlusion is normalized based on the max normalization. The fused score also needed to be normalized between [0, 1] with 0 implying bad quality and 1 good quality. The normalization of the quality score is based on the modified form of min-max normalization:

$$Q_s = \frac{Q_{old} - Q_{\min}}{Q_{\max} - Q_{\min}} \quad (22)$$

With Q_{old} represents the raw quality score.

5. DATASET USED FOR ANALYSIS

In this paper, the CASIA and UBIRIS databases which are available free online and Internal Iris Database (IID) database were used, to estimate the quality parameters and their scores. For CASIA a subset of images called 'interval' was used. It contained 525 images which were captured at a resolution of 320 x 280 pixels. UBIRIS consists of 802 images captured at a resolution of 200 x 150 pixels. IID consists of 116 images captured at a resolution of 640 x 480 pixels.

Table I DATABASE DESCRIPTION

Database	Image No.	Bad Quality	Good Quality
UBIRIS	802	685	117
CASIA	525	69	456
IID	116	35	81

6. IMAGE DESCRIPTION

For UBIRIS database images were captured on two different sessions. For the first session noise factors like reflection, luminosity and contrast were minimized by capturing images inside a dark room. In the second session capturing location was changed to introduce noisy images. This introduced diverse images with respect to reflection, contrast, focus and luminosity problems [18].

For CASIA database images were captured by a closed up iris camera with circular NIR Light-Emitting Diode (LED) array which had suitable luminous flux for iris imaging. The camera captures very clear iris images [19].

The IID iris database which was also used for testing the algorithm is a new database of which its images were collected in the Council of Scientific and Industrial Research (CSIR) campus in Pretoria. A Vista EY2 iris camera was used to collect these images.

7. EXPERIMENTAL RESULTS

Analysis was performed on three databases, namely, CASIA Interval, UBIRIS and IID. The sample iris images in Fig. 2 are from UBIRIS, IID and CASIA databases. Based on human visual assessment sample (a) represents good quality from UBIRIS database and (c) represent good quality from IID database. Sample (e) and (f) represent good and bad quality respectively, from CASIA Interval database. Image (b) and (f) represent degraded image quality that is affected by occlusion, blur and dilation. Sample image (b) is also affected by area ratio quality parameter. Table II illustrates the estimated quality parameters of the images in Fig. 2. The quality scores are normalized to the values between zero and one, with one implying good quality and 0 bad quality. The overall quality distribution for CASIA, UBIRIS and IID databases are illustrated in Fig. 3 respectively. CASIA has the highest quality distribution, followed by IID and the UBIRIS. IID suffers from quality degrading with respect to sharpness, dilation and blur, which is visually evident. These parameters have high weight on the quality score of IID which results in low quality. The reason for this problem is the mere fact that the iris capturing session for this database was done in an environment with light which caused reflections, resulting in the images

being less clear. Moreover, individuals were required to focus their eyes to a mirror for a certain period which caused their pupil to dilate. Also, the camera captured iris images automatically and required individuals to be still which caused some discomfort and as the individual became tired and moved, which resulted in the camera capturing blurred images. For the UBIRIS database was captured in an environment that introduced noisy images affected by diverse problems with respect to reflection, contrast, focus and luminosity. The results of individual parameters also indicate that this database is affected by sharpness, dilation, area ratio and blur which is caused by the environment condition. That is why there are more low quality scores for this database. When grading these data sets in terms of quality scores obtained on the plots, CASIA scores the highest, followed by IID and then UBIRIS.

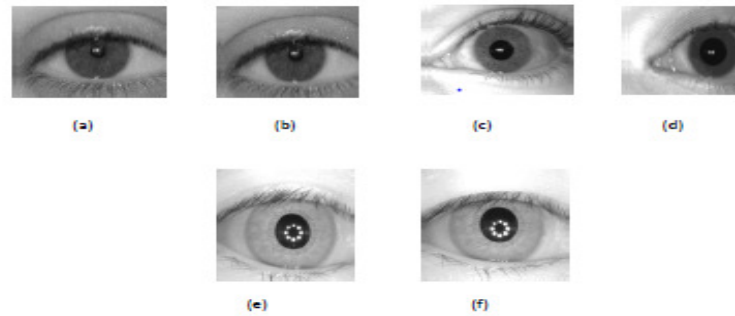


Figure 2 Sample eye images from UBIRIS, IID and CASIA Interval database. (a) - (b) UBIRIS. (c) - (d) IID. (e) - (f) CASIA.

Table II ESTIMATED QUALITY PARAMETERS OF IMAGES IN FIG. 2

M_C	M_{STD}	M_O	M_S	M_A	M_D	M_B	Score
0.2755	0.9897	0.4137	0.0271	0.0729	0.5182	0.3242	0.8089
0.2841	0.7013	0.0316	0.0502	0.0636	0.4352	0.4134	0.0209
0.3375	0.8273	0.0707	0.0052	0.6773	0.3714	0.3103	0.8930
0.3372	0.9810	0.3574	0.0299	0.5958	0.4222	0.285	0.3471
0.2535	0.7456	0.2927	0.0252	0.6689	0.3818	0.3212	0.8339
0.2718	0.8113	0.7552	0.0711	0.0229	0.4286	0.2924	0.5768

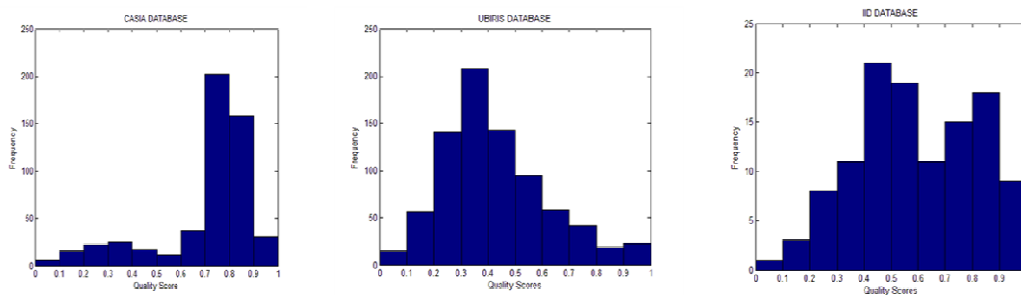


Figure 3 Overall Quality Distribution of CASIA, UBIRIS and IID Databases

8. CLASSIFICATION PERFORMANCE

The performance of the biometric system is typically characterized by error rate. To evaluate the separation performance of the good and bad images of the proposed quality metric, a k-fold cross-

validation technique is employed in a Support Vector Machine (SVM) classifier. SVM is a classifier that performs classification by constructing a hyperplane in a multidimensional space and separating the data points into different classes [20]. K-fold cross-validation is a rotational estimation of which dataset is randomly grouped into k mutually exclusive folds of approximately equal size [21]. Data is divided into k groups and each group is rotated between being a training group and a testing group. In this research $k = 10$ and then the correct rate of the classifier is averaged out. The performance of the assessment is illustrated in Table III. From these results it is clear that the proposed assessment algorithm is significant as the correct rate is above 80 % on both classifiers.

Table III SUMMARY OF PERFORMANCE STATISTICS USING SVM

Database	Correct Rate	Error Rate
CASIA	99.05	0.95
UBIRIS	98.75	1.25
IID	84.48	15.52

9. CONFUSION MATRIX

In CASIA database out of 525 images only 5 were classified wrongly giving an overall average total of 99.05 % accuracy in performance of the classifier. For UBIRIS database none of the images were classified as bad quality while they were actually good quality and 10 images were classified as good quality while they were bad, giving an overall average total of 98.75 % in classifier accuracy. Last, for IID database out of 116 images only 18 images were classified wrongly giving an overall average total of 84.48 % in accuracy. Table IV, V and VI illustrates these results.

In this research, there is no ground truth for all databases used, so human inspection was used to classify the images. However, humans have limited resolution, so they cannot detect quality of the images pixel by pixel. Moreover, humans perceive quality with limited factors such as clearness of features, blurriness and brightness of the image. On the other hand, the proposed algorithm determines quality based on standard deviation, contrast, dilation, blur, sharpness, area ratio and occlusion by calculating each factor pixel by pixel. Therefore, the proposed algorithm can calculate quality better than the human eye hence the misclassification.

Table IV CONFUSION MATRIX FOR CASIA

		Actual class	
		TP	TN
Predicted class	TP	69	0
	TN	5	451

Table V CONFUSION MATRIX FOR UBIRIS

		Actual class	
		TP	TN
Predicted class	TP	675	10
	TN	0	116

Table VI CONFUSION MATRIX FOR IID

Predicted class	Actual class		
	TP	TP	TN
	TN	14	63

10. PREDICTION PERFORMANCE OF PROPOSED ASSESSMENT METHOD

Fig. 4, 5 and 6 illustrate the prediction performance of the proposed quality assessment algorithm using the CASIA, UBIRIS and IID databases. The false positive rate of all three databases is low, implying that fewer images were misclassified than correctly classified. Table VII contains the statistics of the ROC curve analysis. The area under the curve for all databases range between 92 % to 97 %, which indicates a good performance of the classifier. Moreover, the 95% confidence interval (CI) for all databases are fairly high, with UBIRIS having the lowest lower bound of 0.88469, which implies the performance of the classifier is good. This concludes that the proposed algorithm is capable of distinguishing a good sample quality from a bad one. It can also be observed that the proposed quality assessment method can predict the quality of the image for all three databases as the AUC of the above ROC curves is above 90 %. These results imply that the proposed fused quality measure is suitable to be used as the informal measure for ensuring images of sufficient quality are used for feature extraction.

Table VII STATISTICS OF ROC CURVE ANALYSIS

Database	S. E.	AUC	C.I.
CASIA	0.00979	0.92505	0.92505 0.96342
UBIRIS	0.01127	0.96707	0.94499 0.98915
IID	0.02299	0.92975	0.88469 0.97480

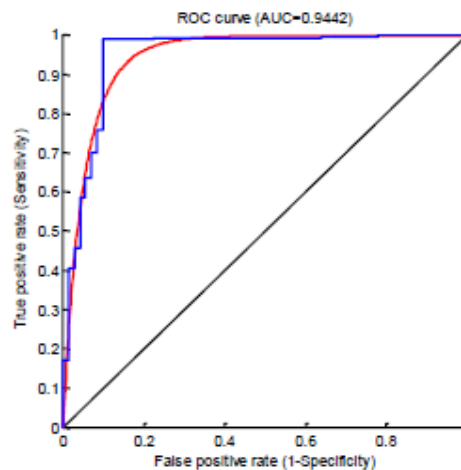


Figure 4 Verification performance of CASIA Database

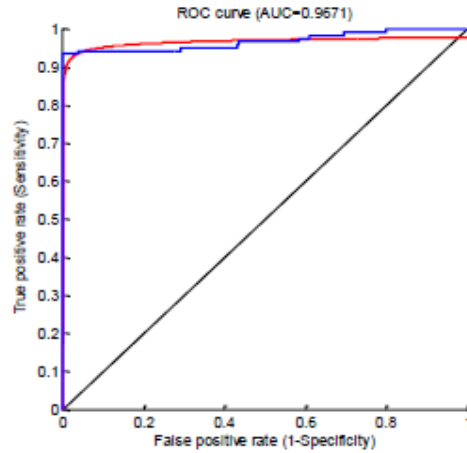


Figure 5 Verification performance of UBIRIS Database

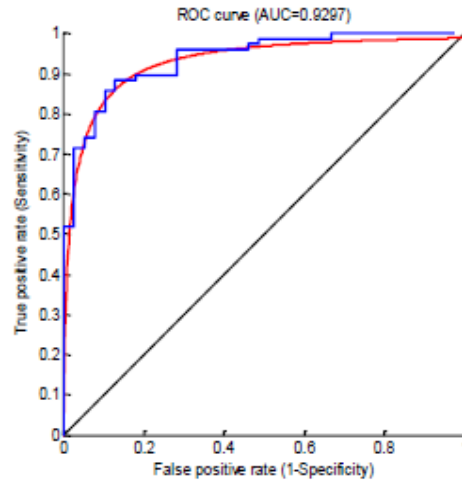


Figure 6 Verification performance of IID Database

11. CONCLUSION

In order to guide the selection of image of good quality, a quality model that evaluates the results of segmented iris image based on richness of the texture, shape and amount of information in the iris image has been developed. We extend iris quality assessment research by analysing the effect of various quality parameters such as standard deviation, contrast, area ratio, occlusion, blur, dilation and sharpness of an iris image. A fusion approach is presented for fusing all quality measures to a quality score. This is necessary because in order to successfully identify an individual on iris recognition systems an iris image must have sufficient features for extraction. The aim of this paper is to present a method that could be used for selection of high quality images, which may improve iris recognition performance. In analysing results the proposed assessment method proved to be capable of quality characterisation as it yields above 84 % in CR. The major benefit of this paper is that assessment is done before feature extraction, so only high quality images will be processed therefore saving time and resources.

ACKNOWLEDGEMENTS

The authors wish to thank CSIR Modelling and Digital Science for sponsoring this research, without your support I would not be able to accomplish my studies. Also, I want to acknowledge Tendani Malumedzha and Prof Fulufhelo Nelwamondo for their patience, wisdom and guidance. Thank you.

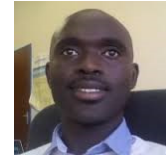
REFERENCES

- [1] Gulmire, K. and Ganorkar, S., (2012), "Iris recognition using Gabor wavelet." *International Journal of Engineering*, Vol. 1, No. 5.
- [2] Masek, L., "Recognition of human iris patterns for biometric identification." PhD thesis.
- [3] Ma, L., Tan, T., Wang, Y. and Zhang, D., (2003) "Personal identification based on iris texture analysis." *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, Vol. 25, No. 12, pp 1519–1533.
- [4] Daugman, J., (2004), "How iris recognition works." *Circuits and Systems for Video Technology*, IEEE Transactions on, Vol. 14, No. 1, pp 21–30.
- [5] Belcher, C., and Du, Y. (2008), "A selective feature information approach for iris image-quality measure". *Information Forensics and Security*, IEEE Transactions on, pp572–577.
- [6] Tabassi, E., (2009), "Biometric Quality Standards" , NIST, Biometric Consortium.,
- [7] Fatukasi, O., Kittler, J., and Poh, N., (2007), "Quality controlled multi-modal fusion of biometric experts.", In *Progress in Pattern Recognition, Image Analysis and Applications*, pp 881–890. Springer.
- [8] Kalka, N. D., Dorairaj, V., Shah, Y. N., Schmid, N. A. and Cukic B., (2002), "Image quality assessment for iris biometric." , In *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation*, pp 445–452. Springer.
- [9] Makinana, S., Malumedzha, T., Nelwamondo, F.V., (2014) "Iris Image Quality Assessment Based on Quality Parameters", *Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems Part I Lecture Notes in Artificial Intelligence*, pp571–580. Springer,
- [10] Kalka, N. D. and Zuo, J. and Schmid, N. A. and Cukic, B., (2006), "Image quality assessment for iris biometric", *Defense and Security Symposium*, International Society for Optics and Photonics, pp62020D–62020D
- [11] Crete, F., Dolmiere, T., Ladret, P. and Nicolas, M., (2007), "The blur effect: perception and estimation with a new no-reference perceptual blur metric.", *Human Vision and Electronic Image in XII*, pp6492:64920I.
- [12] Sandre, S-L and Stevens, M. and Mappes, J., (2010), "The effect of predator appetite, prey warning coloration and luminance on predator foraging decisions", *Behaviour*, vol.147, No. 9., 1121–1143, BRILL.
- [13] Du, Y. and Belcher, C. and Zhou, Z. and Ives, R., (2010), "Feature correlation evaluation approach for iris feature quality measure", *Signal processing*, Vol. 90, No. 4, pp1176–1187, Elsevier.
- [14] Nill, N. B., (2007), "IQF (Image Quality of Fingerprint) Software Application," The MITRE Corporation,
- [15] Bieroza, M. and Baker, A. and Bridgeman, J., (2011), "Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment", *Environmetrics*, Vol. 22, No.3, pp256–270, Wiley Online Library.
- [16] Jeong, D. H. and Ziemkiewicz, C. and Ribarsky, W. and Chang, R. and Center, C. V., (2009), "Understanding Principal Component Analysis Using a Visual Analytics Tool," *Charlotte Visualization Center*, UNC Charlotte, 2009
- [17] Suhr, D. D., (2005), "Principal component analysis vs. exploratory factor analysis," *SUGI 30 Proceedings*, pp 203–230.
- [18] Proena, H. and Alexandre, L.A., (2005), "UBIRIS: A noisy iris image database," *International Conference on Image Analysis and Processing*.
- [19] Chinese Academy of Sciences Institute of Automation., (2012), "CASIA Iris Database, Online:" <http://http://biometrics.idealtest.org/dbDetailForUser.do?id=4>.

- [20] Fauvel, M. and Benediktsson, J. A. and Chanussot, J. and Sveinsson, J. R., (2008), “Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles”, *Geoscience and Remote Sensing IEEE Transactions on*, vol. 46, No. 11, pp3804–3814.
- [21] Kohavi, R., (1995), “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *International Joint Conferences on Artificial Intelligence*, Vol 14, No. 2, pp1137–1145.

AUTHORS

Sisanda Makinana is a Biometrics Researcher at the Council for Scientific and Industrial Research (CSIR), South Africa under Information Security. She holds an MSc in Electrical Engineering from the University of Johannesburg.



Tendani Malumedzha is a Systems Engineer for Information Security at the Council for Scientific and Industrial Research (CSIR), South Africa. He holds an MSc in Electrical Engineering from the University of the Witwatersrand.



Fulufhelo Nelwamondo is a Competency Area Manager for Information Security at the Council for Scientific and Industrial Research (CSIR), South Africa. He holds a PhD in Electrical Engineering from the University of the Witwatersrand and is a visiting professor of Electrical Engineering at the University of Johannesburg.



INTENTIONAL BLANK

APPLICATION OF RHETORICAL RELATIONS BETWEEN SENTENCES TO CLUSTER-BASED TEXT SUMMARIZATION

N. Adilah Hanin Zahri¹, Fumiyo Fukumoto², Matsyoshi Suguru²
and Ong Bi Lynn¹

¹School of Computer and Communication,
University of Malaysia Perlis, Perlis, Malaysia
adilahhanin, drlynn@unimap.edu.my

²Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Yamanashi, Japan
fukumoto@yamanashi.ac.jp, suguru@yamanashi.ac.jp

ABSTRACT

Many of previous research have proven that the usage of rhetorical relations is capable to enhance many applications such as text summarization, question answering and natural language generation. This work proposes an approach that expands the benefit of rhetorical relations to address redundancy problem in text summarization. We first examined and redefined the type of rhetorical relations that is useful to retrieve sentences with identical content and performed the identification of those relations using SVMs. By exploiting the rhetorical relations exist between sentences, we generate clusters of similar sentences from document sets. Then, cluster-based text summarization is performed using Conditional Markov Random Walk Model to measure the saliency scores of candidates summary. We evaluated our method by measuring the cohesion and separation of the clusters and ROUGE score of generated summaries. The experimental result shows that our method performed well which shows promising potential of applying rhetorical relation in cluster-based text summarization.

KEYWORDS

Rhetorical Relations, Text Clustering, Extractive Text Summarization, Support Vector Machine, Probability Model, Markov Random Walk Model

1. INTRODUCTION

The study on rhetorical relations between sentences has been introduced in the late 80's to analyze, understand, and generate natural human-languages. Rhetorical relations hold sentences or phrases in a coherent discourse and indicate the informative relations regarding an event. In general, the rhetorical relations hold primarily between adjacent components with lexical elements. Rhetorical relations are defined functionally, in terms of the effect the writer intends to achieve by presenting two text spans. Up until now, researchers have developed several structures to describe the semantic relations between words, phrases and sentences. Some of the well-known structures are Rhetorical Structure Theory (RST) [1], RST Treebank [2], Lexicalized Tree-Adjoining Grammar based discourse [3], Cross-document Structure Theory (CST) [4][5] and Discourse GraphBank[6]. Each work proposed different kind of methods to distinguish how

events in text are related by identifying the transition point of a relation from one text span to another. Here, similar to the TDT project, an event refers to something that occurs at a specific place and time associated with some specific actions. In many structures, rhetorical relations is defined by the effect of the relations, and also by different constraints that must be satisfied in order to achieve this effect, and these are specified using a mixture of propositional and intentional language. For instance, in RST structure, the *Motivation* relation specifies that one of the spans presents an action to be performed by the reader; the *Evidence* relation indicates an event (claim), which describes the information to increase the reader's belief of why the event occurred [2]. Rhetorical relations also describe the reference to the propositional content of spans and which span is more central to the writer's purposes.

Therefore, the interpretation of how the phrases, clauses, and texts are semantically related to each other described by rhetorical relations is crucial to retrieve important information from text spans. These coherent structures have benefit various NLP applications such as text summarization [7][8][9][10][11][12], question answering [13][14] and natural language generation [15][16]. For instance, Litkowski proposed an approach that makes use of structural information of sentences, such as the discourse entities and semantic relation to generate database for question answering system [13]. In text summarization, discourse relations are used to produce optimum ordering of sentences in a document and remove redundancy from generated summaries. Our work focused on this area where we exploited the structure of rhetorical relations among sentences in multi-document text summarization.

Text summarization is the process of automatically creating a summary that retains only the relevant information of the original document. Generating summary includes identifying the most important pieces of information from the document, omitting irrelevant information and minimizing details. Automatic document summarization has become an important research area in natural language processing (NLP), due to the accelerating rate of data growth on the Internet. Text summarization limits the need for user to access the original documents and improves the efficiency of the information search. Our work focused on extractive summarization in multiple documents, which is finding the most salient sentences for the overall understanding of a given document. The task becomes tougher to accomplish as the system also has to deal with multi-document phenomena, such as paraphrasing and overlaps, caused by repeated similar information in the document sets. In this work, we make use of the rhetorical relations to improve the retrieval of salient sentences and redundancy elimination. We first examined and investigated the definition of rhetorical relations from existed structure, Cross-document Structure Theory (CST) [4][5]. We then redefined the rhetorical relations between sentences in order to perform an automated identification of rhetorical relations using machine learning technique, SVMs. We examined the surface features, *i.e.* the lexical and syntactic features of the text spans to identify characteristics of each rhetorical relation and provide them to SVMs for learning and classification module. We extended our work to the application of rhetorical relations in text clustering and text summarization. The next section provides an overview of the existing techniques. Section 3 describes the basic idea and methodology of our system. Finally, we report experimental result with some discussion.

2. PREVIOUS WORK

Previous work has shown many attempts to construct coherent structures in order to examine how the phrases, clauses, and texts are connected to each other [1][2][3][4][5][6]. In accordance with the development of various coherent structures, there were also many works dedicated to explore the benefit of rhetorical/discourse relations in NLP applications, especially in multi-document text summarization [1][8][9][10][11][12] and question answering [13][14]. The earliest structure of rhetorical relation is defined by Rhetorical Structure Theory (RST) proposed in 1988 [1]. RST

describes a text as hierarchically divided units. The units are asymmetrically related by a certain rhetorical relations that usually consist of a nucleus and satellites. A nucleus refers to the claim or information given regarding an event, while satellites refer to the evidence that supports the claim. RST has been developed into more than 20 definitions of rhetorical relations to describe structural patterns in text spans. On the other hand, Cross-document Structure Theory (CST) [4][5] attempts to describe the relationships exist between two or more sentences from multiple sources regarding the same topic. CST defines 18 types of rhetorical relations that accommodate the relations between sentences from multiple documents. The CST relationship are defined in term of relationship of the first sentence S_1 to the second sentence S_2 . For instance, *Equivalence* relation represents two text spans, S_1 and S_2 as having the same information content disregard the different word usage and sequences. Besides RST and CST, other well-known coherent structures are Lexicalized Tree-Adjoining Grammar Discourse [3], RST Treebank [2], and Discourse GraphBank [6]. Discourse GraphBank represents discourse relation as graph structure, while other works represent them as hierarchical structure between textual units. Each work proposed different kind of method to distinguish how events in text are semantically connected among the sentences.

Meanwhile, clustering of similar text refers to learning method of assigning a set of text into groups, known as clusters. Two or more text spans are considered belong to the same cluster if they are "close" according to a given similarity or distance. The clustering techniques are generally divided into partitioning [17][18], hierarchical [19][20] and graph-based clustering [21]. K-means [17][22][23] is an example of a simple partition based unsupervised clustering algorithm. The algorithm first defines the number of clusters, k to be created and randomly selects k sentences as the initial centroid of each cluster. All sentences are iteratively assigned to the closest cluster given the similarity distance between the sentence and the centroid and ends once all sentences are assigned and the centroid are fixed. Another most used partitioning clustering method is Fuzzy C-Means clustering [18][25]. Fuzzy C-means (FCM) is a method of clustering which allows sentences to be gathered into two or more clusters. This algorithm assigns membership level to each sentence corresponding to the similarity between the sentences and the centroid of the cluster. The closer the sentences to the centroid, the stronger the connection to the particular cluster. After each iteration, the membership grade and cluster center are updated. Other than K-Means and Fuzzy C-Means, hierarchical clustering is also widely used for text classification.

Text classification is one of many approach to multi-document text summarization. Multiple documents usually discuss more than one sub-topic regarding an event. Creating summary with wide diversity of each topic discussed in a multiple document is a challenging task for text summarization. Therefore, cluster-based approaches have been proposed to address this challenge. A cluster-based summarization groups the similar textual units into multiple clusters to identify themes of common information and candidates summary are extracted from these clusters [25][26][27]. Centroid based summarization method groups the sentences closest to the centroid in to a single cluster [9][28]. Since the centroid based summarization approach ranks sentences based on their similarity to the same centroid, the similar sentences often ranked closely to each other causing redundancy in final summary. In accordance to this problem, MMR [29] is proposed to remove redundancies and re-rank the sentences ordering. In contrast, the multi-cluster summarization approach divides the input set of text documents in to a number of clusters (sub-topics or themes) and representative of each cluster is selected to overcome redundancy issue [30]. Another work proposed a sentences-clustering algorithm, *SimFinder*[31][32] clusters sentences into several cluster referred as themes. The sentence clustering is performed according to linguistic features trained using a statistical decision [33]. Some work observed time order and text order during summary generation [34]. Other work focused on how clustering algorithm and representative object selection from clusters affects the multi-document summarization

performance [35]. The main issue raised in multi-cluster summarization is that the topic themes are usually not equally important. Thus, the sentences in an important theme cluster are considered more salient than the sentences in a trivial theme cluster. In accordance to this issue, previous work suggested two models, which are Cluster-based Conditional Markov Random Walk Model (Cluster-based CMRW) and Cluster-based HITS Model [36]. The Markov Random Walk Model (MRWM) has been successfully used for multi-document summarization by making use of the “voting” between sentences in the documents [37][38][39]. However, MRWM uniform use of the sentences in the document set without considering higher-level of information other than sentence-level information. Differ with former model, Cluster-based CMRW incorporates the cluster-level information into the link graph, meanwhile Cluster-based HITS Model considers the clusters and sentences as hubs and authorities. Wan and Yang considers the theme clusters as hubs and the sentences as authorities [36]. Furthermore, the coherent structure of rhetorical relations has been widely used to enhance the summary generation of multiple documents [40][41][42]. For instance, a paradigm of multi-document analysis, CST has been proposed as a basis approach to deal with multi-document phenomenon, such as redundancy and overlapping information during summary generation[8][9][10][11][12]. Many of CST based works proposed multi-document summarization guided by user preferences, such as summary length, type of information and chronological ordering of facts. One of the CST-based text summarization approaches is the incorporation of CST relations with MEAD summarizer [8]. This method proposes the enhancement of text summarization by replacing low-salience sentences with sentences that have maximum numbers of CST relationship in the final summary. They also observed the effect of different CST relationships against summary extraction. The most recent work is a deep knowledge approach system, CST-based SUMMarizer or known as CSTSumm [11]. Using CST-analyzed document, the system ranks input sentences according to the number of CST relations exist between sentences. Then, the content selection is performed according to the user preferences, and a multi-document summary is produced CSTSumm shows a great capability of producing informative summaries since the system deals better with multi-document phenomena, such as redundancy and contradiction.

3. FRAMEWORK

3.1. Redefinition of Rhetorical Relations

Our aim is to perform automated identification of rhetorical relations between sentences, and then apply the rhetorical relations to text clustering and summary generation. Since that previous works proposed various structure and definition of rhetorical relations, the structure that defines rhetorical relations between two text spans is mostly appropriate to achieve our objective. Therefore, we adopted the definition of rhetorical relation by CST [5] and examined them in order to select the relevant rhetorical relations for text summarization. According to the definition by CST, some of the relationship presents similar surface characteristics. Relations such as *Paraphrase*, *Modality* and *Attribution* share similar characteristic of information content with *Identity* except for the different version of event description. Consider the following examples:

Example 1

S_1 : Airbus has built more than 1,000 single-aisle 320-family planes.
 S_2 : It has built more than 1,000 single-aisle 320-family planes.

Example 2

S_3 : Ali Ahmedi, a spokesman for Gulf Air, said there was no indication the pilot was planning an emergency landing.

S_4 : *But Ali Ahmedi said there was no indication the pilot was anticipating an it emergency landing.*

Example 1 and 2 demonstrate an example of sentences pair that can be categorized as *Identity*, *Paraphrase*, *Modality* and *Attribution* relations. The similarity of lexical and information in each sentences pair is high, therefore these relations can be concluded as presenting the similar relation. We also discovered similarity between *Elaboration* and *Follow-up* relations defined by CST. Consider the following example:

Example 3

S_5 : *The crash put a hole in the 25th floor of the Pirelli building, and smoke was seen pouring from the opening.*

S_6 : *A small plane crashed into the 25th floor of a skyscraper in downtown Milan today.*

Example 3 shows that both sentences can be categorized as *Elaboration* and *Follow-up*, where S_5 describes additional information since event in S_6 occurred. Another example of rhetorical relations that share similar pattern is *Subsumption* and *Elaboration*, as shown in Example 4 and Example 5, respectively.

Example 4

S_7 : *Police were trying to keep people away, and many ambulances were at the scene.* S_8 : *Police and ambulance were at the scene.*

Example 5

S_9 : *The building houses government offices and is next to the city's central train station.*

S_{10} : *The building houses the regional government offices, authorities said.*

S_7 contains additional information of S_8 in Example 4, hence describes that sentences pair connected as *Subsumption* can also be defined as *Elaboration*. However, the sentences pair belongs to *Elaboration* in Example 5 cannot be defined as *Subsumption*. The definition of *Subsumption* denotes the second sentence as the subset of the first sentence, however, in *Elaboration*, the second sentence is not necessary a subset of the first sentence. Therefore, we keep *Subsumption* and *Elaboration* as two different relations so that we can precisely perform the automated identification of both relations.

We redefined the definition of the rhetorical relations adopted from CST, and combined the relations that resemble each other which have been suggested in our previous work [43]. *Fulfillment* relation refers to sentence pair which asserts the occurrence of predicted event, where overlapped information present in both sentences. Therefore, we considered *Fulfillment* and *Overlap* as one type of relation. As for *Change of Perspective*, *Contradiction* and *Reader Profile*, these relations generally refer to sentence pairs presenting different information regarding the same subject. Thus, we simply merged these relations as one group. We also combined *Description* and *Historical Background*, as both type of relations provide description (historical or present) of an event. We combined similar relations as one type and redefine these combined relations. Rhetorical relations and their taxonomy used in this work is concluded in Table 1.

Table 1. Type and definition of rhetorical relations adopted from CST.

Relations by CST	Proposed Relations	Definition of Proposed Relation
Identity, Paraphrase, Modality, Attribution	Identity	Two text spans have the same information content
Subsumption, Indirect Speech, Citation	Subsumption	S_1 contains all information in S_2 , plus other additional information not in S_2
Elaboration, Follow-up	Elaboration	S_1 elaborates or provide more information given generally in S_2 .
Overlap, Fullfillment	Overlap	S_1 provides facts X and Y while S_2 provides facts X and Z; X, Y, and Z should all be non-trivial
Change of Perspective, Contradiction, Reader Profile	Change of Topics	S_1 and S_2 provide different facts about the same entity.
Description, Historical Background	Description	S_1 gives historical context or describes an entity mentioned in S_2 .
-	No Relations	No relation exists between S_1 and S_2 .

By definition, although *Change of Topics* and *Description* does not accommodate the purpose of text clustering, we still included these relations for evaluation. We also added *No Relation* to the type of relations used in this work. We combined the 18 types of relations by CST into 7 types, which we assumed that it is enough to evaluate the potential of rhetorical relation in cluster-based text summarization.

3.2. Identification of Rhetorical Relations

We used a machine learning approach, Support Vector Machine (SVMs)[44] which have been proposed by our previous work [43] to classify type of relations exist between each sentence pairs in corpus. We used CST-annotated sentences pair obtained from CST Bank [5] as training data for the SVMs. Each data is classified into one of two classes, where we defined the value of the features to be 0 or 1. Features with more than 2 value will be normalized into [0,1] range. This value will be represented by 10 dimensional space of a 2 value vector, where the value will be divided into 10 value range of [0.0,0.1], [0.1,0.2], ..., [0.9,1.0]. For example, if the feature of text span S_j is 0.45, the surface features vector will be set into 0001000000. We extracted 2 types of surface characteristic from both sentences, which are lexical similarity between sentences and the sentence properties. Although the similarity of information between sentences can be determined only with lexical similarity, we also included sentences properties as features to emphasis which sentences provide specific information, *e.g.* location and time of the event. We provided the surface characteristics to SVMs for learning and classification of the text span S_i according to the given text span S_j .

3.2.1 Lexical Similarity between Sentences

We used 4 similarity measurements to measure the amount of overlapping information among sentences. Each measurement computes similarity between sentences from different aspects.

1. Cosine Similarity

Cosine similarity measurement is defined as follows:

$$\cos(S_1, S_2) = \frac{\sum_i (s_{1,i} \times s_{2,i})}{\sqrt{\sum_i (s_{1,i})^2} \times \sqrt{\sum_i (s_{2,i})^2}}$$

where S_1 and S_2 represents the frequency vector of the sentence pair, S_1 and S_2 , respectively. The cosine similarity metric measures the correlation between the two sentences according to frequency vector of words in both sentences. We observed the similarity of word contents, verb tokens, adjective tokens and bigram words from each sentences pair. The cosine similarity of bigram s is measured to determine the similarity of word sequence in sentences. The words ordering indirectly determine the semantic meaning in sentences.

2. Overlap ratio of words from S_1 in S_2 , and vice versa

The overlap ratio is measured to identify whether all the words in S_2 are also appear in S_1 , and vice versa. This measurement will determine how much the sentences match with each other. For instance, given the sentences pair with relations of *Subsumption*, the ratio of words from S_2 appear in S_1 will be higher than the ratio of words from S_1 appear in S_2 . We add this measurement because cosine similarity does not extract this characteristic from sentences. The overlap ratio is measured as follows:

$$WOL(S_1) = \frac{\#commonword(S_1, S_2)}{words(S_1)}$$

where “*#commonword*” and “*#words*” represent the number of matching words and the number of words in a sentence, respectively. The feature with higher overlap ratio is set to 1, and 0 for lower value. We measured the overlap ratio against both S_1 and S_2 .

3. Longest Common Substring

Longest Common Substring metric retrieves the maximum length of matching word sequence against S_1 , given two text span, S_1 and S_2 .

$$LCS(S_1) = \frac{len(MaxComSubstring(S_1, S_2))}{length(S_1)}$$

The metric value shows if both sentences are using the same phrase or term, which will benefit the identification of *Overlap* or *Subsumption*.

4. Ratio overlap of grammatical relationship for S_1

We used a broad-coverage parser of English language, MINIPAR [45] to parse S_1 and S_2 , and extract the grammatical relationship between words in the text span. Here we extracted the number of *surface subject* and the *subject of verb (subject)* and *object of verbs*

(*object*). We then compared the grammatical relationship in S_1 which occur in S_2 , compute as follows:

$$SubjOve(S_1) = \frac{\#commonSubj(S_1, S_2)}{Subj(S_1)}$$

$$ObjOve(S_1) = \frac{\#commonObj(S_1, S_2)}{Obj(S_1)}$$

The ratio value describes whether S_2 provides information regarding the same entity of S_1 , i.e. *Change of Topics*. We also compared the *subject* in S_1 with *noun* of S_2 to examine if S_1 is discussing topics about S_2 .

$$SubjNounOve(S_1) = \frac{\#commonSubj(S_1)Noun(S_2)}{Obj(S_1)}$$

The ratio value will show if S_1 is describing information regarding subject mention in S_2 , i.e. *Description*.

3.2.2 Sentences Properties

The type of information described in two text spans is also crucial to classify the type of discourse relation. Thus, we extracted the following information as additional features for each relation.

1. Number of entities

Sentences describing an event often offer information such as the place where the event occurs (location), the party involves (person, organization or subject), or when the event takes place (time and date). The occurrences of such entities can indicate how informative the sentence can be, thus can enhance the classification of relation between sentences. Therefore, we derived these entities from sentences, and compared the number of entities between them. We used Information Stanford NER (CRF Classifier: 2012 Version) of Named Entity Recognizer [46] to label sequence of words indicating 7 types of entities (*PERSON*, *ORGANIZATION*, *LOCATION*, *TIME*, *DATE*, *MONEY* and *PERCENT*). Based on the study of training data from CSTBank, there are no significant examples of annotated sentences indicates which entity points to any particular discourse relation. Therefore, in the experiment, we only observed the number of sentences entities in both text spans. The features with higher number of entities are set to 1, and 0 for lower value.

2. Number of conjunctions

We observed the occurrence of 40 types of conjunctions. We measured the number of conjunctions appear in both S_1 and S_2 . The feature with higher number of entities is set to 1, and 0 for lower value.

3. Lengths of sentences

We defined the length of S_j as follows:

$$Length(S_j) = \sum_i w_i$$

where w is the word appearing in the corresponding text span.

4. Type of Speech

We determined the type of speech, whether the text span, S_j cites another sentence by detecting the occurrence of quotation marks to identify *Citation* or *Indirect Speech* which are the sub-category of *Identity*.

3.3. Rhetorical Relation-based Text Clustering

The aim of this work is to expand the benefits of rhetorical relations between sentences to cluster-based text summarization. Rhetorical relation between sentences not only indicates how two sentences are connected to each other, but also shows the similarity patterns in both sentences. Therefore, by exploiting these characteristics, our idea is to construct similar text clustering based on rhetorical relations among sentences. We consider that *Identity*, *Subsumption*, *Elaboration* and *Overlap* relations are most appropriate for this task. These relations indicates either equivalence or partial overlapping information between text spans, as shown in Table 1. Connections between two sentences can be represented by multiple rhetorical relations. For instance, in some cases, sentences defined as *Subsumption* can also be define as *Identity*. Applying the same process against the same sentence pairs will be redundant. Therefore to reduce redundancy, we assigned the strongest relation to represent each connection between 2 sentences according to the following order:

- (i) whether both sentences are identical or not
- (ii) whether one sentence includes another
- (iii) whether both sentences share partial information
- (iv) whether both sentences share the same subject of topic
- (v) whether one sentence discusses any entity mentioned in another

The priority of the discourse relations assignment can be concluded as follows:

$$Identity > Subsumption > Elaboration > Overlap$$

We then performed clustering algorithm to construct groups of similar sentences. The algorithm is summarized as follows:

- i) The strongest relations determined by SVMs is assigned to each connection (refer to Figure 1(a)).
- ii) Suppose each sentence is a centroid of its own cluster. Sentences connected to the centroid as *Identity* (*ID*), *Subsumption* (*SUB*), *Elaboration* (*ELA*) and *Overlap* (*OVE*) relations¹ is identified and sentences with these connections are evaluated as having similar content, and aggregated as one cluster (refer Figure 1(b)).
- iii) Similar clusters is removed by retrieving centroids connected as *Identity*, *Subsumption* or *Elaboration*.
- iv) Clusters from (iii) is merged to minimize the occurrence of the same sentences in multiple clusters (refer Figure 1(c)).
- v) Step (iii) and (iv) are iterated until the number of clusters is convergence

We performed 2 types of text clustering, which are:

- i) *RRCluster 1*, which consist of *Identity* (*ID*), *Subsumption* (*SUB*), *Elaboration* (*ELA*) and *Overlap* (*OVE*)
- ii) *RRCluster2*, which consist of *Identity* (*ID*), *Subsumption* (*SUB*) and *Elaboration* (*ELA*)

The algorithm of similar text clustering is illustrated in Figure 1.

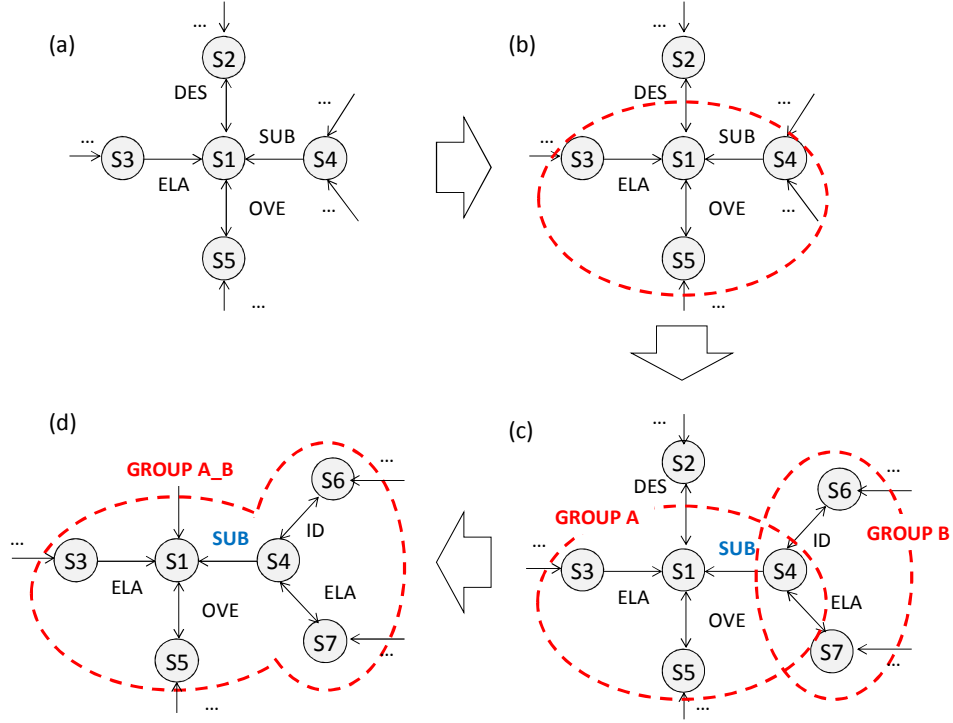


Figure 1. Rhetorical relation-based clustering algorithm

3.4. Cluster-based Summary Generation

We performed a cluster-based text summarization using clusters of similar text constructed by exploiting rhetorical relations between sentences. We used Cluster-based Conditional Markov Random Walk Model [36] to measure the saliency scores of candidates summary. Here we defined the centroid as relevant candidate summary since each centroid represents the whole cluster. The Conditional Markov Random Walk Model is based on the two-layer link graph including both the sentences and the clusters. Therefore, the presentation of the two layer graph are is denoted as $G^* = \langle V_s, V_c, E_{ss}, E_{sc} \rangle$. Suppose $V_s = V = v_i$ is the set of sentences and

$V_c = C = c_j$ is the set of hidden nodes representing the detected theme clusters, where

$E_{ss} = E = e_{ij} \mid v_i \in V_s$ corresponds to all links between sentences

$E_{sc} = e_{ij} \mid v_i \in V_s, c_j \in V_c, c_j = \text{clus}(v_i)$ corresponds to the correlation between a sentence and its cluster. The score is computed measured as follows:

$$\text{SenScore} = \mu \cdot \sum_{\text{all } j \neq i} \text{SenScore}(v_j) \cdot \tilde{M}_{ij,i}^* + \frac{(1-\mu)}{|V|}$$

μ is the damping factor set to 0.85, as defined in the PageRank algorithm. $\tilde{M}^*_{j,i}$ refers to row-normalized matrix $\tilde{M}^*_{j,i} = (\tilde{M}^*_{j,i})_{|V| \times |V|}$ to describe \tilde{G}^* with each entry corresponding to the transition probability, shown as follows:

$$\tilde{M}^*_{ij,i} = p(i \rightarrow j \mid \text{clus}(v_i), \text{clus}(v_i))$$

Here, $\text{clus}(v_i)$ denotes the theme cluster containing sentence v_i . The two factors are combined into the transition probability from v_i to v_j defined as follows:

$$p(i \rightarrow j \mid \text{clus}(v_i), \text{clus}(v_i)) = \frac{f(i \rightarrow j \mid \text{clus}(v_i), \text{clus}(v_i))}{\sum_{k=1}^{|V|} f(i \rightarrow k \mid \text{clus}(v_i), \text{clus}(v_k))}, \text{ if } \sum f \neq 0$$

$f(i \rightarrow j \mid \text{clus}(v_i), \text{clus}(v_i))$ denotes the new affinity weight between two sentences v_i and v_j , where both sentences belong to the corresponding two clusters. The conditional affinity weight is computed by linearly combining the affinity weight conditioned on the source cluster, i.e. $f(i \rightarrow j \mid \text{clus}(v_i))$ and the affinity weight conditioned on the target cluster i.e. $f(i \rightarrow j \mid \text{clus}(v_j))$, defined in the following equation.

$$\begin{aligned} f(i \rightarrow j \mid \text{clus}(v_i), \text{clus}(v_i)) &= \lambda \cdot (f(i \rightarrow j \mid \text{clus}(v_i)) + (1 - \lambda) \cdot f(i \rightarrow j \mid \text{clus}(v_i))) \\ &= \lambda \cdot f(i \rightarrow j) \cdot \pi(\text{clus}(v_i)) \cdot \omega(v_i, \text{clus}(v_i)) \\ &\quad + (1 - \lambda) \cdot f(i \rightarrow j) \cdot \pi(\text{clus}(v_j)) \cdot \omega(v_j, \text{clus}(v_j)) \\ &= f(i \rightarrow j) \cdot (\lambda \cdot \pi(\text{clus}(v_i)) \cdot \omega(v_i, \text{clus}(v_i)) \\ &\quad + (1 - \lambda) \cdot \pi(\text{clus}(v_j)) \cdot \omega(v_j, \text{clus}(v_j))) \end{aligned}$$

Where $\lambda \in [0,1]$ is the combination of weight controlling the relative contributions from the source cluster and the target cluster¹. $\pi(\text{clus}(v_i)) \in [0,1]$ refers the importance of cluster $\text{clus}(v_i)$ in the whole document set D and $\omega(v_i, \text{clus}(v_i)) \in [0,1]$ denotes the strength of the correlation between sentence v_i and its cluster $\text{clus}(v_i)$. In this work, $\pi(\text{clus}(v_i))$ is set to the cosine similarity value between the cluster and the whole document set, computed as follows:

$$\pi(\text{clus}(v_i)) = \text{sim}_{\text{cosine}}(\text{clus}(v_i), D)$$

Meanwhile, $\omega(v_i, \text{clus}(v_i))$ is set to the cosine similarity value between the sentence and the cluster where the sentence belongs, computed as follows:

$$\pi(v_i, \text{clus}(v_i)) = \text{sim}_{\text{cosine}}(v_i, \text{clus}(v_i))$$

The saliency scores for the sentences are iteratively computed until certain threshold, θ is reached².

4. EXPERIMENT

4.1. Data

¹We set $\lambda = 0.5$ for fair evaluation with methods adopted from (Wan and Yang, 2008)

²In this study, the threshold, θ is set to 0.0001

CST-annotated sentences are obtained from Cross-document Structure Theory Bank (Radevet. *al*, 2004). Our system is evaluated using 2 data sets from Document Understanding Conference, which are DUC'2001 and DUC'2002 [47].

4.2. Result and Discussion

4.2.1 Identification of Rhetorical Relations

The rhetorical relations assigned by SVMs are manually evaluated by 2 human judges. Since no human annotation is available for DUC data sets, 5 times of random sampling consisting 100 sentence pairs is performed against each document set of DUC'2001 and DUC'2002). The human judges performed manual annotation against sentence pairs, and assessed if SVMs assigned the correct rhetorical relation to each pair. The correct rhetorical relation refers to either one of the relations assigned by human judges in case of multiple relations exist between the two sentences. As a baseline method, the most frequent relation in each set of sampling data is assigned to all sentence pairs. We evaluated the classification of rhetorical relations by measuring the Precision, Recall and F-measure score.

Table2 shows the macro average of Precision, Recall and F-measure for each data set. *Identity* shows the most significant performance of Precision, where the value achieved more than 90% in all data sets. Meanwhile, the Precision value for *Citation* and *Description* performed worse compared to others in most data sets. Evaluation result shows that sentence pairs with quotation marks mostly classified as *Citation*. As for Recall value, *Identity*, *Subsumption*, *Elaboration* and *Description* yield more than 80%, meanwhile *Change of Topic* and *No Relation* performed the worst with Recall of 60% in both data sets. We found that SVMs was unable to identify *Change of Topics*, when multiple subjects (especially contained personal pronoun) occurred in a sentence. According to F-Measure, SVMs performed well during the classification of *Identity*, *Subsumption* and *Elaboration* with the Precision values achieved are above 70% for most data set. Overall, compared to other relations, the *Identity* classification by SVMs performed the best in each evaluation metric as expected. Sentence pair with *Identity* relation shows significant resemblance in similarity value, grammatical relationship and number of entities. For instance, the similarity between sentence pair is likely close to 1.0, and there are major overlap in subject and the object of the sentences. *Citation*, *Subsumption* and *Elaboration* indicate promising potential of automated classification using SVMs with F-measure achieved higher than 70%. We observed that characteristics such as similarity between sentences, grammatical relationship and number of entities are enough to determine the type of rhetorical relation of most data sets. Therefore, we considered the ratio of rhetorical relations except for *No Relations* show a great potential for automated classification with small number of annotated sentences.

Table 2. Evaluation result for identification of rhetorical relations

Relations	DUC'2001			DUC'2002		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Baseline	0.875	0.114	0.201	0.739	0.108	0.188
Identity	0.980	1.000	0.989	0.849	1.000	0.917
Citation	0.583	1.000	0.734	0.617	1.000	0.763
Subsumption	0.721	0.984	0.830	0.685	0.900	0.773
Elaboration	0.664	0.952	0.778	0.652	0.901	0.743
Overlap	0.875	0.532	0.653	0.739	0.556	0.633
Change of Topics	0.591	0.709	0.640	0.618	0.589	0.597
Description	0.841	0.947	0.886	0.817	0.856	0.826
No Relations	1.000	0.476	0.632	0.966	0.475	0.628

We found that the lack of significant surface characteristic is the main reason of misclassification of relations such as *Citation*, *Overlap*, *Change of Topics* and *Description*. Therefore, we conducted further analysis using confusion matrix [48] to determine the accuracy of classification by SMVs. Confusion matrix compares the classification results by the system and actual class defined by human, which useful to identify the nature of the classification errors. Table 3 and 4 describe the evaluation result of DUC'2001 and DUC'2002, respectively. The analysis is done against each relation independently. Each table shows the classification nature of rhetorical relations according to the number of sentences pair. We also included the accuracy and reliability value of every relations. For instance, according to evaluation of DUC'2001 in Table 3, from 44 pairs of sentences with *Identity* relation, our system has been able to classify 43 pairs of them as *Identity* correctly, while 1 pair misclassified as *Subsumption*. As a result, the Accuracy and Reliability value achieved for *Identity* are 1.000 and 0.977, respectively.

Despite the errors discovered during the identification of rhetorical relations, the classification by SVMs shows a promising potential especially for *Identity*, *Subsumption*, *Elaboration* and *No Relation*. In future, the increment of annotated sentences with significant characteristics of each relation will improve the identification of rhetorical relation. For instance, in this experiment, *Overlap* refers to sentences pair that shares partial information with each other. Therefore, we used Bigram similarity and Longest Common Substring metric to measure the word sequences in sentences. However, these metrics caused sentences with long named entity, e.g. "President George Bush" and "Los Angeles", as having consecutive words which contributed to false positive result of *Overlap* relation. The increment of annotated sentences consists of consecutive common nouns and verbs will help to precisely define *Overlap* relation. Moreover, improvement such as the usage of lexical database to extract lexical chain and anaphora resolution tool can be used to extract more characteristics from each relation.

Table 3. Evaluation of Confusion Matrix for DUC'2001

		Classification by System								Accuracy
		ID	CIT	SUB	ELA	OVE	CHT	DES	NOR	
Actual Class	ID	43	0	0	0	0	0	0	0	1.000
	CIT	0	27	0	0	0	0	0	0	1.000
	SUB	1	0	61	0	0	0	0	0	0.984
	ELA	0	0	2	48	0	0	1	0	0.941
	OVE	0	20	3	12	57	3	2	0	0.533
	CHT	0	0	5	6	6	51	3	0	0.718
	DES	0	0	0	0	0	2	59	0	0.967
	NOR	0	0	3	5	3	30	2	35	0.449
Reliability		0.977	0.574	0.726	0.676	0.864	0.593	0.881	1.000	

Table 4. Evaluation of Confusion Matrix for DUC'2002

		Classification by System								Accuracy
		ID	CIT	SUB	ELA	OVE	CHT	DES	NOR	
Actual Class	ID	55	0	0	0	0	0	0	0	1.000
	CIT	0	31	0	0	0	0	0	0	1.000
	SUB	6	0	51	0	0	0	0	0	0.895
	ELA	0	0	4	35	0	0	0	0	0.897
	OVE	2	19	12	6	54	2	2	0	0.557
	CHT	1	0	4	9	10	40	2	1	0.597
	DES	0	0	0	0	0	8	70	0	0.886
	NOR	0	0	3	6	10	13	7	36	0.480
Reliability		0.859	0.620	0.689	0.614	0.730	0.635	0.864	0.973	

Table 5. Evaluation result for cohesion and separation of clusters

Data Set	Evaluation	Clustering Method		
		K-Means	RRCluster1 (ID,SUB,ELA,OVE)	RRCluster2 (ID, SUB, ELA)
DUC'2001	Average SSE	7.271	4.599	4.181
	Average SSB	209.111	397.237	308.153
	Average SC	0.512	0.652	0.628
DUC'2002	Average SSE	6.991	3.927	3.624
	Average SSB	154.511	257.118	214.762
	Average SC	0.510	0.636	0.639

4.2.2 Rhetorical Relation-based Clustering

We evaluated our method by measuring the cohesion and separation of the constructed clusters. The cluster cohesion refers to how closely the sentences are related within a cluster, measured using Sum of Squared Errors (SSE) [49]. The smaller value of SSE indicates that the sentences in clusters are closer to each other. Meanwhile, Sum of Squares Between (SSB) [49] is used to measure cluster separation in order to examine how distinct or well-separated a cluster from others. The high value of SSB indicates that the sentences are well separated with each other. Cosine similarity measurement is used to measure the similarity between sentences in both SSE and SSB evaluation. We also obtained the average of Silhouette Coefficient (SC) value to measure the harmonic mean of both cohesion and separation of the clusters [49][50]. The value range of the Silhouette Coefficient is between 0 and 1, where the value closer to 1 is the better.

Table 5 shows the evaluation results for cohesion and separation of the clusters. *RRCluster1* refers to the clusters constructed by *Identity*, *Subsumption* and *Elaboration*, while *RRCluster2* refers to the clusters constructed by *Identity*, *Subsumption*, *Elaboration* and *Overlap*. We also used K-Means clustering for comparison [17]. K-means iteratively reassigns sentences to the closest clusters until a convergence criterion is met. Table 5 indicates that *RRCluster2*, which generates clusters of sentences with strong connections *Identity*, *Subsumption* and *Elaboration*, demonstrates the best SSE value (4.181 for DUC'2001 and 3.624 for DUC'2002), which shows the most significant cohesion within clusters. In contrast, *RRCluster1* which includes *Overlap* during clustering indicates the most significant separation between clusters with the best SSB value (397.237 for DUC'2001 and 257.118 for DUC'2002). *RRCluster1* generated bigger clusters, therefore resulted wider separation from other clusters. The average Silhouette Coefficient shows that our method, *RRCluster1* (0.652 for DUC'2001 and 0.636 for DUC'2002) and *RRCluster2* (0.628 for DUC'2001 and 0.639 for DUC'2002) outranked K-Means (0.512 for DUC'2001 and 0.510 for DUC'2002) for both data sets.

Table 6. Evaluation result for pair-wise

Data Set	Evaluation	Clustering Method		
		K-Means	RRCluster2 (ID, SUB, ELA)	RRCluster1 (ID, SUB, ELA, OVE)
DUC'2001	Precision	0.577	0.805	0.783
	Recall	0.898	0.590	0.758
	F-Measure	0.702	0.678	0.770
DUC'2002	Precision	0.603	0.750	0.779
	Recall	0.885	0.533	0.752
	F-Measure	0.716	0.623	0.766

In addition, we examined the clusters by performing a pair-wise evaluation. We sampled 5 sets of data consisting 100 sentences pairs and evaluated if both sentences are actually belong to the same clusters. Table 6 shows the macro average Precision, Recall and F-measure for pair-wise evaluation. *RRCluster2*, which excludes *Overlap* relation during clustering, demonstrated a lower Recall value compared to *RRCluster1* and K-Means. However, the Precision score of *RRCluster2* indicates better performance compared to K-Means. Overall, *RRCluster1* obtained the best value for all measurement compared to *RRCluster2* and K-Means for both data sets. We achieved optimum pair-wise results by including *Overlap* during clustering, where the F-measure obtained for DUC'2001 and DUC'2002 are 0.770 and 0.766, respectively.

We made more detailed comparison between clusters constructed by K-Means and our method. The example of the clustered sentences by each method from the experiment is shown in Table 7. K-Means is a lexical based clustering method, where sentences with similar lexical often be clustered as one group although the content semantically different. The 5th sentences from K-Means cluster in Table 7 demonstrates this error. Meanwhile, our system, *RRCluster1* and *RRCluster2* performed more strict method where not only lexical similarity, but also syntactic similarity, *i.e* the overlap of grammatical relationship is taken into account during clustering. According to Table 5, Table 6 and Table 7, the connection between sentences can allow text clustering according to the user preference. For instance, *RRCluster2* performed small group of similar sentences with strong cohesion in a cluster. In contrast, *RRCluster1* method performed clustering of sentences with *Identity*, *Subsumption*, *Elaboration* and *Overlap*, which are less strict than *RRCluster2*, however presents strong separation between clusters. In other words, the overlapping information between clusters are lower compared to *RRCluster2*. Thus, the experimental results demonstrate that the utilization of rhetorical relations can be another alternative of cluster construction other than only observing word distribution in corpus.

4.2.3 Cluster-based Summary Generation

We generated short summaries of 100 words for DUC'2001 and DUC'2002 to evaluate the performance of our clustering method, and to observe if rhetorical relation-based clustering benefits the multi-document text summarization. The experimental results also include the evaluation of summaries based on clusters generated by Agglomerative Clustering, Divisive Clustering and K-Means as comparison, adopted from [36]. The ROUGE-1 and ROUGE-2 score of clustering method shown in Table 8.

Table 7. Comparison of sentences from K-Means and proposed methods clusters

K-Means		
√	Centroid	<i>Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</i>
√	1	<i>Earlier Wednesday Gilbert was classified as a Category 5 storm, the strongest and deadliest type of hurricane.</i>
√	2	<i>Such storms have maximum sustained winds greater than 155 mph and can cause catastrophic damage.</i>
√	3	<i>As Gilbert moved away from the Yucatan Peninsula Wednesday night , the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again.</i>
√	4	<i>Only two Category 5 hurricanes have hit the United States the 1935 storm that killed 408 people in Florida and Hurricane Camille that devastated the Mississippi coast in 1969, killing 256 people.</i>
x	5	<i>"Any time you contract an air mass , they will start spinning . That's what makes the tornadoes , hurricanes and blizzards , those winter storms",Bleck said.</i>
RRcluster2		
√	Centroid	<i>Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</i>
√	1	<i>On Saturday , Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.</i>
√	2	<i>The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night.</i>
√	3	<i>It reached tropical storm status by Saturday and a hurricane Sunday.</i>
√	4	<i>Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</i>
RRcluster1		
√	Centroid	<i>Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</i>
√	1	<i>On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.</i>
√	2	<i>The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night.</i>
√	3	<i>Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs , uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel.</i>
√	4	<i>It reached tropical storm status by Saturday and a hurricane Sunday.</i>
√	5	<i>Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</i>

Table 8. Comparison of ROUGE score for DUC'2001 and DUC'2002

Method	DUC'2001		DUC'2002	
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
Agglomerative	0.3571	0.0655	0.3854	0.0865
Divisive	0.3555	0.0607	0.3799	0.0839
K-Means	0.3582	0.0646	0.3822	0.0832
RRcluster2	0.3359	0.0650	0.3591	0.0753
RRcluster1	0.3602	0.0736	0.3693	0.0873

For DUC'2001 data set, our *RRcluster1* performed significantly well for ROUGE-1 and ROUGE-2 score, where we outperformed others with highest score of 0.3602 and 0.0736, respectively. Divisive performed the worst compared to other methods. As for DUC'2002 data set, Agglomerative obtained the best score of ROUGE-1 with 0.3854, while *RRcluster2* yield the lowest score of 0.3591. In contrast, *RRcluster1* gained the best score of ROUGE-2 with 0.0873.

We observed that our proposed *RRCluster1* performed significantly well with ROUGE-2. During the classification of rhetorical relations, we also considered word sequence of Bigram to determine rhetorical relations, therefore resulted a high score of ROUGE-2. However, the ROUGE-1 score of our proposed methods performed poorly for DUC'2002 data sets, especially for *RRCluster2*. This technique, which considers *Identity*, *Subsumption* and *Elaboration* during text clustering certainly constructed clusters with high cohesion, but also limits the clustering to sentences with only strong connections. This led to the construction of many small clusters with possibility of partial overlaps of information with other clusters. As a result, the structure of clusters in *RRCluster2* caused the low value of both ROUGE-1 and ROUGE-2 scores.

Although our method only achieved good ROUGE-2 score, we considered that rhetorical relation-based clustering shows a great potential since that our clustering method is at initial stage yet already outperformed some of the well-established clustering method. Clearly, rhetorical relation-based cluster need some further improvement in future in order to produce better result. However, the result we obtained from this experiment shows that rhetorical relation-based clustering can enhance the cluster-based summary generation.

5. CONCLUSIONS

This paper investigated the relevance and benefits of the rhetorical relation for summary generation. We proposed the application of rhetorical relations exist between sentences to improve extractive summarization for multiple documents, which focused on the extraction of salient sentences and redundancy elimination. We first examined the rhetorical relations from Cross-document Theory Structure (CST), then selected and redefined the relations that benefits text summarization. We extracted surfaces features from annotated sentences obtained from CST Bank and performed identification of 8 types of rhetorical relations using SVMs. Then we further our work on rhetorical relations by exploiting the benefit of rhetorical relation to similar text clustering. The evaluation results showed that the rhetorical relation-based method has promising potential as a novel approach for text clustering. Next, we extended our work to cluster-based text summarization. We used ranking algorithm that take into account the cluster-level information, Cluster-based Conditional Markov Random Walk (Cluster-based CMRW) to measure the saliency score of sentences. For DUC'2001, our proposed method, *RRCluster1* performed significantly well for ROUGE-1 and ROUGE-2 score with highest score of 0.3602 and 0.0736, respectively. Meanwhile, *RRCluster1* gained the best score of ROUGE-2 with 0.0873 for DUC'2002. This work has proved our theory that rhetorical relations can benefit the similar text clustering. With further improvement, the quality of summary generation can be enhanced. From the evaluation results, we concluded that the rhetorical relations are effective to improve the ranking of salient sentences and the elimination of redundant sentences. Furthermore, our system does not rely on fully annotated corpus and does not require deep linguistic knowledge.

ACKNOWLEDGEMENTS

This research is supported by many individuals from multiple organization of University of Yamanashi, Japan and University of Perlis, Malaysia.

REFERENCES

- [1] Mann, W.C. and Thompson, S.A., "Rhetorical Structure Theory: Towards a Functional Theory of Text Organization", *Text*, 8(3), pp.243-281, 1988.
- [2] Carlson, L., Marcu, D. and Okurowski, M.E., "RST Discourse Treebank", *Linguistic Data Consortium* 1-58563-223-6, 2002.

- [3] Webber, B.L., Knott, A., Stone, M. and Joshi, A., "Anaphora and Discourse Structure", *Computational Linguistics* 29 (4), pp. 545–588, 2003.
- [4] Radev, D.R., "A Common Theory of Information Fusion from Multiple Text Source Step One: Cross-Document", In *Proc. of 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, 2000.
- [5] Radev, D.R., Otterbacher, J. and Zhang, Z., *CSTBank: Cross-document Structure Theory Bank*, <http://tangra.si.umich.edu/clair/CSTBank/phase1.htm>, 2003.
- [6] Wolf, F., Gibson, E., Fisher, A. and Knight, M., "DiscourseGraphbank", *Linguistic Data Consortium*, Philadelphia, 2005.
- [7] Marcu, D., "From Discourse Structures to Text Summaries", In *Proc. of the Association for Computational Linguistics (ACL) on Intelligent Scalable Text Summarization*, pp. 82-88, 1997.
- [8] Zhang, Z., Blair-Goldensohn, S. and Radev, D.R., "Towards CST-enhanced Summarization", In *Proc. of the 18th National Conference on Artificial Intelligence (AAAI)*, 2002.
- [9] Radev, D.R., Jing, H., Stys, M., Tam, D., "Centroid-based Summarization of Multiple Documents", *Information Processing and Management* 40, pp. 919–938, 2004.
- [10] Uzeda, V.R., Pardo, T.A.S., Nunes, M.G.V., "A Comprehensive Summary Informativeness Evaluation for RST-based Summarization Methods", *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)* ISSN: 2150-7988 Vol.1, pp.188-196, 2009.
- [11] Jorge, M.L.C and Pardo, T.S., "Experiments with CST-based Multi-document Summarization", *Workshop on Graph-based Methods for Natural Language Processing*, Association for Computational Linguistics (ACL), pp. 74-82, 2010.
- [12] Louis, A., Joshi, A., and Nenkova, A., "Discourse Indicators for Content Selection in Summarization", In *Proc. of 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 147-156, 2010.
- [13] Litkowski, K., "CL Research Experiments in TREC-10 Question Answering", *The 10th Text Retrieval Conference (TREC 2001)*. NIST Special Publication, pp. 200-250, 2002.
- [14] Verberne, S., Boves, L., and Oostdijk, N., "Discourse-based Answering of Why-Questions", *Traitement Automatique des Langues*, special issue on Computational Approaches to Discourse and Document Processing, pp. 21-41, 2007.
- [15] Theune, M., "Contrast in Concept-to-speech Generation", *Computer Speech and Language*, 16(3-4), ISSN 0885-2308, pp. 491-530, 2002.
- [16] Piwek, P. and Stoyanchev, S., "Generating Expository Dialogue from Monologue Motivation, Corpus and Preliminary Rules", In *Proc. of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.
- [17] McQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations", In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
- [18] Dunn, J.C., "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics*, pp. 32-57, 1973.
- [19] Johnson, S.C., "Hierarchical Clustering Schemes", *Psychometrika*, pp. 241-254, 1967.
- [20] D'andrade, R., "U-Statistic Hierarchical Clustering", *Psychometrika*, pp. 58-67, 1978.
- [21] Ng, A. Y., Jordan, M. I., and Weiss, Y., "On Spectral Clustering: Analysis and an Algorithm", In *Proc. of Advances in Neural Information Processing Systems (NIPS 14)*, 2002.
- [22] Hartigan, J. A., Wong, M. A., "Algorithm AS 136: A K-Means Clustering Algorithm", *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28 (1), pp. 100–108, 1979.
- [23] Hamerly, G. and Elkan, C., "Alternatives to the K-means Algorithm that Find Better Clusterings", In *Proc. of the 11th International Conference on Information and Knowledge Management (CIKM)*, 2002.
- [24] Bezdek, J.C., "Pattern Recognition with Fuzzy Objective Function Algorithms", *Plenum Press*, New York, 1981.
- [25] McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R. and Eskin, E., "Towards Multi-document Summarization by Reformulation: Progress and prospects", In *Proc. of the 16th National Conference of the American Association for Artificial Intelligence (AAAI)*, pp. 453-460, 1999.
- [26] Marcu, D., and Gerber, L., "An Inquiry into the Nature of Multidocument Abstracts, Extracts, and their Evaluation", In *Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, *Workshop on Automatic Summarization*, pp. 1-8, 2001.
- [27] Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Wise, G.B., and Zhang, X., "Cross-document Summarization by Concept Classification", In *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 121-128, 2002.

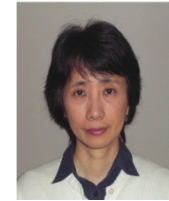
- [28] Radev, D.R., Jing, H., and Budzikowska, M., "Centroid-based Summarization of Multiple Documents: Sentence extraction, Utility-based Evaluation, and User Studies", In ANLP/NAACL Workshop on Summarization, 2000.
- [29] Carbonell, J.G. and Goldstein, J., "The Use of MMR, Diversity-based Re-ranking for Reordering Documents and Producing Summaries," In Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335-336, 1998.
- [30] Stein, G.C., Bagga, A. and Wise, G.B., "Multi-Document Summarization: Methodologies and Evaluations", In Conference TALN, 2000.
- [31] Hatzivassiloglou, V., Klavans, J., and Eskin, E., "Detecting Text Similarity Over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning", In Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP), 1999.
- [32] Hatzivassiloglou, V., Klavans, J., Holcombe, M.L., Barzilay, R., Kan, M.-Y., and McKeown, K.R., "SimFinder: A Flexible Clustering Tool for Summarization", In Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Workshop on Automatic Summarization, 2001.
- [33] Cohen, W., "Learning Trees and Rules with Set-valued Features", In Proc. of the 14th National Conference on Artificial Intelligence (AAAI), 1996.
- [34] Barzilay, R., Elhadad, N., and McKeown, R.K., "Sentence Ordering in Multi-document Summarization", In Proc. of the Human Language Technology Conference, K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA - International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, (ISSN 0974-3375), pp. 325-335, 2009.
- [35] Sarkar, K., "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA - International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, (ISSN 0974-3375), pp. 325-335, 2009.
- [36] Wan, X. and Yang, J., "Multi-Document Summarization Using Cluster-Based Link Analysis", In Proc. of the 31st Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR) Conference, pp. 299-306, 2008.
- [37] Erkan, G. and Radev, D.R., "LexPageRank: Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research 22, pp. 457-479, 2004.
- [38] Mihalcea, R., and Tarau, P., "A language Independent Algorithm for Single and Multiple Document Summarization", In Proc. of International Joint Conference on Natural Language Processing (IJCNLP), 2005.
- [39] Wan, X. and Yang, J., "Improved Affinity Graph based Multi-document Summarization", In Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL), 2006.
- [40] Otterbacher, J., Radev, D. and Luo, A., "Revisions that Improve Cohesion in Multidocument Summaries: A Preliminary Study", In Proc. of Conference on Association of Computer Linguistics (ACL), Workshop on Automatic Summarization, pp. 27-36, 2002.
- [41] Teufel, S. and Moens, M., "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Structure", Computational Linguistics 28(4): 409-445, 2002.
- [42] Pardo, T.A.S. and Machado Rino, L.H., "DMSumm: Review and Assessment", In Proc. of Advances in Natural Language Processing, 3rd International Conference (PorTAL 2002), pp. 263-274, 2002.
- [43] Nik Adilah Hanin Binti Zahri, Fumiyo Fukumoto, Suguru Matsuyoshi, "Exploiting Discourse Relations between Sentences for Text Clustering", In Proc. of 24th International Conference on Computational Linguistics (COLING 2012), Advances in Discourse Analysis and its Computational Aspects (ADACA) Workshop, pp. 17-31, December 2012, Mumbai, India.
- [44] Vapnik, V. : The Nature of Statistical Learning Theory, Springer, 1995.
- [45] Lin, D., "PRINCIPAR- An Efficient, Broad-coverage, Principle-based Parser", In Proc. of 15th International Conference on Computational Linguistics (COLING), pp. 482-488, 1994.
- [46] Finkel, J.R., Grenager, T. and Manning, C., "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", In Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 363-370, 2005.
- [47] Buckland, L. & Dang, H., Document Understanding Conference Website, <http://duc.nist.gov/>
- [48] Kohavi, R. and Provost, F., "Glossary of Terms", Machine Learning 30, No. 2-3, pp. 271-274, 1998.
- [49] IBM SPSS Statistic Database, "Cluster Evaluation Algorithm" <http://publib.boulder.ibm.com>, 2011.
- [50] Kaufman, L. and Rousseeuw, P., "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons, London. ISBN: 10: 0471878766, 1990

AUTHORS

N. Adilah Hanin Zahri graduated from Computer Science and Media Engineering, University of Yamanashi in 2006. She received MSc in 2009 and PhD in Human Environmental Medical Engineering in 2013 from Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan. Currently, she is working at Department of Computer Engineering, School of Computer and Communication Engineering in University of Malaysia Perlis, Malaysia.



Fumiyo Fukumoto graduated from Department of Mathematics in the faculty of Sciences, Gakushuin University, 1986. From 1986 to 1988, she joined R&D Department of Oki Electric Industry Co., Ltd. From 1988 to 1992, she joined Institute for New Generation Computer Technology (ICOT). She was at Centre for Computational Linguistics of UMIST (University of Manchester Institute of Science and Technology), England as a student and a visiting researcher, from 1992 to 1994, and awarded MSc. Since 1994, she has been working at University of Yamanashi, Japan. She is a member of ANLP, ACL, ACM, IPSJ and IEICE.



Suguru Matsuyoshi received the B.S. degree from Kyoto University in 2003, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2005 and 2008, respectively. Prior to 2011, he was a Research Assistant Professor in Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan. Since 2011, he has been an Assistant Professor in Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan.



Ong Bi Lynn graduated with B. Eng. (Hons) Electrical and Electronics from Universiti Malaysia Sabah (UMS) in the year 2001. She received her Master of Business Administration from Universiti Utara Malaysia (UUM) in 2003. She obtained her Ph.D. in the field of Computer Network in the year 2008 from Universiti Utara Malaysia (UUM). Currently, she is working with Department of Computer Network Engineering, School of Computer and Communication Engineering in Universiti of Malaysia Perlis (UniMAP), Perlis, Malaysia.



AN EMPIRICAL EVALUATION OF CRYPTOOL IN TEACHING COMPUTER SECURITY

Mabroka Maeref¹ and Fatma Algali²

¹Department of Computer Science, Sebha University, Sebha, Libya
roka.mayouf@yahoo.com

²Department of Computer Science, Sebha University, Sebha, Libya
fatma.algali@yahoo.com

ABSTRACT

In the area of network security, the fundamental security principles and security practice skills are both required for students' understanding. Instructors have to emphasize both; the theoretical part and practices of security. However, this is a challenging task for instructors' teaching and students' learning. For this reason, researchers are eager to support the lecture lessons by using interactive visualization tools. The learning tool CrypTool 2 is one of these tools that mostly cover all of the above. In fact, the evaluations of the effectiveness of the tools in teaching and learning are limited. Therefore, this paper provides an overview of an empirical evaluation for assessing CrypTool 2 tool. The effectiveness of this tool was tested using an empirical evaluation method. The results show that this visualization tool was effective in meeting its learning objectives.

KEYWORDS

Computer Security, Cryptographic Protocols, Visualization and Animation, Empirical Evaluation

1. INTRODUCTION

The visualization and animation approach is increasingly being adopted in Computer Science education with the promise of enhancing student understanding of complex concepts. Using this approach, tools were developed using visualization and animation techniques to interactively help students gain knowledge and acquire skills about a subject. If these tools are exploited efficiently, they can facilitate the education process, thus minimizing the learning/teaching time for both lecturers and students.

In the area of network security, fundamental security principles and security practice skills are both required for a student to understand the subject matter. Instructors have to emphasize both the theoretical and practical aspects of security. However, this area poses a challenge for instructors to teach and for students to learn. For this reason, researchers have been eager to support lectures by offering interactive visualization and animation tools that facilitate student understanding and shorten the time consumed in long-term teaching [1-8].

In response to the rising number of security crimes and attacks, specific security courses have been developed by colleges and universities [9]. Although the Model Curricula for Computing CC-2008 [10] describes a cryptographic algorithm as an elective unit— with topics that include private and public key cryptography, key exchanges, digital signatures and security protocols—security experts, including Bishop [11], Hoglund [12] and Howard [13], emphasize the need to incorporate security into the undergraduate curriculum.

Cryptographic protocols mostly combine both theory and practice [14, 15] and as such, interactive visualization tools are essential [7, 8] to support a student's understanding of the subject matter. However, experiences with these kinds of tools are limited. A justification of these tools' effectiveness is highly required in order to declare the values of these tools. In this paper, we describe our experience with a CrypTool 2 [16] as an experimental procedure and evaluate the tool using empirical evaluation approach. The following section describes the most related works to our paper while section 3 explains the CrypTool description. Experimental procedure and results are described in section 4. A discussion of this paper is explained in section 5 and the conclusion is provided in section 6.

2. RELATED WORK

Researchers have developed various kinds of interactive visualization tools for teaching/learning cryptographic protocol behaviour and concepts. One of these tools is the Kerberos tool, which developed for visualizing one specific protocol: Kerberos protocol [17]. Another tool is the GRACE tool [3], the Game tool [18], GRASP tool [19] and crypTool [20, 21]. CrypTool is a freeware Program with graphical user interface for applying and analyzing cryptographic algorithms with extensive online help. Literature on related visualization tools, together with comparisons between them, is available in our papers [22] and [23].

The main goal of this paper is to evaluate quantitatively the effectiveness of CrypTool. For the purpose of this paper, effectiveness refers to the ability of this tool in enhancing student's understanding. This goal is evaluated using an empirical evaluation approach (without animation vs. animation with CrypTool tool). We have chosen this tool because it covers the most aspects of computer security. With respect to this chosen tool, the question is, "*Is teaching using CrypTool more effective than traditional teaching medium?*"

Various studies have been carried out for evaluating interactive mediums. From the literature, a study conducted by Kehoe et al. [24] used an interactive animation to teach algorithm animation and data structure. Their results showed in scores on a post-test used to evaluate the understanding with 12 students divided into two groups. The results showed that the animation group significantly outperformed the non animation group. Moreover, Yuan et al. [8] used Kerberos as an interactive animation tool to teach Kerberos protocol. His results showed in scores on pre-post tests used to evaluate the understanding with 16 students. The *t*-test results show that the improvement from pre-test to post-test is statistically significant. Hundhausen et al. [25] also considered 24 experiments used different concept of animation to teach algorithm animation and data structures. Twenty two of the experiments used post-test or pre-post tests to evaluate the understanding. Their results are various according to the interactivity of animation.

3. CRYPTOTOOL DESCRIPTION

Cryptool is a freeware Program with graphical user interface for applying and analysing cryptographic algorithms with extensive online help. It can be understandable without deep crypto knowledge. It contains nearly all state of the art crypto algorithms with "playful"

introduction to modern and classical cryptography. Learning through CrypTool is almost can be done by everyone either through the internet or by download and install the tool from the website (www.cryptool.org). The features of CrypTool include cryptography and cryptanalysis. Both of them constitute the science of cryptology. Figure 1 shows the main menu of the tool.

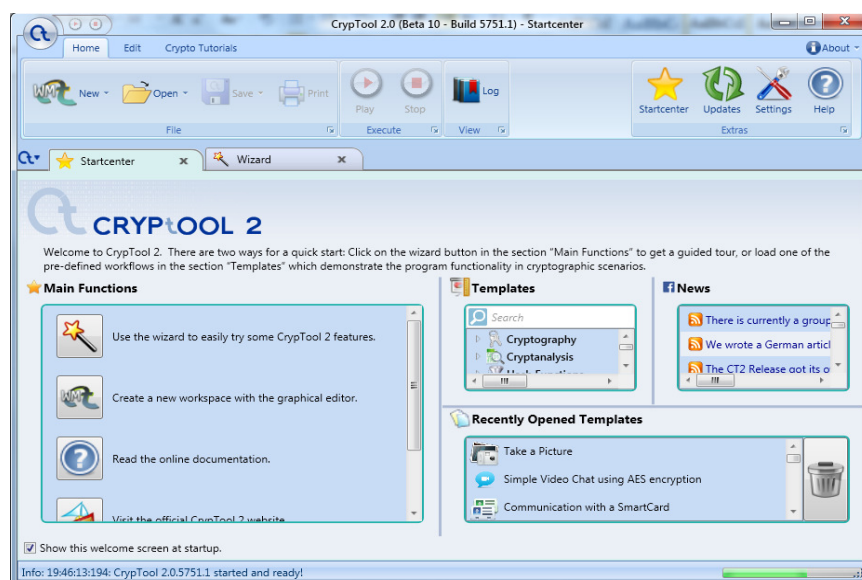


Figure 1. CrypTool main menu

4. CRYPTOL EVALUATION

We carried out an experiment consists of one group of the same lesson taught to the undergraduate Computer Science students of the Network Security course at Sebha University of Libya during the semester II of 2013-2014 year. The experiment was conducted in two stages where each stage uses a different learning medium approach; the first stage uses only text-based materials (no animation), the second stage uses CrypTool in the final part of the lesson. The student will be given a same test throughout the two stages. They are allowed to improve their answer after each stage. The results of the tests after each stage of the medium approach are compared.

The same topics of the lessons are given during all of the two stages. These topics are: symmetric-key and Asymmetric-key cryptographic protocol, Diffie-Hellman protocol with respect to the possible attack to Diffie-Hellman protocol, the concept of hash function, digital signature and digital certificate.

In this experiment, the tool SPSS [26] is used to statistically evaluate the effectiveness of CrypTool using t-test and p-value.

4.1. Experimental Procedure

A total of 20 students participated in the experiment. The students are final year of Computer Science students (undergraduate students) at Sebha University of Libya. We follow the pre-test to post-test accuracy [8, 25, 27] in order to evaluate the effectiveness of CrypTool. The same students were given the same lesson but using different medium each time. The experiment was

conducted using the learning medium approach (no animation vs. animation with CrypTool). The students were given the lesson using only text-based materials followed by a pre-test, then, the same students were introduced to CrypTool followed by a post-test.

The experiment was controlled by delivering the same lesson to all of the students by the same teacher during the two consecutive sessions. The topics were: symmetric-key and Asymmetric-key cryptographic protocol, Diffie-Hellman protocol with respect to the possible attack to Diffie-Hellman protocol and the concept of hash function, digital signature and digital certificate.

In the first session of the three hours, only text-based materials were used during the lesson time with the help of electronic slides. At the end of the session, the students were given a pre-test of ten multiple choice questions with a time limit of 30 minutes to answer them.

In the second session, after the pre-test, students were introduced to CrypTool and to its visual interface. They were asked to experiment with simple symmetric and asymmetric-key cryptographic protocols and to recreate Diffie-Hellman protocol. They were also asked to experiment with the concepts of hash function, digital signature, digital certificate and their usages of avoiding possible attack. At the end of the session, the students were given a post-test of the same questions as in the first session with a time limit of 30 minutes to answer them.

Again, to control the tasks performance, the same test of ten multiple questions were given to all students with a specific time. During the test, the students were not allowed to consult books or use any materials. Then the results of pre-test and post-tests were compared. The following points describe the details of the ten multiple questions:

- The first question dealt with the communication components of asymmetric-key cryptographic protocol.
- The second question dealt with the differences between symmetric-key and asymmetric-key cryptography.
- The third question dealt with Diffie-Hellman protocol steps.
- The fourth question dealt with the communication components of Diffie-Hellman protocols.
- The fifth question dealt with digital signature.
- The sixth question dealt with digital certificate.
- The seventh question dealt with Diffie-Hellman possible attack.
- The eighth question dealt with a hash function.
- The ninth question dealt with avoiding Diffie-Hellman protocol attack.
- The last question dealt with a hybrid system (using of both symmetric and asymmetric-key cryptography).

4.2. Experimental Results

To determine the effectiveness of CrypTool, a pre-test and post-test accuracy is used. Table 1 describes the students' scores for the pre-test and post-tests. Notice that the maximum score for each student is 10. In the other side, the Table 2 describes the mean of the group tested and Figure 2 explains the idea.

Table 1. The students' scores of pre-test and post-test

No.	Pre-test scores No animation	Post-test scores Using CrypTool
1	4	6
2	4	6
3	5	5
4	4	6
5	4	4
6	5	5
7	5	5
8	5	7
9	4	7
10	4	5
11	4	6
12	5	7
13	4	6
14	4	4
15	4	4
16	5	5
17	5	5
18	5	7
19	4	7
20	5	7

Table 2. The students' scores means of pre-test and post-test

Time	Treatment	No.	Mean
Sebha University	No animation	20	4.45
	CrypTool	20	5.70

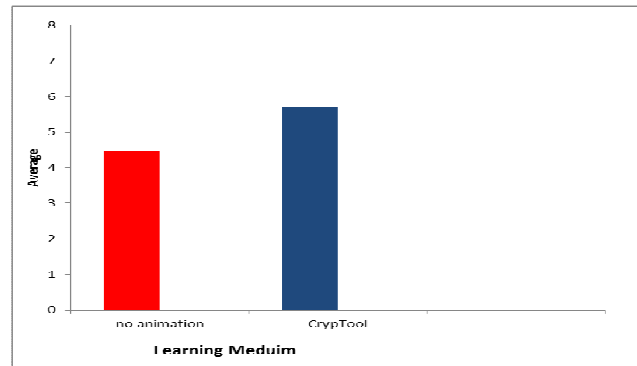


Figure 2. The means of the students' scores

The adopted statistical analysis of this experiment is that:

- Null Hypothesis (H_0): the conducted hypothesis is that there is no difference in the mean of pre-test and post-tests scores. In other words, the pre-test and post-tests scores will have equal means.
- Alternative hypothesis (H_1): the alternative hypothesis is that there is at least one difference in the mean of the pre-test and post-test scores in the group tested.

- p-value: the return value of the statistical test which indicates the probability of getting a mean difference between the groups as high as what is observed by chance. The lower the P-value, the more significant difference between the groups. The typical significance level that has been chosen in this experiment is 0.05.
- t-test: this test was run on the pre-test and post-test scores. In this experiment, the result t-test shows that there is a difference between the pre-test and post-test according to the p-value which is 0.0 and less than the significance level 0.05. table 3 shows the result of t-test.

Table 3. The results of t-test

Treatment	No. of student	Mean	p-value	t- test
CrypTool	20	5.7	0.0	CrypTool > No animation
No animation	20	4.45		

The test shows that there is a difference between no animation and CrypTool based on the p-value which equal to 0.0. The p-value is less than the significance level (0.05) and that means the improvement from pre-test to post-test is statistically significant.

5. DISCUSSION

The results in this experiment indicate that CrypTool is more effective and efficient than traditional learning medium. According to our hypothesis testing, there is a significant difference between using CrypTool as a teaching/learning medium and text-only material. It shows that CrypTool (interactive visualization and animation tool) significantly outperformed text-only material (no animation). The overall improvement of enhancing the students' ability for understanding the cryptographic protocols and computer security concepts using CrypTool is demonstrated and achieved.

6. CONCLUSION

Regardless of the advancement in the area of educational techniques, the area needs to be further tested with more empirical evaluation, especially of using the teaching/learning interactive visualization and animation tools. Currently, a few researches dealt with the problem of the lack of using these kinds of tools. The missing of a clear and complete principle design for interactive tools is seldom discussed and yet plays a crucial role in the tool development. The principle design is important because a tool without a base is inadequate even if it is supplied with good structures. Furthermore, studies have shown that visualization and animation educationally enhanced students' understanding if they were supported by active learning. This paper was motivated by these observations. In particular, this paper suggested more experiments of other interactive visualization tools through empirical evaluation in order to improve their effectiveness and teaching/learning support.

REFERENCES

- [1] Asseisah, M. S., Bahig, H. M., & Daoud, S. S., (2010) Interactive Visualization System for DES. Berlin Heidelberg: Springer-Verlag
- [2] Catrambone, R. & Seay, A. F., (2002) "Using Animation to Help Students Learn Computer Algorithms," The Journal of the Human Factors and Ergonomics Society, vol. 44, pp. 495-511.
- [3] Cattaneo, G., Santis, A. D., & Petrillo, U. F., (2008) "Visualization of cryptographic protocols with GRACE," Journal of Visual Languages and Computing, vol. 19 pp. 258-290.
- [4] Holliday, M. A., (2003) "Animation of computer networking concepts," ACM Journal on Educational Resources in Computing (JERIC), vol. 3, pp. 1-26.

- [5] Kazemi, N. & Azadegan, S., "IPsecLite: a tool for teaching security concepts," in SIGCSE '10 Proceedings of the 41st ACM technical symposium on Computer science education NY, USA, 2010.
- [6] Kerren, A. & Stasko, J. T., (2002) "Algorithm animation," Software Visualization, LNCS 2269, pp. 1-15.
- [7] Schweitzer, D. & Brown, W., (2009) "Using Visualization to Teach Security," JCSC, vol. 24, pp. 143-150.
- [8] Yuan, X., Vega, P., Qadah, Y., Archer, R., Yu, H., & Xu, J., (2010) "Visualization Tools for Teaching Computer Security," ACM Transactions on Computing Education, vol. 9, pp. 147-155.
- [9] Taylor, B. & Azadegan, S., "Moving Beyond Security Tracks: Integrating Security in CS0 and CS1," in SIGCSE '08: Proceedings of the 39th SIGCSE technical symposium on Computer science education, 2008, pp. 320-324.
- [10] CC2008, "Computer Science 2008, An Interim Revision of CS 2001."
- [11] Bishop, M. & Frincke, D., (2005) "Teaching Secure Programming," IEEE Security and Privacy, vol. 3, pp. 54-56.
- [12] Hoglund, G. & McGraw, G., (2004) Exploiting Software:How to Break Code. Boston: Addison-Wesley.
- [13] Howard, M. & LeBlanc, D., (2003) Writing Secure Code. Redmond, WA: Microsoft Press.
- [14] Stallings, W., (2006) Cryptography and Network Security: Principles and Practices, 4 ed. Upper Saddle River, NJ: Prentice Hall.
- [15] Forouzan, B. A., (2008) Cryptography and Network Security, 1 ed. New York, NY: McGraw-Hill Higher Education.
- [16] Deutsche, A., "CrypTool," 2009.
- [17] Yuan, X., Qadah, Y., Xu, J., Yu, H., Archer, R., & Chu, B., (2007) "An animated learning tool for Kerberos authentication architecture," Journal of Computing Sciences in Colleges, the twelfth annual CCSC Northeastern Conference, vol. 22, pp. 147 – 155.
- [18] Hamey, L. G. C., "Teaching Secure Communication Protocols Using a Game Representation," in Australasian Computing Education Conference (ACE2003), Adelaide, Australia, 2002.
- [19] Schweitzer, D., Baird, L., Collins, M., Brown, W., & Sherman, M., "GRASP: a visualization tool for teaching security protocols," in the Tenth Colloquium for Information Systems Security Education, Adelphi, MD, 2006, pp. 1-7.
- [20] Eckert, C., Clausius, T., Esslinger, B., Schneider, J., & Koy, H., "CrypTool," 2003.
- [21] Esslinger, B., "The CrypTool Script: Cryptography, Mathematics, and More," 10 ed: Frankfurt am Main, Germany, 2010.
- [22] Mayouf, M. A. & Shukur, Z., (2008) "Animation of Natural Language Specifications of Authentication Protocol," Journal of Computer Science, vol. 4, pp. 503-508
- [23] Mayouf, M. A. & Shukur, Z., (2009) "Using Animation in Active Learning Tool to Detect Possible Attacks in Cryptographic Protocols," LNCS 5857, pp. 510-520.
- [24] Kehoe, C., Stasko, J., & Taylor, A., (2001) "Rethinking the evaluation of algorithm animations as learning aids: an observational study," International Journal of Human Computer Studies, vol. 54, pp. 265-284.
- [25] Hundhausen, C. D., Douglas, S. A., & Stasko, A. T., (2002) "A meta-study of algorithm visualization effectiveness," Journal of Visual Languages and Computing, vol. 13, pp. 259-290.
- [26] Pallant, J., (2010) SPSS Survival Manual: A step by step guide to data analysis using SPSS. Berkshire UK: McGraw-Hill Education.
- [27] Hansen, S. R., Narayanan, N. H., & Douglas, S., (2000) "Helping Learners Visualize and Comprehend Algorithms Interactive Multimedia Electronic " Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, vol. 2,

AUTHORS

Mabroka Maeref: received her BSc degree in Computer Science from University of Sebha, Libya, MSc in Computer Science from Universiti Sains Malaysia, and PhD in Software Engineering from Universiti Kebangsaan Malaysia. Her interests span a wide range of topics in the area of Software Engineering, Networking, Computer Security, Visual Informatic and Computer Education. she is currently working as a lecturer at the departement of computer science, Faculty of Sciences in Sebha University of Libya.



Fatma Abdullah Alghali received a Ph.D. in Computer Science (Software Engineering) from University of AL-Neelain SUDAN 2006, Master of Computer Science from Warsaw University of Technology, Poland , 1997, BSc of Computer Science from Sebha University, Libya, 1991, Her research interest includes Software Engineering, Human Computer Interactive (HCI) , E-Learning, Cloud Computing, She is working as Assistant Professor. In Computer Science Department of Sebha University LIBYA



ENTERPRISE DATA PROTECTION: MEETING REQUIREMENTS WITH EFFICIENT AND COST-EFFECTIVE METHODS

Khaled Aldossari

EXPEC Computer Center, Saudi Aramco, Saudi Arabia

dosskm01@aramco.com

ABSTRACT

This paper addresses the major challenges that large organizations face in protecting their valuable data. Some of these challenges include recovery objectives, data explosion, cost and the nature of data. The paper explores multiple methods of data protection at different storage levels. RAID disk arrays, snapshot technology, storage mirroring, and backup and archive strategies all are methods used by many large organizations to protect their data. The paper surveys several different enterprise-level backup and archive solutions in the market today and evaluates each solution based on certain criteria. The evaluation criteria cover all business needs and help to tackle the key issues related to data protection. Finally, this paper provides insight on data protection mechanisms and proposes guidelines that help organizations to choose the best backup and archive solutions.

KEYWORDS

Data Protection, Data Loss, Data Recovery, Backup, Archive

1. INTRODUCTION

In any organization, the requirement to store digital data has been growing exponentially year after year. To cope with this increasing data requirement, larger amounts of bigger and faster storage devices need to be installed in data centers around the world. The downside with having more hardware installed is that it also increases the chance of losing data due to user and hardware error or malfunction. Losing data can be costly for organizations both legally and financially. Below are some statistics that show the potential results from losing data:

- The cost associated with lost data for the energy business is \$2.8 million of lost revenue per hour. [1]
- The cost of recreating just 20 MB of engineering data is 42 days and \$98,000. [2]
- In less than a year after they faced a major data loss, 70 percent of small companies stop business permanently. [2]
- Among companies that lost data in 2012, only 33 percent were able to recover 100 percent of their data. [3]

To avoid such impacts, a successful data protection strategy has to keep data available and accessible against possible losses caused for any reason. In fact, data loss can happen for different reasons, including:

- Hardware or system malfunctions, such as power failure, media crash and controller failure
- Human errors, such as accidental deletion of files or physical damage caused by dropping storage devices
- Software corruption, including software bugs and software crashes while editing
- Computer viruses and malware
- Natural disasters, such as earthquakes, floods and fires

The chart below shows the percentage of data loss incidents due to each leading cause according to Kroll Ontrack Inc. [1]

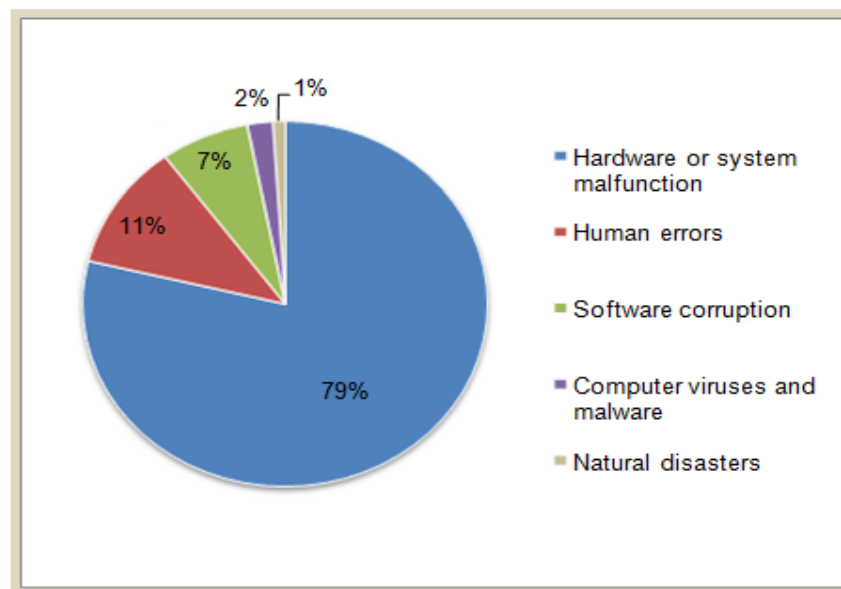


Figure 1. Causes of data loss

The next sections cover, in detail, the availability and recoverability aspects of data protection solutions at the enterprise level.

2. DATA RECOVERY CHALLENGES

Today, large organizations face challenges when they plan and implement data recovery solutions. These challenges could make the recovery of data in the event of data loss more difficult. As a result, it is important to understand and address these challenges before implementing a data protection strategy. Major challenges are listed and explained below:

2.1. Recovery Objectives

The recovery time objective (RTO) and recovery point objective (RPO) are two critical business concepts related to data recovery. RTO is the maximum time period by which the data must be restored after a data loss event. On the other hand, RPO is defined as a point in time prior to a

data loss event where data can be restored. Each organization has its own RTO and RPO that should be defined clearly and carefully based on the business needs and regulatory compliance requirements. Meeting recovery objectives is one of the challenges that faces data protection solutions.

2.2. Data Explosion

As the volumes of data continue to grow exponentially, the scalability and performance of data protection solutions become significant challenges. The backup and restore system has to have enough performance to meet the recovery objectives. It also has to be able to scale out to encounter future data expansion.

2.3. Cost

The cost associated with data protection is one of the greatest challenges. The total cost of ownership of the backup and recovery infrastructure and its operational cost are proportional to the amount of data to be protected. As a result, the data protection solution should be cost-effective while maintaining the recovery objectives.

2.4. Nature of Data

Enterprise organizations usually have different types of data that reside on heterogeneous systems. Structured data, such as databases, requires backup and recovery technologies that are different from those used with unstructured data, such as images and videos. The change rate of data affects the frequency of its backup. With equal total size, a large number of small files has more of an impact on the performance of backup and recovery systems than a small number of large files. Therefore, different types of data require different backup and recovery technologies, which make data protection a challenge.

3. LEVELS OF DATA PROTECTION

To address the aforementioned challenges, enterprises implement data protection at different levels. Data protection technologies use the principle of redundancy to prevent a total loss of data by creating another copy of it [4].

The most common technologies of data protection are explained below.

3.1. RAID Disk Arrays

Redundant array of inexpensive/independent, disks (RAID) is a storage technology that combines multiple disks, known as a RAID set, and presents the disks as a single logical disk. [5] With this technique, data spread across the RAID set enhances reliability and increases I/O performance. [5] Different architectures of RAID exist to provide different levels of fault tolerance and performance. The most common levels of RAID include RAID 0, RAID 1, RAID 5, and RAID 10. [6]

RAID is perfect for protecting against hardware failures but cannot protect against other types of risks such as human errors, software corruption, malware, virus attacks and natural disasters.

3.2. Snapshot Technology

Snapshot technology is an instance copy of a defined collection of data—a set of files, directories, or volumes—at a particular point in time. [7] Apparently, a snapshot provides another level of redundancy. Different storage vendors offer various implementations of snapshot. The most popular snapshot implementations are copy-on-write, redirect-on-write, split-mirror, log structure file architecture, copy-on-write with background copy and continuous data protection. [8]

Snapshots provide protection against human errors, software corruption, malware and virus attacks but still cannot protect against natural disasters.

3.3. Storage Mirroring (Replication)

As mentioned above, RAID and snapshot cannot protect data against natural disasters. Here is where storage mirroring, also known as replication, comes in handy. Today, many storage manufacturers offer replication solutions, which can be used for disaster recovery. Typically, replication solutions copy data from the primary storage system over long distances to another storage system. [4] With this technique, an up-to-date copy of the data is maintained at a remote site in case the primary copy is lost. Replication solutions operate in either synchronous or asynchronous mode. With synchronous replication, data is written to the primary and secondary storage systems at the same time. [9] Asynchronous replication, on the other hand, writes data to the secondary storage system with a delay. [9]

3.4 Backup and Archive Strategies

Although storage replication enables business continuity in case one site is lost, it introduces extra overheads on writing. Also, it requires more disk space, which needs more power and space. Most enterprises implement backing up and archiving to cheaper media, mainly tapes, to keep data protected and, at the same time, lower associated cost.

The backup process creates a redundant copy of the original data on a different location, or storage media, for the purpose of recovery in case of data loss. The most common types of backup are full and incremental backup. With full backup, the entire file system is copied to the backup destination. It allows fast recovery of the file system in case it becomes inaccessible. Backing up the entire file system is slow and requires more backup media. Incremental backup, on the other hand, provides faster backup with less capacity by copying only files that are created or modified since the last backup. [10]

Archiving is the process of moving a selected collection of data, usually inactive, to another storage system for long-term retention. Archives are kept for long periods of time to meet regulatory compliance and/or for future reference.

Backup and archive play a major role in almost every data protection plan. Today, most backup systems also provide archive capabilities. Large organizations typically use low-price disks or tapes as the destination storage system for their backups and archives to reduce the cost. It has always been a good practice to send the backup/archive media to a remote place as a part of the disaster recovery plan.

The table below summarizes the possible data recovery methods to protect against each type of data loss threat:

Table 1. Recommended Data Recovery Methods

Data Loss Cause	Data Recovery Method
Hardware or system malfunctions	RAID, storage mirroring (replication), backup and archive
Human errors	Snapshots, backup and archive
Software corruption	Snapshots, backup and archive
Computer viruses and malware	Snapshots, backup and archive
Natural disasters	Storage mirroring (replication), backup and archive

4. EVALUATING BACKUP AND ARCHIVE SOLUTIONS

Backup and archive strategies are the most common data recovery methods used by most enterprises. Therefore, in this section, we propose different criteria to help enterprises choose backup and archive solutions that meet their requirements. Then, we see how the EXPEC Computer Center at Saudi Aramco Oil Company used these criteria to evaluate some of the popular backup and archive systems available in the market. These systems are IBM Tivoli Storage Manager (TSM), CommVault® Simpana, and Interica Intelligent Data Store (IDS).

We have identified 18 criteria that cover every aspect of data backup and recovery. The importance of each criterion varies from one organization to another depending on its requirements. These criteria are:

4.1. Architecture

Different backup and archive solutions have different architecture. Certainly, the system architecture has an impact on its performance, scalability, reliability and other features.

In one-tier architecture, all components of the system exist on a single server. The advantage of using a one-tier architecture backup and archive system is its simplicity and ease of management. The scalability and overall performance of this type of architecture does not help large organizations meet their backup and archive requirements. Two-tier architecture consists mainly of clients and a server residing on different hosts. In this type of backup and archive system, the client moves backup or archive data to the server and the server only keeps track of metadata. Client-server systems have the advantage of flexibility and can provide better performance, yet they have the single server bottleneck. Three-tier architecture, on the other hand, involves clients, a server and data movers. Clients send their backup or archive data to the data movers, which move them to the backup storage. The role of the server is just to monitor the whole backup and archive environment and to execute some administrative tasks. This architecture delivers better performance and more scalability but it might increase the complexity of the solution. Sharing resources like tape drives or the host memory in three-tier architecture requires more effort.

Interica IDS, by itself, is a single-tier backup and archive system, which provides centralized tape storage management. Because of its architecture, IDS cannot scale very well to be able to protect a complicated environment with large storage capacity. IBM TSM can be configured to operate as a two-tier system or a three-tier system. But, even with three-tier configuration, the TSM server has to do more of the tasks such as generating a second copy of the backup or archive, migrating data from one storage pool to another, and maintaining the system database. CommVault® Simpana provides a clear three-tier solution where its data movers transfer data from clients to

backup storage. Its server is just responsible for monitoring the entire environment and collecting statistics and reporting about it.

4.2. Scalability

The backup and archive solution has to be scalable enough to meet the continuing growth of data. As mentioned before, the architecture of the solution has a direct impact on its scalability. In addition, the internal structure of the software and its associated database affects solution scalability. Our evaluation shows that both TSM and Simpana have better scalability over IDS.

4.3. Reliability

Is the solution highly available with no single point of failure? By looking into each component of the solution, you can pinpoint the possible cause of failure and, therefore, service disruption. Both TSM and Simpana support cluster configurations that provide automatic recoverability and increase availability. IDS does not have the concept of clustering, but it can manually fail over the server to another host and still point to the same database.

4.4. Performance

The overall performance of the solution should be sufficient to cover the RTO and RPO requirements. Two measurements can be used to evaluate the performance of the system: backup or archive speed (TB/hr) and restore speed (TB/hr). TSM, Simpana and IDS have no limits and can push to the maximum what the storage systems can deliver.

4.5. Supported operating systems

Each organization has its own preferable operating system. Therefore, it is important to know operating system platforms for which backup and archive solutions support every component. TSM is compatible with most operating systems including Linux, AIX and Windows. The data mover and client components of Simpana can run on all operating systems but the server component runs only on Windows. IDS runs only on Unix or Linux machines.

4.6. Simplicity

The backup and archive solution should be easy enough to implement and to manage. Losing data by itself is troublesome — the data recovery should not be. During the evaluation, we found that all solutions are not hard to deal with. Nevertheless, since TSM is a large system, it might be, for some administrators, more complicated than others.

4.7. Security

Data security is a critical feature in any backup and archive system. The system shall provide different levels of user access control such as administrator, operator and users. In some cases, integrating the operating system permissions and ownership with the backup system is necessary. Unlike TSM and Simpana, IDS, by itself, does not support any kind of security. It requires another product named PARS (Project Archive and Retrieval System) to solve the security issue.

4.8. Intelligence

Is the backup and archive system smart enough to fix damaged data in its backup storage? Is it capable of performing some data analysis and share the result with the administrator? All evaluated systems can fix damage within its storage media to a certain limit. Simpana goes further and enables data analysis, such as classification of data based on its access date, type, size and others.

4.9. Data Policy Management

Depending on the business requirements, particular data policies are needed. Examples of these policies include flexible data retention policy per dataset, ability to extend data retention on the fly, automatic data expiration process, automatic data or tape media replication feature for disaster recovery, and automatic media transcription. Our evaluation shows that Simpana software has a more flexible data policy management than the other systems.

4.10. Open Standards

Does the backup and archive solution support open standards? More specifically, does the system write data to tape media in open format readable by other applications? Can the system export and import data to/from an open format such as LTFS? Is the system capable of rebuilding the system from tape without any additional outside information? TSM and Simpana use proprietary formats unlike IDS, which uses an open format (tar). Only TSM can support LTFS. TSM also requires additional outside information to rebuild the system; Simpana and IDS do not depend on additional outside information for rebuilding the system.

4.11. Metadata Search Engine Capability

In many cases, especially for archives, a fast and reliable search engine capability is very helpful. Users can navigate through the command line interface (CLI) or Web interface to identify and retrieve any archived data by searching one or more key fields in the database such as dataset name, size, tags and age. The evaluation shows clearly that Simpana is more capable in providing an efficient and flexible metadata search engine.

4.12. Tape Vault Management

Most large organizations use tape as their backup storage media. Usually, the organization maintains a large number of tape media that exceed the automated tape library capacity. As a result, the backup and archive solution has to be able to vault tapes for disaster recovery, and to track and report data on these tapes. Also, it has to track tape media outside the library and notify the operator console when there is a need to insert a tape into the tape library. Only Simpana has this capability as a built-in feature. TSM and IDS do not support this feature. But there are few products that can integrate with TSM to do the tape tracking management task.

4.13. Tape Operator Console

Besides the previous requirement, the solution shall provide a centralized tape operator console. The console can be a Web or GUI-based interface that provides real-time monitoring of solutions. The console should present helpful information to the operator such as the health status of the tape library, tape drives, online tape media, capacity (online/vaulted), tape drive activities (busy/idle), errors and alerts, and a list of tapes on the shelf. It also has to show actions waiting

for operator input such as inserting tapes. All evaluated solutions support tape operators consoles with different capabilities.

4.14. Reporting

One of the important features of any enterprise solution is its reporting capability. In the backup and archive environment, the system shall be able to report the status of backup and archive jobs. It also has to generate reports about the health status of its storage systems, servers and clients. Customized reporting on data utilization per user group or class will also be useful. The evaluation shows that TSM and Simpana provide more advanced reports than IDS.

4.15. Support Services

Regardless of the level of expertise an organization has, the vendor should support the solution. In addition, having official documentation is necessary for any proposed solution. Large community support can also help system administrators to resolve related problems and come up with new ideas. Official training is required to build the administrator's skills and expertise for the solution. Vendors of all evaluated systems provide professional services on their products. TSM has the oldest and largest community support. The community of Simpana is becoming larger and experienced. IDS software has weak community support. Regarding training, both TSM and Simpana have an excellent training path.

4.16. System Popularity

Some enterprises look for a solution that is more popular and used by a wide range of businesses. It provides an enterprise with more confidence and allows it to find needed resources easily and cost-effectively. Both TSM and Simpana are very popular backup and archive enterprise solutions. Fewer customers, mainly with oil and gas exploration and production business lines, use IDS as a project-based archival solution.

4.17. Other Features

Depending on business needs, different organizations might require other features. In some cases, supporting multiple automated tape libraries within the system domain is needed. Limitations on the number of files to manage or the maximum size of a single file might impact the selection of the system. Spanning a single large file over two or more tape media is important to consider if the environment has very large files. Unlike Simpana and IDS, TSM does not provide a virtualization of automated tape libraries to appear as a single library. All evaluated solutions do not have issues with spanning a single file over tape media. During the evaluation, all systems were able to backup 10TB files without failures.

4.18. Total Cost of Ownership

The capital and operational expanses of the backup and archive system are important factors when evaluating different solutions. It is also important to consider the license model of the solution (per TB, number of servers, number of hosts to backup or others), which indeed affects its cost in the long term. The key point here is to choose the most cost-effective and affordable solution that meets the minimum backup and archive requirements.

5. SUMMARY

The value of digital data in any organization has increased. Without successful data protection strategies, data loss can be costly to an organization. We discussed challenges that face data recovery. These challenges include recovery objectives, data explosion, associated costs, and the nature of data. Then, we reviewed the different technologies used by most enterprises to overcome these challenges at different levels. Such technologies are RAID disk arrays, snapshot technology, storage replication, and backup and archive strategies. Because backup and archive systems are used by most enterprises, we focused on this method. We proposed 18 criteria that cover every aspect of backup and archive systems. These criteria can be used by any organization as a template when weighing different backup and archive solutions on the market. Finally, we showed how the EXPEC Computer Center at Saudi Aramco used these criteria to evaluate three backup and archive systems: IBM Tivoli Storage Manager (TSM), CommVault® Simpana, and Interica Intelligent Data Store (IDS).

ACKNOWLEDGEMENTS

The author would like to thank his colleagues in Saudi Aramco, Hussain Al-Raqa for his encouragement to write this paper and Edward Liu for his valuable comments.

REFERENCES

- [1] Kroll Ontrack, 'Understanding Data Loss'. [Online]. Available: <http://www.ontrackdatarecovery.com.au/understanding-data-loss/>. [Accessed: 23- Sep- 2014].
- [2] M. Foster, 'Save your business with data backup', NetSource Technologies. [Online]. Available: <http://www.netsourceinc.com/blog/save-your-business-with-data-backup>. [Accessed: 23- Sep- 2014].
- [3] Kroll Ontrack, 'Kroll Ontrack study reveals 40 percent of companies lose data annually from their virtual environments', 2013. [Online]. Available: <http://www.krollontrack.com/company/news-releases/?getPressRelease=62077>. [Accessed: 23- Sep- 2014].
- [4] C. Chang, 'A Survey of Data Protection Technologies', 2005 IEEE International Conference on Electro Information Technology, p. 6, 2005.
- [5] M. Dutch, A Data Protection Taxonomy. The Storage Networking Industry Association, 2010, p. 20.
- [6] R. Natarajan, 'RAID 0, RAID 1, RAID 5, RAID 10 Explained with Diagrams', The Geek Stuff, 2010. [Online]. Available: <http://www.thegeekstuff.com/2010/08/raid-levels-tutorial/>. [Accessed: 23- Sep- 2014].
- [7] M. Staimer, 'Backup in a snap: A guide to snapshot technologies', Storage Technology Magazine, 2009. [Online]. Available: <http://searchstorage.techtarget.com/magazineContent/Backup-in-a-snap-A-guide-to-snapshot-technologies>. [Accessed: 23- Sep- 2014].
- [8] StoneFly, 'Exploring Storage Snapshot technology'. [Online]. Available: <http://www.iscsi.com/resources/Storage-Snapshot-Technology.asp>. [Accessed: 23- Sep- 2014].
- [9] D. Bradbury, 'Remote replication: Comparing data replication methods', ComputerWeekly, 2011. [Online]. Available: <http://www.computerweekly.com/feature/Remote-replication-Comparing-data-replication-methods>. [Accessed: 23- Sep- 2014].
- [10] A. Chervenak, V. Vellanki and Z. Kurmas, 'Protecting file systems: A survey of backup techniques', in Joint NASA and IEEE Mass Storage Conference, 1998.
- [11] P. Dorion, 'Backup vs. archive', Search Data Backup, 2008. [Online]. Available: <http://searchdatabackup.techtarget.com/tip/Backup-vs-archive>. [Accessed: 23- Sep- 2014].
- [12] H. Garcia-Molina, C. Polyzois and R. Hagmann, in Compcon Spring '90. Intellectual Leverage. Digest of Papers. Thirty-Fifth IEEE Computer Society International Conference, 1990, pp. 573-577.
- [13] L. Black, 'The Importance of Data Backup', The Livingston Business Journal, 2014. [Online]. Available: <http://www.livingstonbusiness.com/2014/07/20/the-importance-of-data-backup/>. [Accessed: 23- Sep- 2014].
- [14] Software Testing Class, 'What is Difference Between Two-Tier and Three-Tier Architecture?', 2013. [Online]. Available: <http://www.softwaretestingclass.com/what-is-difference-between-two-tier-and-three-tier-architecture/>. [Accessed: 23- Sep- 2014].

AUTHOR

Khaled M. Aldossari works with the data storage support group at the EXPEC Computer Center, Saudi Aramco. For more than eight years of experience, Khaled led major projects to evaluate, design, and implement different data protection solutions. He worked also on supporting large-scale high performance storage. Khaled attained a distinguished Bachelor Degree in Computer Engineering from KFUPM University. He also received his Master's degree in Computer Science from California State University. In addition, Khaled is a SNIA Certified Storage Professional.



E-EDUCATION WITH FACEBOOK – A SOCIAL NETWORK SERVICE

Mohammad Derawi

Smar Wireless Systems, Gjøvik University College, Norway
mohammad.derawi@hig.no

ABSTRACT

In this paper, we study the social networking website, Facebook, for conducting courses as a replacement of high-cost classical electronic learning platforms. At the early stage of the Internet community, users of the Internet used email as the main communication mean. Although email is still the essential approach of communication in a suitable but offline mode, other services were introduced, such as many Instant Messaging (IM) software applications like ICQ, Skype, Viber, WhatsApp and MSN, which enable people to connect in a real-time mode. However, the communication between people was further improved to the next phase, when Facebook came to reality as a social networking homepage that wires many features. People do not only link with others, but also establish all kinds of connections between them. Facebook offers rich functions for forming associations. The framework of Facebook actually delivers without charge software that were provided by traditional electronic learning. This paper looks at how people apply Facebook for teaching and learning, together with recommendations provided.

KEYWORDS

Facebook, Education, Social Network Services

1. INTRODUCTION

Internet provides software applications of the necessary communication media for computation purposes. Due to its popularity, it has become a necessity of modern people for communication and information sharing purposes. For example, casual Internet users are using email as a replacement of sending letters via postal [1]. Although emails arrive at the mailboxes of recipients instantly, emails are to be read only when the recipients check their accounts. At the early stage, computers allow instant messaging among Internet users using software talk on UNIX operation system. It enables Internet users to communicate in real-time by sending textual data character by character. However, these applications were not popular among casual Internet users, because they must access to host machines [2].

The extensive use of Facebook is not only due to its popularity, but also due to the support by various devices. Facebook is a web application that can be accessed via any web browser. Besides, many mobile phones are equipped with web browsers, such as Opera Mini (a mobile phone version of Opera web browser), and some are even equipped with dedicated software solely for accessing Facebook, such as Apple iPhone, Samsung, Ultra-mobile PC's, various

netbooks and the Apple iPad. The support of Facebook by these mobile devices is a definite advantage of using Facebook for education purposes[3].

2. FACEBOOK FUNCTION

Facebook is a social networking web application that supports the following functions, which are for education purposes[4][5]:

- *No Cost* - The use of Facebook is free of charge.
- *No prerequisite* - Any Internet user with a valid email address is allowed to register
- *Group* - It supports user-defined groups so that users can be divided into groups. There are private groups and public groups. The former can only be joined by users via invitation and the latter is open to all. On the other hand, Facebook page enables any student to join the page for accessing the teaching materials and to be notified by any update of the page.
- *Page* - It enables users to create Facebook pages for particular organizations, so that other users can join the group and will be informed of all updates to the Facebook pages.
- *Privacy* - It supports the control of privacy in terms of items posted, users and groups. In other words, it is possible to set the access control privileges of individual items posted, users and groups.
- *Notifications* - It supports user notifications of all updates of items, users and groups via emails. If there is any update of an item, a user or a group, emails are sent to the related users for notifications.
- *Photo albums* - It supports user and group level photo albums.
- *Discussions* - It supports discussions with respect to a message, a photo, a photo album or an article.
- *Emails* - It supports internal emails between any two Facebook users, and it is possible to send an email to all users of a group.
- *Events* - It supports events and is possible to create events for a group. Users to indicate whether they will be present or absent from the events.
- *User main page* - The main page of a Facebook user shows all the updates to friends, the groups joined, and all the upcoming events.
- *Chatting* - Facebook support real-time chatting through the web browser.
- *User-defined software* - There is a well-defined Facebook API (Application Program Interface) so that software developers can develop software to be executed within the Facebook webpage. For example, quiz creator software enables any Facebook user to create a survey, questionnaire or quiz easily. Furthermore, applications for file sharing allows users to share their own documents with any other users. Besides, some Facebook applications are educational[10] .
- *Activity log* - All operations by any Facebook user are logged with timestamps and can be traced.

3. EDUCATIONAL PLATFORM

It is conceivable to use Facebook as a social network for education purpose as follows:

- *User creations* - Teaching staff and students need to access the Facebook website for registration. Preferably, they all use their email accounts granted by the universities, so that it is easier for them to locate one another. Furthermore, each of them can keep their

own personal Facebook accounts for their own casual uses, whereas the Facebook student users created by using university accounts are for teaching and learning only if they would like to prevent lectures from accessing their private life in the social networking website. The limitation is that the students may not log on their Facebook that is associated with their university email account daily.

- *Course preparations* - Teaching staff can create a Facebook page for each course with their Facebook accounts. Each Facebook page can create multiple photo albums and multiple discussions. Therefore, teaching staff can make use of the facilities provided by Facebook to enrich their Facebook page for the course, such as adding links to references materials, discussions or photo albums.
- *Teaching materials preparation* - For teaching purposes, the most important teaching materials to be distributed are lecture notes or slides. Usually, teaching staff uses Microsoft PowerPoint to prepare the PowerPoint files for students to download. Although there are free Microsoft PowerPoint viewer applications for Windows platforms released by Microsoft, there are platforms and mobile devices that cannot display PowerPoint files properly. Instead, image file format is the universal format for display purposes. It is therefore preferable to convert all PowerPoint files into sequences of images, and upload them as Facebook page photo albums. There are freeware applications that can convert Microsoft Office files into sequences of images. Then, Facebook users will be notified the existing of new slides, which can be accessed by any web-browsing enabled devices. For presentation files other than Microsoft PowerPoint, lectures can also using different applications to convert the files into images for uploading. Most mobile devices can be used for web browsing. Some mobile phones, such as Apple iPhone, are equipped with dedicated components for accessing Facebook. With the existence of mobile network technologies, such as GPRS and HSDPA, students can view the lecture notes as photo albums on the Facebook page for the course anywhere. The teaching materials in Microsoft Office formats can be uploaded to a web server and their URLs can be posted to the Facebook pages. As a result, Facebook student users can determine whether to download the original files. Besides, it is possible to post links of videos or upload video files to the course Facebook page, such as the videos for the lectures or demonstrations.
- *Conducting lectures and tutorials* - Teaching staff can use the PowerPoint or other presentation files to conduct lectures and tutorials. With Facebook, they have an alternate way to present the notes, which is showing Facebook photo albums for the presentation files. There is an extra benefit of showing a photo album compared with presenting a presentation file, which supports discussions on the entire photo albums and individual slides. Furthermore, while showing a slide as Facebook image, students can add comments to the slide which will notify the teaching staff the existence of comments for immediate feedbacks. It facilitates the discussions among teaching staff and students, especially those who are unwilling to speak in front of other students. Furthermore, if a student has any problem on any slide, he or she can add a comment, and a notification email will be sent to the teaching staff. Then the teaching staff can simply click the link embedded in the email to locate the slide (image) the student mentioned and provide feedbacks.
- *Discussions* - Whenever there is any update to the course group or course page, all involved Facebook student users are notified and can access those changes, such as a posting of links referring to online reference materials, videos and a creation of photo albums. Then, all users can access to those items and leave comments which can be read by other users for discussions. By consolidating the reference materials which originally scattered in the Internet, students time for searching the materials by themselves can be saved. For example, lecture notes can be released as Facebook photo albums, so that all students can access these albums for viewing them. Whenever they have any comments or questions regarding any slides, they can leave comments or questions to them.

Teaching staff and other students will be notified of such comments or questions by emails, and leave responses on the slide. Since Facebook is informal, users are more willing to leave messages on them. It actually motivates students to share and discuss for peer-to-peer learning. In fact, there are many interactive applications developed for Facebook. Lectures can make appropriate use of those external applications to facilitate interactions among lecturers and students.[6]

- *Assessments* - Facebook provides application programming interface (API) for software developers to develop Facebook applications. As such, there have been a lot of applications available for Facebook users. There are several Facebook applications which enable Facebook users to create quizzes, such as the Quizzes and Quiz Creator applications. By using these software, teaching staff can create a quiz, such as for each lecture, and post the link to the course page, and inform students to take the quiz to examine their understandings on the course materials. In addition, file sharing applications allow students submit assignments to teachers easily. As Facebook can be accessed by any web browser, students can increase their understanding on the course materials, anytime and anywhere. [7]
- *Personal notes and private files* - Students can make use of notes function in Facebook to keep their personal study notes. They can either keep the notes private or share the notes with others. Private files can be sent using the private message function with attachment. [8]
- *Privacy, security and legal issues* - Facebook provides customization in course account setting that protect privacy and ensure security of course access. The course creator can set the access of content to their students only by using the “add friend” function and “controlling how you share” function properly. Account and privacy setting can be performed under the “Account” session in Facebook. Regarding the legal issues of posting teaching materials in the social networking website, the lecturers should well aware of the terms and agreements listed in Facebook. By using Facebook appropriately, education functions can be delivered via this platform effectively.[9]

4. CASE STUDY

In this case study, a course is used for illustration purposes. Upon the creation of Facebook account by a teaching staff, the lecturer can create a Facebook page for the course with the given course code. For creating the course account, the teaching staff, clicked the Drop down arrow and "Create Page" to create a new page for the course as shown in Figure 1

By clicking “Create Page to start creating the page for the course”, the teaching staff can specified the course details on the webpage as in Figure 2 below. Finally, the teaching staff clicked “Create Page” to create the course page. Then, the lecturer could create photos albums for the lecture notes. The lecturer clicked “Photos” to create a new photo album

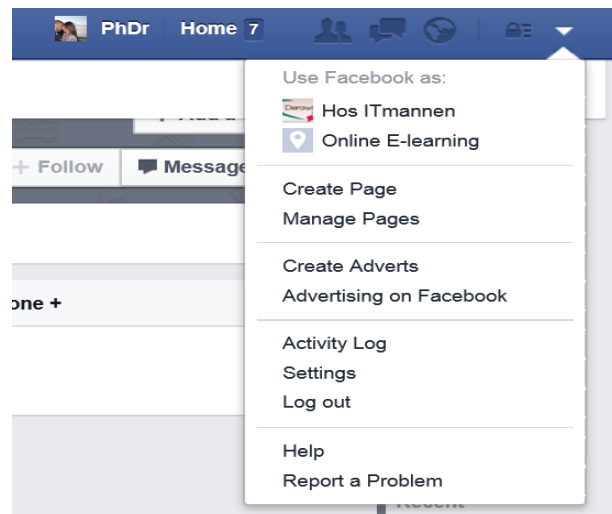


Figure 1. Facebook page for personal profile creation

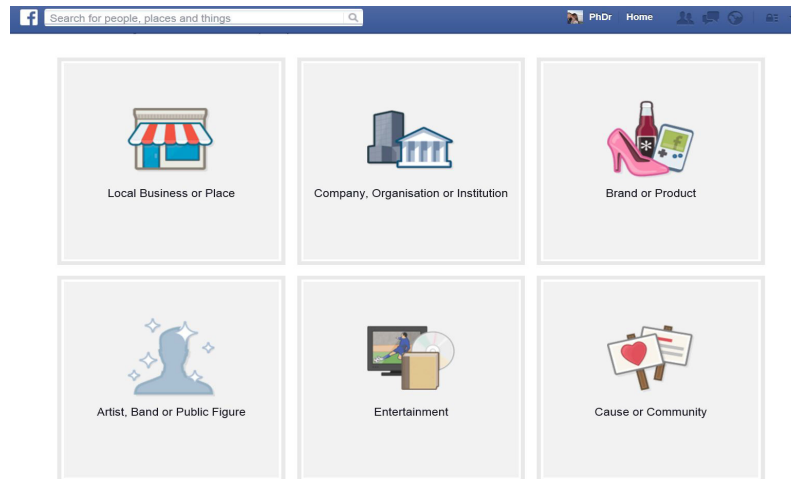


Figure 2. Facebook page for course main menu

The lecturer then started uploading the lecture notes images to the photo album. Since different web browsers support different approaches of uploading images to a photo album, for example, Microsoft Internet Explorer and Google Chrome can make use of a Facebook plugin whereas Firefox uses a Java based component, the lecturer would experience different interface when using different web browsers. Once the photo album was created, the lecturer reviewed the images and rearranged the sequence of the images as necessary. Then, the photo album with lecture notes was ready to be accessed by students.

For those students who would like to receive notification of course notes publishing, the lecturer could instruct them to use the function of adding themselves as fans of the course page. The lecturer could either rearrange the images or add new images by clicking “Organize Photos” or “Add Photo” buttons. For any further updates of the album, students with the role of fans of the course page would receive new notifications about the changes. The overview of the album is shown in Figure 4 and a screen showing the course content is shown in Figure 5.

Set up Online E-learning

1 About
2 Profile Picture
3 Add to Favourites
4 Reach More People

Add categories, a description and a website to improve the ranking of your Page in search.
Fields marked by asterisks (*) are required.

Course Wireless Systems

Add a few sentences to tell people what your Page is about. This will help it show up in the right search results. You will be able to add more details later from your Page settings.

155

*Tell people what your Page is about...

Website (e.g.: your website, Twitter or Yelp links)

Choose a unique Facebook web address to make it easier for people to find your Page. Once this is set, it can only be changed once.

<http://www.facebook.com/wirlessonline>

Is Online E-learning a real organisation, school or government? ☐ Yes ☐ No
This will help people find this organisation, school or government more easily on Facebook.

[Save Info](#) [Skip](#)

Figure 3. Facebook page for course description

Photos 1 - 20 out of 37 | [Back to CSJ462 - Introduction to Database Systems's Photos](#) | [Edit Photos](#) | [Organize Photos](#) | [Add More Photos](#)

1 2 next

Extended Entity Relationship Model

Added 13 minutes ago • [Comment](#) • [Like](#)

[Share This Album](#)
[Post Album to Profile](#)

[Write a comment...](#)

Figure 4. Facebook page for course slides

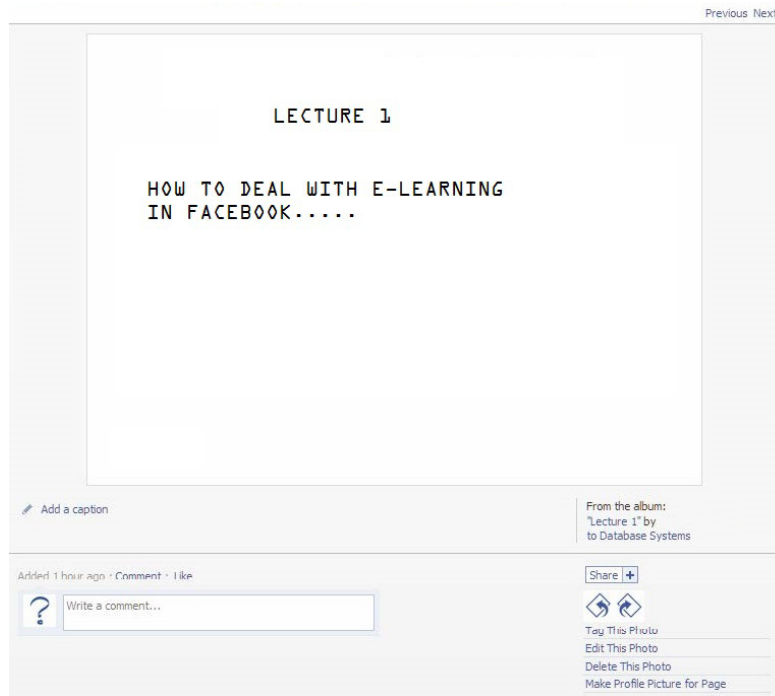


Figure 5. Facebook page for course slide presentation

The lecturer could make use of other Facebook features in the main profile of the course page to provide further support to the students:

- **Link** – teaching staff can add the reference materials on the web page with a link, such as reference articles, videos and so on.
- **Event** – teaching staff can create events for lectures and tutorials, so that student users will be notified and their main page will show the schedules of the lectures and tutorials whenever the students users log on Facebook.
- **Video** – if the lecture, tutorial or demonstration is recorded, it is possible to update it to the Facebook page, so that it is accessible easily by the students.

If the presentation file does not involve any transition effects, teaching staff could use the Facebook photo album web page to conduct the lecture/tutorial. The benefit was that if students wanted to raise any question and provide any feedback on the slide, they could post their comments for such slide and the teaching staff would be notified immediately. Such feature was especially useful to students who were passive in the class. The comments posted were specific to individual slide and it therefore facilitates the discussion among teaching staff and students.

Students could also access to the photo album with their own mobile devices, such as mobile phones. Although the devices were small, they support zooming and enabled the students to provide feedbacks or comments similar to a computer. For example, Figure 6 shows the same lecture note slide to be shown by an Apple iPhone and a LG mobile phone respectively.

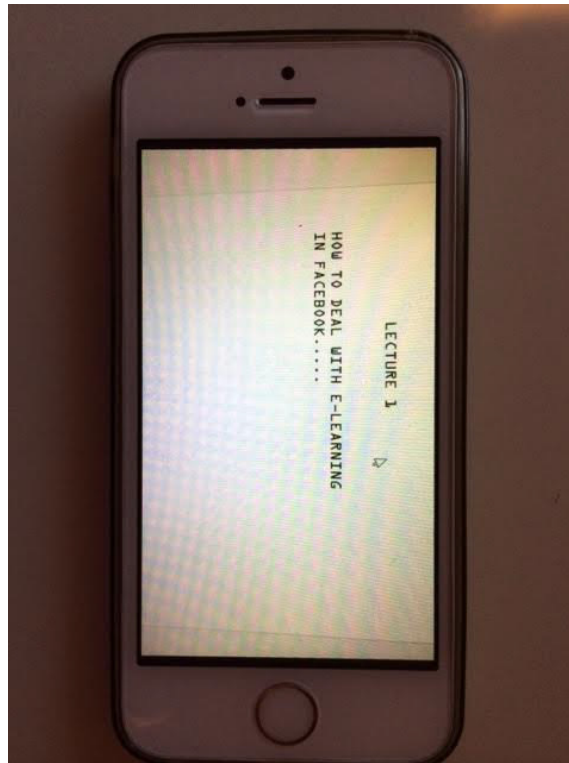


Figure 6. The image for a lecture note slide is shown by an Apple iPhone 5S. The same will also appear in a LG mobile phone

For slide with text in smaller typeface, most mobile phones enable users to zoom the images for better readability. When students wanted to leave comments or questions regarding the slide, they used their mobile device to do so. For example, Figure 7 shows the user interfaces of an Apple iPhone and a LG mobile phone, which enables Facebook student users to post comments to a slide.

In fact, mobile devices are capable of viewing the slide and enable students to leave comments or questions to particular slide. Upon receiving comments or questions, all members in the course, including teaching staff, would be notified. As soon as teaching staff received a notification emails from Facebook, they could click the embedded link that navigates the web browser to the referred slide, and leave another comment for the same slide as responses. Teaching staff could create quizzes to assess students' understandings of the lecture. For example, Figure 8 illustrate the use of Quiz Creator Facebook application by a teaching staff to create a quiz.

Create your own Quiz App!

Let's start by entering some basic info about your quiz. [Click here](#) to see what it looks like in the quiz.

What type of quiz do you want to make? (step 1)

☐ **Personality** Tells you what type of person you are. Example: What's your kissing style?

☒ **Trivia *new!*** Has right & wrong answers. Example: How well do you know Twilight?

Name Your Quiz:

This will be the title of your quiz app.

Quiz Description:

This shows up in the quiz directory.

Quiz Language:

This Quiz is For: ☒ Everybody ☐ Boys ☐ Girls ☐ Only my friends

Contains Alcohol: ☒ No ☐ Yes. Please restrict minors from viewing this quiz.

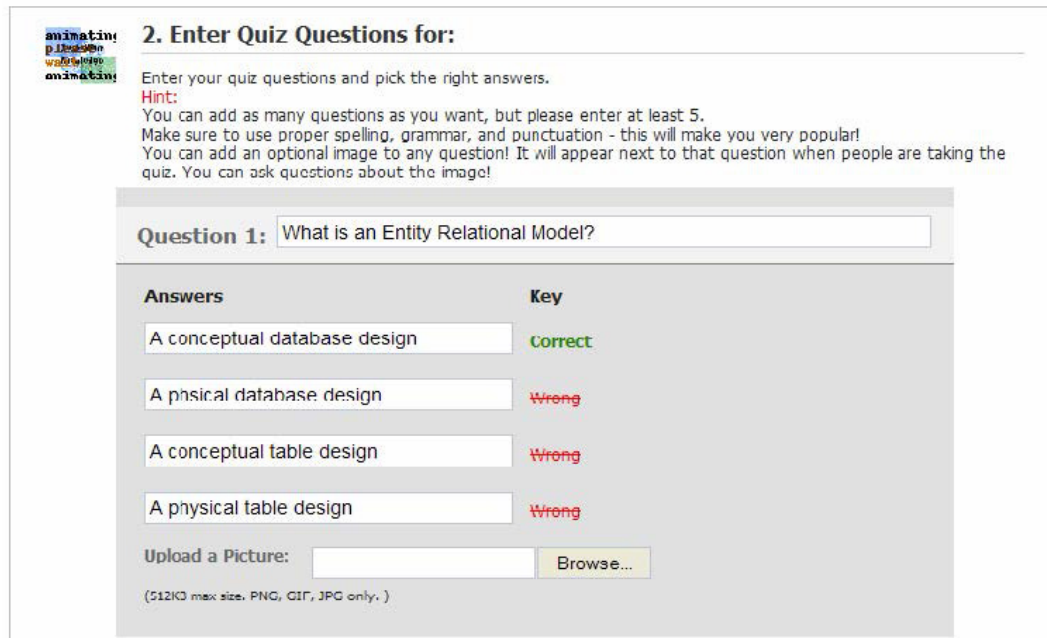
Upload a Picture:

Upload a picture for your quiz to make your quiz more popular! Max 512KB. PNG, GIF, JPGs only.

Facebook guidelines: applications may not promote, or contain content (including any advertising content) referencing, facilitating, promoting or using, the following: Adult content, including nudity, sexual terms and/or images of people in positions or activities that are excessively suggestive or sexual; Obscene, defamatory, libelous, slanderous and/or unlawful content; Content that infringes upon the rights of any third party, including copyright, trademark, privacy, publicity or other personal or proprietary right, or that is deceptive or fraudulent; Sale of liquor, beer, wine, tobacco products, ammunition and/or firearms. Exception: liquor, beer and wine are permitted in apps that are marked as "containing alcohol."; Gambling, including without limitation, any online casino, sports books, bingo or poker.

By pressing Next, I certify that I have read and agree to the Quiz Creator Terms of Service and the Platform Application Guidelines, and that I have the right to distribute these pictures and that they do not violate Facebook's Terms of Use.

Figure 7. Specify the quiz name and details with Quiz Creator



2. Enter Quiz Questions for:

Enter your quiz questions and pick the right answers.

Hint:
You can add as many questions as you want, but please enter at least 5.
Make sure to use proper spelling, grammar, and punctuation - this will make you very popular!
You can add an optional image to any question! It will appear next to that question when people are taking the quiz. You can ask questions about the image!

Question 1:

Answers	Key
<input type="text" value="A conceptual database design"/>	Correct
<input type="text" value="A physical database design"/>	Wrong
<input type="text" value="A conceptual table design"/>	Wrong
<input type="text" value="A physical table design"/>	Wrong

Upload a Picture:

(512KB max size. PNG, GIF, JPG only.)

Figure 8. Specify the quiz questions and answers with Quiz Creator

5. CONCLUSION

Facebook is the most popular social networking web site, and student Facebook users do not need any training on its usage. Besides, any computer and mobile device can easily access it with web browsing capability. Therefore, Facebook is therefore an excellent supplementary education framework that can replace some features of traditional classroom learning. In summary, the use of Facebook for education has a number of advantages. First, true cross platforms and cross devices such as computers and mobile devices support Facebook. Second, course-teaching materials are easily distributed. Third, blog-like discussion on individual items as well as online quizzes and assessments are supported. Fourth, it is user-friendly and no special trainings are required. All these provide some insights for one to develop a student friendly information-sharing platform.

REFERENCES

- [1] E-learning Systems, <http://www.bapsis.com/elearningsystems.htm>, [On-line; accessed 19-September-2014].
- [2] Craciunas, S. & Elsek, I, (2009) The standard model of an e-learning platform, Bucharest, Romania, (Chapter 2).
- [3] Dobre, I., (2010) Critical Study of the present e-learning systems, Academia Romana, Romania, (Chapter 2).
- [4] Edgar, R. W., (2005) Security in e-learning, Springer. Vienna University of Technology, Austria, (Chapter 1).
- [5] Iacob, N., (2010). Data replication in distributed environments, Proceedings of International Scientific Conference ECO-TREND: Brancusi University Targu Jiu, 629-634.
- [6] Jalal, A. & Ahmad, M., (2008). Security Enhancement for E-Learning Portal, Proceedings of International Journal of Computer Science and Network Security, Department of Computer Science City University, Peshawar, Pakistan, 41-45.

- [7] Kritzinger, E. & Solms S., (2006). E-learning: Incorporating Information Security Governance, Proceeding of Informing Science and IT Education Conference, Salford (Greater Manchester), England, 319-325.
- [8] Kumar, S. & Kamlesh, D., (2011). Investigation on Security in LMS Moodle, Proceedings of International Journal of Information Technology and Knowledge Management, Kurukshetra University, Kurukshetra, India, 233-238.
- [9] Przemek, S. (2007), PHP Session Security, Poland, (Chapter 1).
- [10] Smeureanu, I. & Isaila, N, The Knowledge Transfer Through E-Learning in Business Environment, Economy Informatics, 97-98.

AUTHORS

Mohammad Derawi received his diplomas in Computer Science engineering from the Technical University of Denmark where he received both a BSc (2007) and MSc (2009) degree. Derawi has pursued his PhD in information security at the Norwegian Information Security Laboratory (NISLab), Gjøvik University College (Norway). In the beginning of his PhD studies, he was a visiting researcher at the “Center for Advanced Security Research Darmstadt” (CASED, www.cased.de), Germany for an 8 months period. His PhD research interest included biometrics with specialization on gait recognition in mobile devices. Derawi was active in the 7th Framework European project “TrUsted Revocable Biometric IdeNtitiEs” (TURBINE, www.turbine-project.eu) and other main interests of areas include fingerprint recognition. Today he holds an Associate Professorship within Electronic Engineering and is specialised within Information Security, Biometrics and Micro-Controllers.



INTENTIONAL BLANK

A NEW HYBRID METRIC FOR VERIFYING PARALLEL CORPORA OF ARABIC-ENGLISH

Saad Alkahtani, Wei Liu, and William J. Teahan

School of Computer Science, Bangor University, Bangor, United Kingdom
{s.alkahtani,w.liu,w.j.teahan}@bangor.ac.uk

ABSTRACT

This paper discusses a new metric that has been applied to verify the quality in translation between sentence pairs in parallel corpora of Arabic-English. This metric combines two techniques, one based on sentence length and the other based on compression code length. Experiments on sample test parallel Arabic-English corpora indicate the combination of these two techniques improves accuracy of the identification of satisfactory and unsatisfactory sentence pairs compared to sentence length and compression code length alone. The new method proposed in this research is effective at filtering noise and reducing mis-translations resulting in greatly improved quality.

KEYWORDS

Parallel Corpus, Sentence Alignment for Machine Translation, Prediction by Partial Matching Compression

1. BACKGROUND AND MOTIVATION

The history of translation between natural languages can be traced back to the beginning of human culture, with its major mission being to expand the informativeness of one language, decrease the misunderstanding in dialogue, and even contribute to the growth of cultures [5]. Language translation is seen as a valuable social science oriented industry to help people develop international relationships.

Early pioneering machine translation systems were developed in the 1950s and 1960s [8]. Machine translation requires the development of computing technologies in the areas of computational linguistics and natural language processing. Machine and machine-aided translation are gaining in accuracy and popularity. Computer technology is essential to manage the large amounts of text available that may need to be translated.

Our research specifically explores the development of resources for Arabic-English translation, two important global languages. Arabic is spoken throughout the world and is the official language of 27 states, the third most after French and English. The term ‘Arabic’ when used in this paper refers to a variety of dialects belonging to the Central Semitic languages. English is the essential language of commerce and science and is the common lingua franca between many nations. Both languages are two of the six official languages of the United Nations.

Arabic is the primary language for over 380 million native Arabic speakers worldwide [15]. Computing research and applications designed for Arabic has increased in recent years. In particular, there have been a significant number of people accessing the Internet in Arabic. The majority of these users do not speak any language other than Arabic, which means they cannot easily access the vast variety of English information available. At the same time, global interest in Arabic countries, in culture, politics, economics, and other areas has expanded worldwide.

Language corpora have become increasingly important in natural language processing, and machine translation in particular. Corpora are an important resource often used for training purposes for statistical-based language modelling and machine translation. Large-scale parallel corpora are needed to construct statistical machine translation systems. Given the large number of Arabic speakers and the global importance of English, it is vital that translation between these languages be facilitated by the use of high quality parallel corpora. However, the structural differences between these languages present a challenge for machine translation. Arabic requires an altogether different treatment than European languages because of its unique morphology. Arabic and English are also different in a number of graphology aspects as Table 1 shows.

Table 1. A list of differences between the Arabic and English languages.

Graphology Aspects	Arabic Language	English Language
Written and Read	From right to left	From left to right
Capitalization	No	Yes
Size of Alphabet	28 letters	26 letters
Gender Differentiation	Verbs and sentence structures	No differentiation
Types of Sentences	Nominal and verbal	Verbal
Plural Forms	Singular, dual and plural	Singular and plural
Position of Adjective	After the noun	Before the noun

The use of parallel Arabic-English corpora to train statistical MT models provides an effective way for building MT systems. However, Arabic-English parallel texts of high quality are still very limited and are not available in satisfactory quantities, therefore most translations are performed manually, a time consuming and often error-filled process. Limitations of existing parallel corpora include incomplete data, untagged entries, with only limited text genres being available (such as news stories). In addition, many of the better quality corpora are not available for public use with fees in the thousands of dollars. For example, a list of corpora that were available from the Linguistic Data Corporation (LDC) in 2013 at the beginning of our research project is shown in Table 2 [12]. These costs are often unaffordable for most students, and also for many researchers or small research groups.

Table 2. Parallel Arabic-English Corpora as provided by the LDC in 2013 [12].

Corpora	Size (Words)	Price (US \$)
ISI Arabic-English Automatically Extracted Parallel Text	31M	\$4000
Uma Arabic English Parallel News Text	2M	\$3000
Arabic-English Parallel Translation	42K	\$3000
Multiple Translation Arabic (Part 1)	23K	\$1000
Multiple Translation Arabic (Part 2)	15K	\$1000
Arabic Newswire English Translation Collection	551K	\$1500
Arabic News Translation Text (Part 1)	441K	\$3000
GALE Phase 1 Arabic Broadcast News Parallel Text (Part 1)	90K	\$1500
IGALE Phase 1 Arabic Broadcast News Parallel Text (Part 2)	56K	\$1500
UN Bidirectional Multilingual	1M	\$4000

Another motivation behind our research is to develop techniques that would allow the construction of high quality parallel corpora that are free for everyone to use by improving the quality of the data as well as combining existing corpora, and by constructing much larger corpora, for example by using web scraping techniques. In order to achieve this task, we believe that a more accurate and robust metric than existing methods (such as sentence length) is needed for matching sentence pairs between languages.

This paper is organised as follows. In the next section, we review some of the work that is related to the present work. Note that not all of the related work has been included (especially for the use of sentence length as a metric for alignment) due to the many publications in this area. In section 3, the new hybrid metric is described. The experimental evaluation is described in section 4, with the conclusion in the final section.

2. RELATED WORK

In a parallel bilingual corpus, textual elements (e.g. paragraphs, sentences, phrases, words) alignment is an essential job for statistical machine translation. There have been a number of different approaches for sentence alignment such as sentence length, word co-occurrence, cognates, dictionaries and parts of speech etc. for a parallel bilingual corpus [13].

The sentence length metric assumes that the length for each sentence will be kept the same when it is translated from the source language into the target language. Gale and Church [6] aligned parallel sentences in English-French and English-German corpora based on a sentence length metric that required calculating the character length of all sentences. Gale and Church's [6] overall accuracies were 97% for English-German and 94% for English-French. Wu [21] aligned English-Chinese corpora by using sentence length values and reached an accuracy of 95%. Kay and Röscheisen [9] developed a program that combined word and sentence alignment and calculated the word probabilities by using the dice co-efficient. Haruno and Yamazaki [7] used a similar method plus a bilingual dictionary for aligning English-Japanese corpora. Papageorgiou, Cranias, and Piperidis [17] used the sentence alignment metric based on the highest matching part of speech tags and matches restricted to nouns, adjectives and verbs, and reached 99% accuracy. Simard, Foster and Isabelle [19] used cognate-based approaches and found that sentence length difference worked well for sentence alignment. However, Melamed [14] pointed out that because results were only reported for a relatively easy bilingual text, comparing two algorithms' performances in the literature is difficult. In addition, Brown *et al.* [3] calculated sentence length by using the number of words instead of the number of bytes or characters, which generated similar accuracies between 96% and 97%.

In the last decade, there have been few new proposals for sentence alignment for parallel bilingual corpora [22]. One disadvantage of existing sentence alignment algorithms is that it is less effective when linking corresponding sentences if they are one-to-many or many-to-one mutual translations [11].

Our approach, as well as using sentence length, also makes use of the strong correlation in compression code length (number of bits required to encode the text) between original sentences and accurately translated sentences. We show in this paper that this correlation can be used to evaluate and improve the quality of bilingual parallel corpora. If encoded into bit strings, almost all natural language text contain redundant bits that can be removed without affecting the information they carry. An observation by Behr *et al.* on natural languages indicates that all natural languages have similar cross entropies [2]. According to Shannon's information theory [18], a derived hypothesis from this observation will be that all natural languages can be encoded into the same length of bit strings for the same information if redundant bits are excluded.

Our work with using compression code length metrics for sentence alignment with Chinese-English corpora have shown they can be very effective[13]. Our idea of using compression code length as a sentence alignment metric hinges on the premise that the compression of co-translated text (i.e. documents, paragraphs, sentences, clauses, phrases) should have similar code lengths [2]. This is based on the notion that the information contained in the co-translations will be similar. Since compression can be used to measure the information content, we can simply look at the ratio of the compression code lengths of the co-translated text pair to determine whether the text is aligned. That is, if you have a text string (i.e. document, paragraph, sentence, clause or phrase) in one language, and its translation in another language, then the ratio of the compression code lengths of the text string pair should be close to 1.0. This approach and the new hybrid approach are described in more detail in the next section.

3. A NEW METRIC FOR CHECKING THE QUALITY OF PARALLEL SENTENCE PAIRS

In this section, we describe how we use a distance metric based both on sentence lengths and on compression code lengths (using the Prediction by Partial Match (PPM) compression scheme) in order to check the quality of the sentence pairs.

3.1. PPM Compression Code LengthMetric

The Prediction by Partial Matching (PPM) compression scheme, first proposed by Cleary and Witten in 1984[4], predicts the next symbol or character from a fixed order context. The context models are adaptively updated as the text is processed sequentially using an online adaptive process. For both Arabic and English text [1, 9], experiments have shown that order 5 models (using fixed order contexts of length 5) perform best at compressing the text using the PPMD variant of the algorithm developed by Howard in 1993 based on the PPMC variant devised by Moffat[20]. The main difference between PPMC and PPMD (and other variants PPMA and PPMB) is the calculation of the *escape* probability when the model needs to back off to lower order models if a symbol is not predicted by a higher order model.

Formally, the estimation of the *escape* probability for PPMD is $e = t_d / 2n$ and for the symbol probability is $p(\varphi) = (2c(\varphi) - 1) / 2T_d$, where: d is the current coding order of a model; φ is an upcoming symbol ($\varphi = x_{n+1} \in A$); s_d is the current context $s_d = x_n, \dots, x_{n-d+1}$; $c_d(\varphi)$ is the number of times that the symbol φ in the context s_d ; t_d is the total number of unique symbols that occur after the context s_d ; T_d is the total number of times that the context s_d has been seen with $T_d = \sum c_d(\varphi)$; e is the *escape* probability; and $p(\varphi)$ is the probability of the upcoming symbol φ .

In this paper for our experiments, we use PPMD with $d = 5$ since as stated, experience shows that this is most effective for compressing English and Arabic text and performs better than PPMA, PPMB and PPMC. Table 3 shows how the probabilities are estimated using PPMD when the model has been trained on the sample text string “سبيل السبيل”. The table shows the predictions, frequency counts c and probability estimates p for the order $k = 3$, $k = 2$, $k = 1$, $k = 0$ and $k = -1$ PPMD contexts (where k is the order of the model or context length). For example, only one symbol has been predicted in the single order 3 context – this has occurred once in the training text, and therefore its probability estimate that it will occur again is $3/4$ and the probability estimate that a previously unseen symbol in this context will occur instead is $1/4$ therefore necessitating the use of lower order models in order to estimate the probability of the unseen symbol. The model will keep on escaping down until it encounters a context where the symbol has been seen before or the symbol will be encoded using the default $k = -1$ context where every symbol is estimated with equal probability $1/|A|$ where $|A|$ is the size of the alphabet.

Table 3. PPMD order 3 model after processing the text string “سبيل للسلاسل”.

Order $k = 3$			Order $k = 2$			Order $k = 1$			Order $k = 0$		
Prediction	c	p	Prediction	c	p	Prediction	c	p	Prediction	c	p
ل → سسي	2	3/4	ي → سب	2	3/4	ب → س	2	3/8	→ س	4	7/30
									→ ب	2	3/30
						→ ل	2	3/8	→ ي	2	3/30
						→ esc	2	2/8	→ ل	6	11/30
→ esc	1	1/4	→ esc	1	1/4				→ ا	1	1/30
						ي → ب	2	3/4	→ esc	5	5/30
ا → يل	1	1/4	ا → ي	2	3/4				Order $k = -1$		
→ ا	1	1/4				→ esc	1	1/4	Prediction	c	p
→ esc	2	2/4	→ esc	1	1/4				→ A	1	1/ A
ل → زال	1	1/2	ل → زل	1	1/4	ل → ي	2	3/4			
			→ ا	1	1/4						
→ esc	1	1/2	→ esc	2	2/4	→ esc	1	1/4			
س → للل	1	1/2	ل → لل	1	1/4	ل → ل	2	3/12			
			→ ا	1	1/4	→ س	3	5/12			
			→ esc	2	2/4	→ ا	1	1/12			
→ esc	1	1/2				→ esc	3	3/12			
ل → للس	1	1/2	ل → لس	2	3/6						
			→ ب	1	1/6						
→ esc	1	1/2	→ esc	2	2/6						
س → لمـل	2	3/4	س → مل	2	3/4						
→ esc	1	1/4	→ esc	1	1/4						
ل → ملس	1	1/4									
→ ب	1	1/4									
→ esc	2	2/4									
ي → لسب	1	1/2									
→ esc	1	1/2									

3.2.Code Length Ratio Distance Metric for Matching Sentences

The term code length refers to the size (in bytes) of the compressed output file produced by the PPM compression algorithm. When using PPM to compress Arabic or English text, the code length is a measure of the cross-entropy of the text, which is the average size (in bytes) per character for the compressed output string. Theoretically, the cross-entropy is estimated as follows:

$$H(S) = \frac{1}{k} \log_2 p(S) = -\frac{1}{k} \sum_{i=1}^k -\log_2 p(x_i | x_1, \dots, x_{k-1})$$

where $H(S)$ is the average number of bits to encode the text and k is the order of the model (e.g. 5 for the models used in this paper).

Note that the compression code length, the number of bits required to encode the text string losslessly, so that it can be unambiguously decoded, can be expressed simply as $nH(S_L)$.

The ratio of the compression code lengths of the parallel text strings for languages E (English) and A (Arabic) is defined as follows:

$$R(S^E, S^A) = \frac{n}{m} \times \frac{H(S^E)}{H(S^A)}$$

where S^E is an English text string with length n and S^A is an Arabic text string with length m . The code length ratio (CR) is defined as:

$$CR = \max\{R^{E,A}, R^{A,E}\}$$

Liu *et al.* [13] have shown that CR is a more effective distance metric for sentence alignment of Chinese-English parallel corpora than a distance metric based on sentence length. A primary purpose of the research reported in this paper was to investigate whether this would also be the case for Arabic-English parallel corpora.

3.3. Sentence Length Ratio Distance Metrics for Matching Sentences

Automatically generated bilingual corpora often have a large number of noisy sentence pairs. Consequently, researchers have devised various methods to filter noisy sentences from parallel corpora [10]. However, for our experiments discussed in Section 4, we have found a new technique based on a combination of the compression Code Length Ratio (CR) described above and the standard Sentence Length Ratio (SLR) described by Mújdricza-Maydt [16] is the most effective for Arabic-English sentence pairs in order to achieve a high-quality corpus as a result.

The Sentence Length Ratio (SLR) for a pair of translation sentences for Arabic and English can be calculated as follows:

$$SLR = \max\left\{\frac{L^A}{L^E}, \frac{L^E}{L^A}\right\}$$

where L^A is the length of the text for Language A.

4. EXPERIMENTAL EVALUATION

4.1. Developing the Test Corpora

For our experimental evaluation, two parallel Arabic-English test corpora were created. A large corpus (Corpus A) was first created containing fifty-eight million words that was collected from two online sources Al Hayat (<http://www.alhayat.com>) and OPUS (<http://opus.lingfil.uu.se>) with permission obtained from the owners of the data. OPUS is an open source parallel corpus that provides a large collection of translated texts from the web. All the online data was collected automatically and as a result the original texts are not of high quality. However, a primary purpose of our research is to develop a more reliable collection based on this and other data with poor quality translations filtered out using our sentence matching metrics.

A second much smaller test corpus (Corpus B) was created containing 10,000 translations judged satisfactory and 2,000 translations judged unsatisfactory. These were manually selected from Corpus A and formed the ground truth data for our experiment.

The files in Corpus A were also classified into 13 categories such as Books, Business, Cinema, Conferences, Crimes, Decisions, Economy, Geographies, Issues, Law, Politics, Reports and Stories as described in Table 4. The number of Arabic and English characters and words in each

of the categories are also shown in the table. In total, 58,380,784 words were collected comprising 27,775,663 Arabic words and 30,808,480 English words.

Table 4. Character and word counts for test Corpus A.

Categories	Arabic Characters	English Characters	Arabic Words	English Words
Books	10,574,252	7,242,426	931,836	1,079,699
Business	26,367,126	17,987,925	2,289,276	2,624,274
Cinema	61,557,926	36,482,892	7,919,902	8,127,509
Conferences	21,696,083	15,129,972	1,879,527	2,215,857
Crimes	10,147,866	6,473,170	933,842	1,005,221
Decisions	15,863,975	10,822,315	1,397,181	1,605,851
Economy	25,962,438	17,760,514	2,266,424	2,599,651
Geographies	16,096,053	10,924,063	1,392,099	1,595,115
Issues	11,390,107	6,937,792	1,051,195	1,042,316
Law	16,083,105	10,936,231	1,407,292	1,597,873
Politics	23,427,958	15,675,917	2,035,969	2,304,233
Reports	15,960,285	10,819,195	1,388,457	1,590,056
Stories	29,703,105	20,294,105	2,882,663	3,420,825
Total	284,830,279	187,486,517	27,775,663	30,808,480

4.2. Compression Experiments

Preliminary compression experiments were conducted to determine if the *CR* compression code length measure would be effective as a metric for measuring the quality of translation between sentence pairs of Arabic and English.

Standard PPM is an adaptive technique with its language models starting from null when the beginning of a text string is processed. The context frequency counts from which the probability estimates are made are then updated as the text string is processed sequentially. For longer text strings (such as documents and paragraphs), the PPM algorithm will usually have enough text in order to train its models effectively so that higher order contexts are being used for most predictions with less need to escape down to lower order contexts.

One obvious concern when using PPM code lengths is that sentences may not be long enough in order that more reliable probability estimates can be made for the *CR* calculation. A simple expedient in overcoming this difficulty is to prime the PPM models prior to the compression. We can use a large corpus that is representative of the language (such as the Brown corpus for English and the CCA corpus for Arabic) in order to prime the models prior to the compression being performed (i.e. ‘train’ the models using the priming text). This approach has been found to be very effective, for example, when using compression code length based metrics for sentence alignment between Chinese and English [13].

The purpose of the preliminary experiments described in this section were to determine how effective priming of the PPM models was for compressing Arabic sentences, and also how effective the primed PPM compression method as a sentence matching metric. A key requirement of using the *CR* metric is that the compression code lengths in the two different languages should be the same for sentences that are co-translations of each other. The intuition is that if the sentences are satisfactory co-translations, then they should convey exactly the same amount of information. Since compression code length is an effective method for measuring information (see [20] for several references), then we would expect that roughly 50% of the compression code

lengths of sentences in one language to be longer than compression code lengths of sentences in the other language, and vice versa.

Clearly, this correlation would not be expected for sentence lengths. It is quite common that English sentences are shorter than their co-translation counterparts other languages (although this is not the case when compared with the Arabic sentences as reported below in this section). However, this should not be the case for compression code lengths if our intuition about the correlation between information is correct. If we find that the compression code lengths do not correlate, then the reason for this is more likely to be as a result of a less effective compression algorithm being used for one language resulting in a less accurate estimate of the information contained in the sentence.

In a preliminary experiment, 10 sample sentence pairs in Arabic and English were randomly chosen from Corpus A. The 10 sample sentence pairs that we used are shown in Table 5

Table 5. Sample sentence pairs that were used in the initial compression experiments.

Sent. ID	Arabic Sentence	English Sentence
1	موضوعي اليوم جدِّي ولكن أبدأه بطريقة قديمة استدرأجا للقارئ.	My topic today is a serious one, but I will begin with an old anecdote, to lure the reader in.
2	الوقوف في الجانب الصحيح من التاريخ محاولة لتبرير الحروب العادلة.	Standing on the right side of history represents an attempt to justify just wars.
3	كنت أهاذرها إلا أنها فكرت، وسألتني هل أعتقد حقاً أن البكاء وسيلة أفضل لكسب الأصوات.	I was joking with her, but she took it seriously and asked me whether I really believed that crying was a better way to win votes.
4	هكذا الدنيا، جنازة أو جواز كما يقول اللبنانيون.	Such is life, a wedding or a funeral, as the Lebanese say.
5	هذا الرجل يقول: إنه يعرف ما لا يعرف قضاة لجنة الانتخابات	This man is saying that he knows something the judges on the Election Commission do not know.
6	فلندع مجدداً رباتنا الربيع، ونحصي الخيبات، ومرارات صيفٍ يائس.	So let us once again claim to be the precursors of the spring, count the disappointments and tally the bitterness of a wretched summer.
7	وأن الذين توجهوا بعدها إلى القصر تصورا أن الرجل يجلس خلفه في انتظارهم!	Those who subsequently headed to the palace, truly imagined that the man was sitting there, waiting for them!
8	هو أخيراً ارتاح، بعد رحلة الآلام والأمال والنكبات والانتصارات، وترك لنا جميعاً مثلاً يُحتذى.	He has finally rested, after a journey of pains, hopes, disasters and triumphs, and left us all an example to be followed.
9	سيكون هناك شيء جديد تسمعه.	You will have something new to listen to it.
10	العسكريون أكثر تمسكاً بالدولة المدنية الديمقراطية والعلمانية.	The militaries are not more persistent on the civil, democratic and secular state.

The results of compressing these sentences using the PPM compression scheme are shown in Table 6. The table lists the number of bytes that various variants of PPM produced as compressed output. For example, for sentence pair with id 1 (i.e. the first in Table 5), the WOT variant required 69 bytes to compress the Arabic sentence, compared to 69 bytes to compress the English sentence. In contrast, the sentence lengths are very different – the Arabic sentence is 59 characters (bytes) long compared to 95 characters for the English sentence.

Order 5 PPMD (as described above in Section 3.1) was used for these experiments. The WOT variant is for PPM without priming (i.e. no training). The WT variant is for PPM with priming. In

this case, the PPM model was trained on the Brown Corpus prior to compressing the English text, whereas the PPM model was trained on the CCA Corpus prior to compressing the Arabic text. The WTPP variant used the same priming approach as for the WT variant, but also adopted a pre-processing algorithm to convert the UTF-8 encoded Arabic text into a number string before it was compressed by the PPMD5 compressor. This approach is described in detail in [1]. This leads to significantly better compression as a result for Arabic text and therefore leads to a better estimate of the information contained in the Arabic sentence.

Table 6. Compression results of the sample sentences. The PPMD5 compression code length results list the size in bytes of the compressed output produced by various variants of the PPMD5 compressor.

Sentence ID	Sentence Length		PPMD5 Compression CodeLength					
			(WOT)		(WT)		(WTPP)	
	Arabic	English	Arabic	English	Arabic	English	Arabic	English
1	59	95	69	69	32	29	26	29
2	68	82	62	62	31	22	24	22
3	84	132	83	88	41	35	31	35
4	53	59	55	50	32	19	27	19
5	58	94	59	64	25	24	20	24
6	67	136	70	93	39	37	31	37
7	72	110	72	76	33	30	26	30
8	93	123	87	86	43	37	35	37
9	27	45	36	38	13	14	11	14
10	63	83	61	61	28	23	20	23

From the table, we can see there is a clear mis-match as expected between the Arabic and English sentence lengths. This provides clear evidence that metrics based on techniques well founded in information theory (as is the case for compression code length based metrics) have merit since they lead to better correlation.

The WOT variant does a surprisingly good job of matching the sentences with the Arabic bytes size being close to the English bytes size. However, again in most cases, the number of bytes of the compressed English output is greater than the number of bytes of the compressed Arabic output. For the WT variant, the opposite story is now the case – the compressed Arabic bytes is now usually greater the compressed English bytes. This indicates that the compression method being used for the Arabic text is probably not as well tuned as is the case for the English scheme (since the use of PPM for compressing English text has been fine-tuned over many years [20]). This problem was addressed in recent research on the compression of Arabic text [1] where it was found that using pre-processing techniques significantly improves PPM-based compression for Arabic in many cases by over 25%. When we apply these techniques (i.e. this is what we call the WTPP variant), then a more desired set of mixed results is achieved (where code lengths are sometimes greater for Arabic and sometimes greater for English).

In order to investigate this further and to confirm whether we have a compression method for Arabic text that produces compression code lengths that correlate well with those produced by the compression method being used for English text, we conducted further experiments using the WTPP PPM variant on the whole of the test Corpus A and in each of the categories as well. The results are listed in Table 7. The percentage of sentence pairs for which the Arabic sentence lengths are greater than for their English counterparts is shown in column 2. For example, for the Books category, it was found that only 8.55 % of the Arabic sentences are longer. In contrast, the results in the third column, which lists the percentage of sentence pairs for which the Arabic compression code lengths are greater than for their English counterparts, show that the comparison is more even, with most results being around the 50% mark, except for the Crimes

category with 62.48%. Due to this result, the nature of the sentences in this category should be investigated further.

Table 7. Percentage of Arabic sentences lengths or compression code lengths greater than their English sentence counterparts for the test Corpus A.

Categories	% of Arabic sentence lengths that are greater	% of Arabic compression code lengths that are greater
Books	8.55	54.96
Business	16.58	56.93
Cinema	35.43	44.94
Conferences	17.09	56.26
Crimes	21.88	62.48
Decisions	7.93	52.74
Economy	16.80	57.02
Geographies	14.79	58.97
Issues	16.30	53.71
Law	15.40	56.73
Politics	16.23	55.25
Reports	16.53	58.06
Stories	11.77	48.79
Average	16.56	55.14

These results provide reassuring evidence that the compression methods we have adopted produce the desired (and necessary) correlated data for the subsequent experiments we conducted that are described in the next section.

4.3. Analysing the quality of translations in the test corpora

Experiments were performed using the ground truth data in Corpus B to determine the best thresholds and combinations for the *CR* and *SLR* metrics in order to accurately filter out the unsatisfactory translations. For the *CR* calculations listed there, the WTPP PPMD5 variant (as stated, which was primed on the CCA corpus) was used to compress the Arabic text sentences, whereas standard PPMD5 primed on the Brown corpus was used for the English text sentences.

Various thresholds were applied firstly using *SLR* by itself, secondly using *CR* by itself, and thirdly by applying the same threshold to both *SLR* and *CR* together. If the calculated distance metric exceeded the threshold value(s), then the translation sentence pair was judged to be unsatisfactory, otherwise it was judged to be satisfactory.

The results of how accurate the filtering process was against the ground truth data in test Corpus B are shown in Table 8. The table shows the threshold values that were used for both the *SLR* and *CR* calculations in the leftmost column. The accuracy results are then provided in the subsequent columns. (This is the percentage of correct classifications made by the *SLR*, *CR* or *SLR&CR* metrics where a correct classification is made when the metric at a specific threshold judges the sentence pair to be satisfactory or unsatisfactory and this matches the ground truth judgment). The results are split for the satisfactory and unsatisfactory sentence pairs, with the average results provided in the final columns.

Table 8 shows, for example, that *SLR* with a threshold of 2.5 or higher is able to accurately classify 100% of the satisfactory translations whereas the threshold where this occurs for *CR* is at 2.25. For the unsatisfactory translations, 100% of these will be identified using *SLR* if the threshold is set at 1.5 or less, whereas the highest accuracy for *CR* is 97.45% when the threshold

is set as low as 1.25 (meaning most sentence pairs will be rejected). The only calculation that results in an average accuracy of 100% for all sentence pairs (both satisfactory and unsatisfactory) occurs when both *SLR* and *CR* are combined together with a threshold of 2.5. Figure 1 shows the tendencies of the classification of the satisfactory translations and unsatisfactory translations for test Corpus B using different threshold values.

Table 8. Comparison of accuracies among different threshold values when using the different sentence matching metrics on test Corpus B.

Threshold Values	10000 Satisfactory Translations Accuracies(%)			2000 Unsatisfactory Translations Accuracies(%)			Average Accuracies(%)		
	<i>SLR</i>	<i>CR</i>	<i>SLR&CR</i>	<i>SLR</i>	<i>CR</i>	<i>SLR&CR</i>	<i>SLR</i>	<i>CR</i>	<i>SLR&CR</i>
1.25	20.29	89.91	17.11	100	97.45	100	60.15	93.68	58.56
1.50	62.35	97.86	61.10	100	78.05	100	81.18	87.96	80.55
1.75	88.15	99.24	87.58	99.95	43.40	100	94.05	71.32	93.79
2.00	96.50	99.76	96.30	99.35	24.85	100	97.93	62.31	98.15
2.25	98.90	100	98.90	98.45	15.00	100	98.68	57.50	99.45
2.50	100	100	100	97.20	11.40	100	98.60	55.70	100
2.75	100	100	100	70.25	7.35	72.30	85.13	53.68	86.15
3.00	100	100	100	50.40	4.95	51.80	75.20	52.48	75.90
3.25	100	100	100	31.85	3.30	32.80	65.93	51.65	66.40
3.50	100	100	100	19.85	2.15	20.35	59.93	51.08	60.18

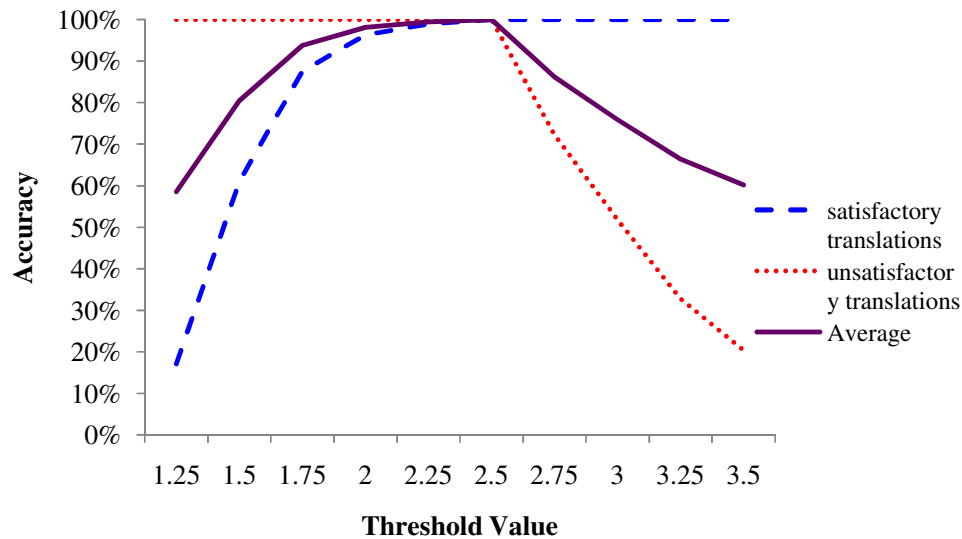


Figure 1. Tendencies of the classification of satisfactory translations and unsatisfactory translations for test Corpus B with different threshold values.

A further experiment was conducted to investigate whether different threshold values are more effective when using the combined *SLR&CR* technique. Table 9 displays the accuracy results matrix of the experiments on the overall accuracy averages on the same 10000 satisfactory translations and 2000 unsatisfactory translations in test Corpus B used in the previous experiment. In the table, the *SLR* threshold value is shown across the top row, and the *CR* threshold value is shown down the left column, both ranging from 1.25 up to 3.50. The table shows that 100%

accuracy is achieved using threshold values 2.50 and higher for *SLR* combined with 2.25 and higher for *CR*.

Table 9. The accuracy results matrix for test Corpus B using threshold values of *SLR* and *CR* from 1.25 to 3.50.

<i>SLR</i> <i>/CR</i>	1.25 (%)	1.50 (%)	1.75 (%)	2.00 (%)	2.25 (%)	2.50 (%)	2.75 (%)	3.00 (%)	3.25 (%)	3.50 (%)
1.25	17.11	58.01	82.33	88.42	89.62	89.91	89.91	89.91	89.91	89.91
1.50	19.58	61.10	86.64	94.81	97.05	97.86	97.86	97.86	97.86	97.86
1.75	20.13	61.95	87.58	95.82	98.17	99.24	99.24	99.24	99.24	99.24
2.00	20.27	62.29	88.00	96.30	98.66	99.76	99.76	99.76	99.76	99.76
2.25	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100
2.50	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100
2.75	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100
3.00	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100
3.25	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100
3.50	20.29	62.35	88.15	96.50	98.90	100	100	100	100	100

Another experiment was devised to determine how much of the larger test Corpus A would be classified as satisfactory or unsatisfactory using various *CR* threshold values (from 1.25 to 3.50) when the *SLR* threshold value was set at 2.5. The results of this experiment are shown in Table 10. The table shows the number classified in each category (in the columns labelled “Amount”) and the corresponding percentages. For example, a threshold value of 2.50 for both *SLR* and *CR* results in 8.18% of test Corpus A being labelled unsatisfactory (and therefore candidates for being removed from the corpus).

Table 10: Percentages of satisfactory and unsatisfactory translations for test Corpus A when the *SLR* threshold is set at 2.5.

Threshold <i>CR</i>	Satisfactory Translations		Unsatisfactory Translations	
	Amount	Percentage (%)	Amount	Percentage (%)
1.25	1313387	72.14	507234	27.86
1.50	1559275	85.65	261346	14.35
1.75	1626973	89.36	193648	10.64
2.00	1650374	90.65	170247	9.35
2.25	1665709	91.49	154912	8.51
2.50	1671768	91.82	148853	8.18
2.75	1674677	91.98	145944	8.02
3.00	1675700	92.04	144921	7.96
3.25	1676166	92.07	144455	7.93
3.50	1676311	92.07	144310	7.93

Figures 2, 3, 4 and 5 show correlations for the sentence length and code length metrics for test Corpus A. Figures 2 and 3 illustrate the sentence lengths and code lengths of Arabic and English sentences classified as unsatisfactory for the test Corpus A and show an obvious split in the plot due to 1:2 and 2:1 type mismatches.

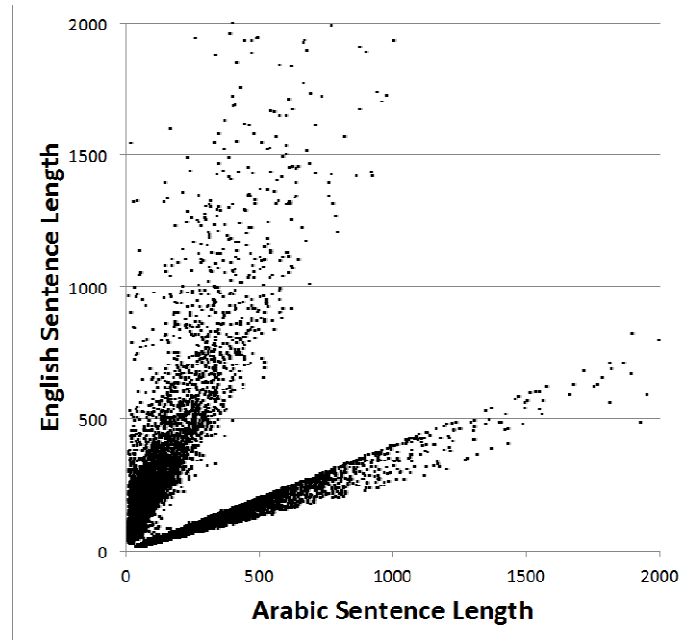


Figure 2. Sentence length correlation for test Corpus A for sentence pairs classified as unsatisfactory.

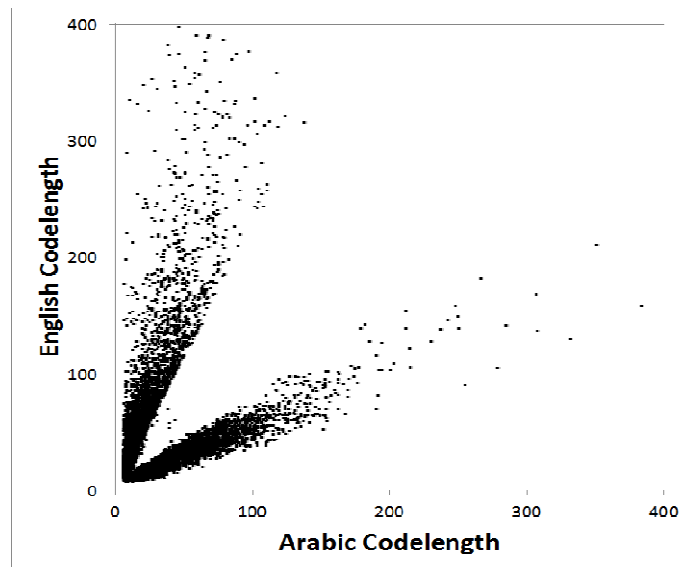


Figure 3. Code length correlation for test Corpus A for sentence pairs classified as unsatisfactory.

In contrast, Figures 4 and 5 illustrate sentence lengths and code lengths of Arabic and English for the translations classified as satisfactory for test Corpus A and show a strong correlation between both sentence lengths and compression code lengths.

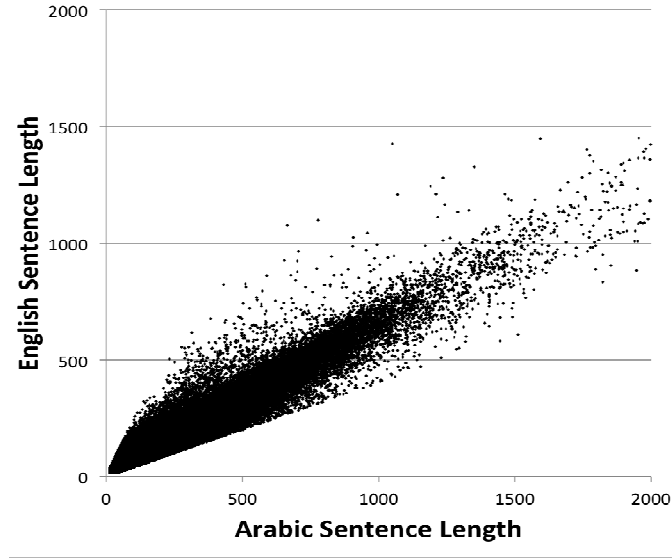


Figure 4. Sentence length correlation for test Corpus A for translations classified as satisfactory.

For defining what is a satisfactory translation in this case, it was decided if the values of *SLR* were less than 2.5 and *CR* less than 2.25 for a pair of translation sentences, then it is classified as a satisfactory translation, otherwise it is classified as an unsatisfactory translation as per Figures 2 and 3.

The unsatisfactory translations might be caused by errors in alignment between Arabic and English sentences which may include non-literal translations and therefore result in significant differences between the sentence pair. English sentences containing websites or abbreviations such as USA (United States of America), UNCTAD (United Nations Conference on Trade and Development Abbreviation) might also lead to mistranslations [10].

The flowchart in Figure 6 shows how the new hybrid metric was applied in this manner. If the *CR* threshold of 2.25 was exceeded, or the *SLR* threshold of 2.5 was exceeded, then the sentence pair would be rejected (i.e. classified as unsatisfactory), otherwise the translation was accepted (i.e. classified as satisfactory).

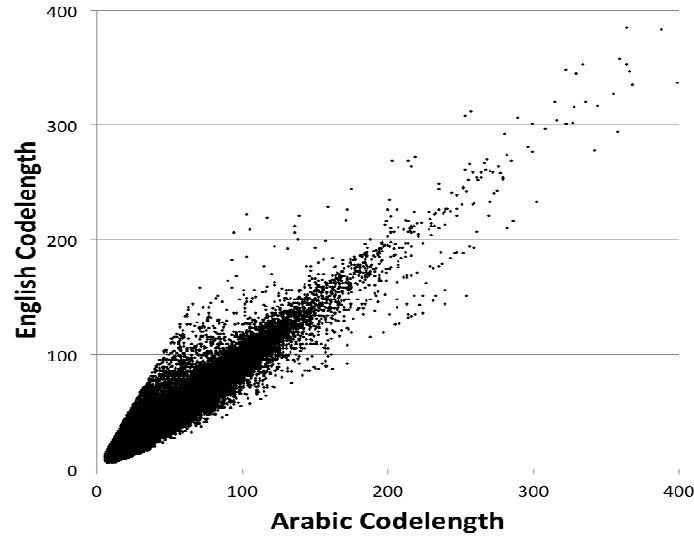


Figure 5. Code length correlation for test Corpus A for translations classified as satisfactory.

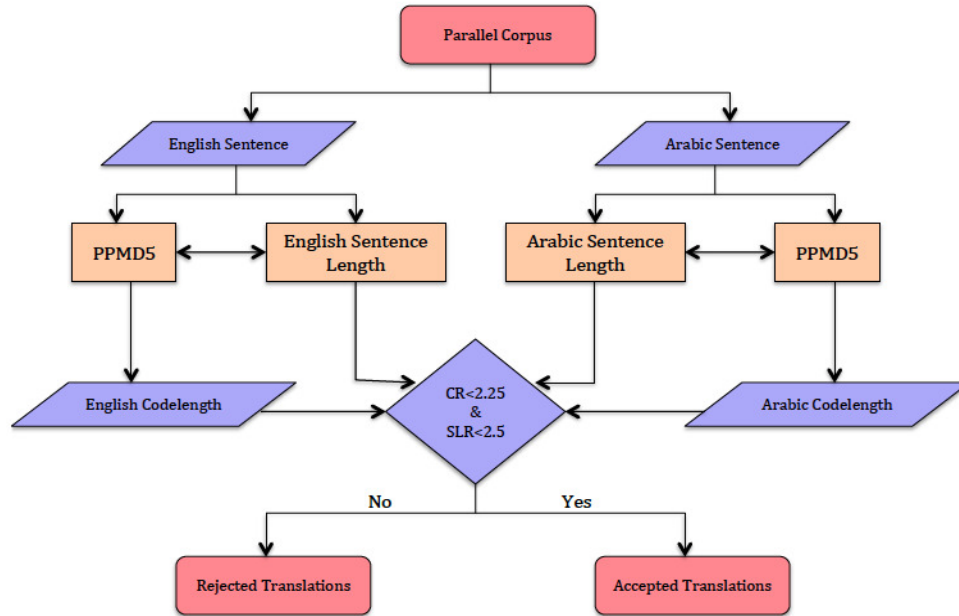


Figure 6. Flow chart of how the new hybrid sentence matching metric based on both compression code length and sentence length was applied to test Corpus A.

5. CONCLUSION

Verification is an essential step in order to ensure a high quality corpus. In this paper, we have described a new method to check how well the sentences match in a parallel corpus. The method is based on the combination of two distance metrics, sentence length ratio (*SLR*) and compression code length ratio (*CR*). A threshold mechanism can be used to filter out unsatisfactory translations when either the *SLR* or *CR* values have been exceeded. Experiments with a small sample of sentence pairs from a test Arabic-English corpus containing ground truth judgments, which were manually judged to be satisfactory or unsatisfactory translations, show that a combination of both

SLR and *CR* distance metrics performs better at classification than using a single distance metric by itself.

There is also other important verification tasks that are often overlooked not described here that need to be done. For example, a single check on document sizes is crucial (e.g. ensuring no zero byte documents, and removing unusually large documents if appropriate). Checking for self-plagiarism (ensuring that documents do not contain strings repeated in other documents) is also essential (especially for corpora containing news stories since it is a common practice for these types of documents to contain material copied from other news stories). We have found that the compression code length metric described here is also effective at classifying the quality of translation not just at the sentence level, but also at the document, paragraph and clause levels, and these should also be checked when verifying a parallel corpus.

REFERENCES

- [1] Alhawiti, K., (2014)“Adaptive Models of Arabic Text”, *PhD Dissertation, Bangor University*.
- [2] Behr F. H., Fossum V., Mitzenmacher M., Xiao D., (2003)“Estimating and Comparing Entropy across Written Natural Languages Using PPM Compression”, *Proceedings of Data Compression Conference*, p416.
- [3] Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R., (1993)“The Mathematics of Machine Translation: Parameter Estimation”, *Computational Linguistics*, Vol. 19, pp263-312.
- [4] Cleary, J. G. & Witten, I. H., (1984)“Data Compression Using Adaptive Coding and Partial String Matching”,*IEEE Transactions on Communications*, Vol. 32, No. 4, pp396-402.
- [5] Fantechi, A., Gnesi, S., Carenini, M., Vanocchi, M., Moreschini, P., (1994)“Assisting Requirement Formalization by Means of Natural Language Translation”, *Formal Methods in System Design*, Vol. 4, No. 3, pp243-263.
- [6] Gale, W.A. & Church, K.W., (1993)“A Program for Aligning Sentences in Bilingual Corpora”,*ACL’93 29th Annual Meeting*, pp177-184.
- [7] Haruno, M. & Yamazaki, T., (1996)“High-performance Bilingual Text Alignment Using Statistical and Dictionary Information”, *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics*, pp131-138.
- [8] Hutchins, W.J.,(1994) “The Encyclopaedia of Languages and Linguistics”, *ed. R.E.Asher, Oxford: Pergamon Press*, Vol. 5, pp2322-2332.
- [9] Kay, M. & Röscheisen, M., (1993)“Text-translation Alignment”,*Computational Linguistics*,Vol. 19, pp121-142.
- [10] Khadivi, S. & Ney, H., (2005)“Automatic Filtering of Bilingual Corpora for Statistical Machine Translation”,*Natural Language Processing and Information Systems*, Vol. 3513, pp263-274.
- [11] Kutuzov, A., (2013)“Improving English-Russian Sentence Alignment through POS Tagging and Damerau-Levenshtein Distance”,*Association for Computational Linguistics*, pp63-68.
- [12] Linguistic Data Consortium, <http://catalog.ldc.upenn.edu>
- [13] Liu, W., Chang, Z., Teahan, W., (2014)“Experiments with Compression-based Methods for English-Chinese Sentence Alignment”,*2nd International Conference on Statistical Language and Speech Processing*,pp70-81.
- [14] Melamed, I.D., (2000)“Models of Translational Equivalence among Words”,*Computational Linguistics*, Vol. 26, No. 2, pp221-249.
- [15] Mubarak, H., Darwish, K., Adly, N., (2014)“Using Twitter to Collect a Multi-Dialectal Corpus of Arabic”,*EMNLP 2014 Workshop on Arabic Natural Language Processing*.
- [16] Mújdricza-Maydt, É., Körkel-Qu, H., Riezler, S., Padó, S., (2013)“High-Precision Sentence Alignment by Bootstrapping from Word Standard Annotations”,*The Prague Bulletin of Mathematical Linguistics*, Vol. 99, pp5-16.
- [17] Papageorgiou, H., Cranias, L., Piperidis, S., (1994) “Automatic Alignment in Corpora”,*Proceedings of 32nd Annual Meeting of Association of Computational Linguistics*, pp334-336.
- [18] Shannon, C.E., (1948)“A Mathematical Theory of Communication”, *Bell System Technical Journal*, Vol. 27, pp379-423 &pp623-656.

- [19] Simard, M., Foster, G.F., Isabelle, P., (1992) "Using Cognates to Align Sentences in Bilingual Corpora", *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pp67–81.
- [20] Teahan, W., (1998) "Modelling English Text", *PhD Dissertation, University of Waikato, New Zealand*.
- [21] Wu, D., (1994) "Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria", *ACL'94 32nd Annual Meeting*, pp80-87.
- [22] Yu, Q., Max, A., Yvon, F., (2012) "Revisiting Sentence Alignment Algorithms for Alignment Visualization and Evaluation", *LREC Workshop*, pp10-16.

INTENTIONAL BLANK

INTRA-CLUSTER ROUTING WITH BACK-UP PATH IN SENSOR NETWORKS

Turki Abdullah¹, Hyeoncheol Zin¹, Mary Wu², and ChongGun Kim^{1*}

¹Department of Computer Engineering, Yeungnam University, Korea
prince.turki.1988@gmail.com , hczin@naver.com, cgkim@yu.ac.kr

²Yongnam Theological University and Seminary, Korea
mrwu@ynu.ac.kr

ABSTRACT

The novel applications of sensor networks impose some requirements in wireless sensor network design. With the energy efficiency and lifetime awareness, the throughput and network delay also required to support emerging applications of sensor networks. In this paper, we propose throughput and network delay aware intra-cluster routing protocol. We introduce the back-up links in the intra-cluster communication path. The link throughput, communication delay, packet loss ratio, interference, residual energy and node distance are the considered factors in finding efficient path of data communication among the sensor nodes within the cluster. The simulation result shows the higher throughput and lower average packet delay rate for the proposed routing protocol than the existing benchmarks. The proposed routing protocol also shows energy efficiency and lifetime awareness with better connectivity rate.

KEYWORDS

Intra-cluster routing; payoff function; wireless sensor network; throughput and energy-aware; back-up path.

1. INTRODUCTION

Wireless sensor network (WSN) is the connection among the tiny mobile or stationary sensor nodes so that they can share data wirelessly [6][18-19]. The use of sensor networks growing rapidly from healthcare to ocean bed monitoring, and from smart home to space shuttle. The innovative applications of sensor networks impose novel challenges on its protocol design. The energy efficiency, lifetime awareness, energy balancing, network throughput and delay minimization, antenna design and sensor miniaturization are the existing research challenges of sensor network design [9].

Clustering is a proven technique to ensure energy efficient communication. Most of the cluster based sensor network research handles the issue of inter-cluster routing and very few of them consider network throughput and network delay of their proposal. Currently, huge network traffic is generated by the sensors and devices of internet of things (IoT) [7][20], smart home and smart grid networks [10-13]. The throughput and delay must be considered for efficient management and control of this type of sensor enabled network.

In this paper, the intra-cluster routing protocol is proposed to ensure reliable data transmission through introducing back-up link in the communication path. Intra-cluster routing is the process

to find out the efficient path to forward network traffic (or data) towards the cluster head within the cluster. We consider the link throughput, delay, packet loss ratio, interference, residual energy [14-17] and node distance to select the links and corresponding sensor node to establish path of the intra-cluster communication. We also use the penalty functions, which helps the sensor nodes not to select the path with lower than the required or expected throughput and delay. And thus in contrast of conventional cost model, we use the payoff function in determining intra-cluster communication path as Midha Surabhi et al. [8] used the payoff function in their game theoretic model.

2. LITERATURE REVIEW AND RELATED WORKS

Routing is the process of selecting best paths in a network. Routing strategy may design without constructing or considering any clusters. Most of the routing algorithms in wired network like Dijkstra's single source shortest path algorithm were developed without considering any clusters. Some of the routing policy like [21] developed a routing algorithm following no-clustering strategy. Intra-cluster routing is the process to find out the efficient path to forward data towards the cluster head within the cluster. Reference [5] proposed a method of intra-cluster routing.

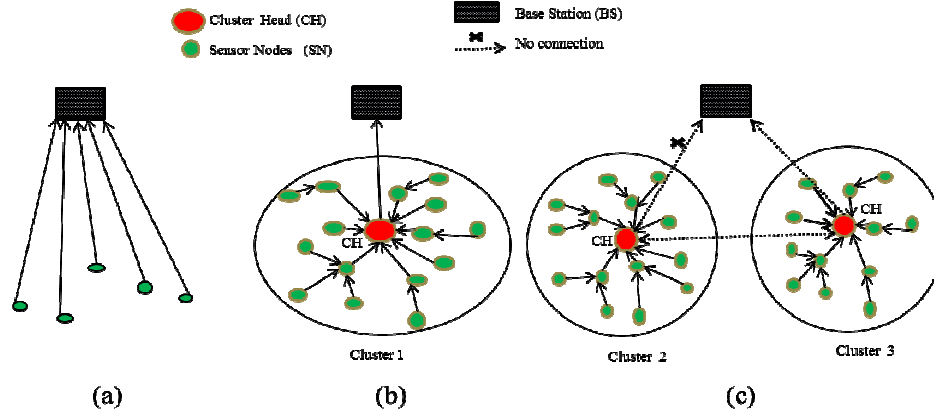


Figure 1: Different routing strategies (a) No Clustering (b) Intra-cluster routing (c) Inter-cluster routing

Inter-cluster routing is the process to find out the efficient path to forward data towards the base station (BS) following cluster to cluster communication. So, inter-cluster routing protocol deals with efficient communication among the clusters. Reference [22] proposed a method of inter-cluster routing. Different routing strategies are shown in Figure 1.

Clustering can be formed using different levels of sensor nodes in perspective of cluster head. If the entire sensor node transmits data directly to the cluster head then this type of clustering is called 1-level clustering. If the sensor nodes of a cluster transmits data to the cluster head through relay of maximum two hop then this type of clustering is called 2-level clustering. Finally, If the sensor nodes of a cluster transmits data to the cluster head through relay of more than two hop then this type of clustering is called N-level clustering. Different levels of clustering in a cluster are shown in Figure 2. In this paper, we consider 2-level clustering for our proposal.

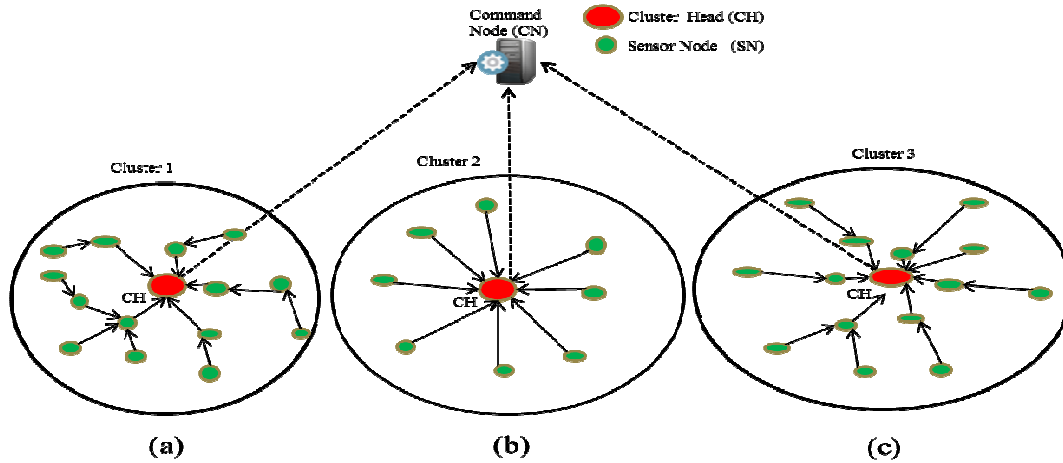


Figure 2: Different level of clustering (a) N- Level Clustering (b) 1-Level Clustering (c) 2-Level Clustering

Some of the protocols are based on hierarchical network structure and some are based on flat network structure, some are QoS based and some are negotiation based. However, the energy efficiency, lifetime awareness, security and network throughputs are the main design goals of WSN routing protocols. A multiple alternating path based on-demand routing protocol is proposed for ad hoc networks, which is an extension of AODV routing protocol [1]. The authors' explored multiple path of data transmission for higher throughput with energy efficiency. This proposal is not based on any clustering mechanism and thus hard to manage the sensor networks in densely deployed environment.

The energy efficient and dual-path routing protocol is proposed in [2], to handle the temporary failure of sensor node after cluster head died in cluster-based WSN. Here, the authors mainly focus on inter cluster routing and try to handle the exception of CH failing. Cluster-based multi-path routing for multi-hop wireless networks is proposed in [3]. The authors' also focused on inter-cluster routing with multiple paths to enhance network throughput. It allows only one path to go over a cluster to reduce interference among two parallel paths. Maintaining routing table for multiple paths causes' higher energy consumption in resource constrained sensor networks. Whereas in our proposed routing protocol we introduce backup links to enhance network throughput in an intra-cluster communication environment.

A proactive routing protocol i.e. Energy Adaptive Clustering Hierarchy (LEACH) is proposed in [4], where a cluster is formed according to the communication range and cluster head is selected by the base station (BS) in a uniform random distributed manner. It is the seminal paper, which shows the way to develop energy efficient WSN through clustering. It is proactive routing protocol, where the node in the network periodically sends data towards CH by following predefined schedule. Easiness of cluster formation and sensor network configuration is one of the best criteria of LEACH protocol. Multi-path or back-up path is not considered in LEACH protocol and thus lower network throughput, higher packet loss and retransmission rate diminish the gain of energy efficiency.

The energy-aware routing in cluster-based sensor networks (EARCBSN) is proposed in [5]. In the proposed method, the authors' assigns a centralized network manager or gateway node to manage intra-cluster communication efficiently. The gateway node sets the route mainly based on energy usage, distance, propagation delay, queuing cost and maximum connections per relay. The authors' show the higher throughput and energy efficiency of their proposed routing scheme

through simulation. In contrast, we proposed a reactive intra-cluster routing protocol based on payoff function with the back-up links for energy efficiency, lifetime awareness and higher network throughput.

3. PROPOSED METHODS

The system model of intra-cluster routing protocol is presented in Figure 3, where we consider non-hierarchical or flat routing topology. The command node (CN) is the control unit of wireless sensor network management. As the CN is a high powered node with permanent electricity supply and having the back-haul communication link with the core network, the CN is responsible for routing table construction and distribution to each of cluster, and also propagate the collected sensor information towards the high end control center. The sensor nodes are deployed in random pattern to collect environmental data. The clusters of sensors nodes are dynamic in nature having a cluster head (CH), which is responsible for gathering environmental data from sensor nodes (SN) within its territory and then transfer the collected data towards the CN directly. The cluster heads are also dynamically changed in each and every round of data transmission. The cluster formation, route determination and data accumulation phases of our proposed intra-cluster routing protocol are discussed in following subsections.

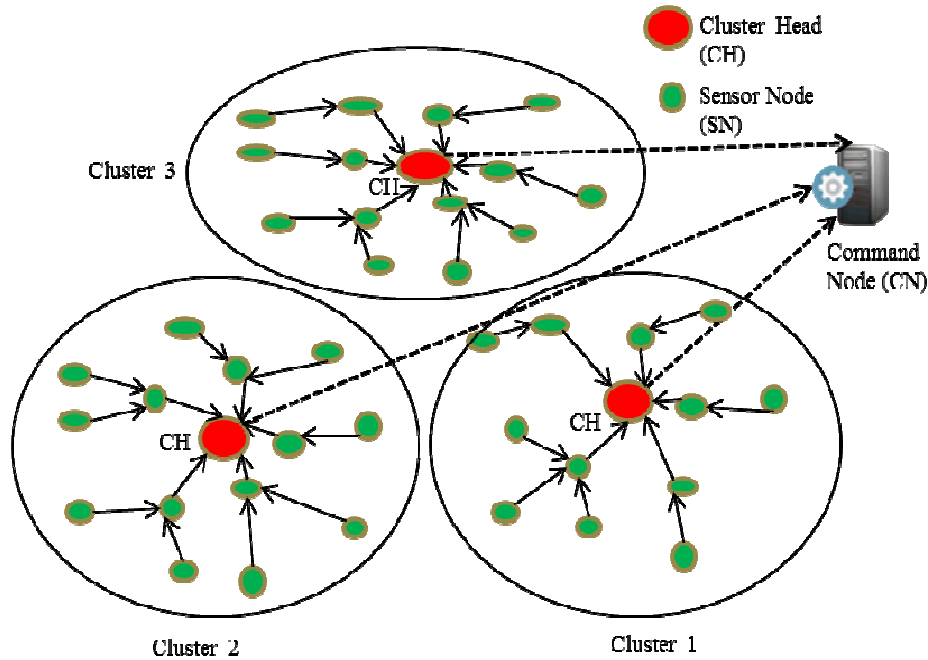


Figure 3. System model of intra-cluster routing protocol.

3.1 Cluster Formation

The cluster of sensor nodes is formed on the basis of nodes within the radio range. Some of the sensor nodes of a cluster can directly send data to the cluster head and some other node uses the relay node to send data to cluster head through multi-hop communication. For efficient communication purpose, the command node (CN) selects the cluster heads (CH) for each and every round of data transmission. The CN selects CH following uniform distribution so that each and every node becomes cluster head by turn. After receiving commands from CN, the CH

broadcast his headship status to the sensor network and expecting membership request from normal sensor nodes (SNs). The SNs send the membership request with their positional and energy level information to the nearest CH. After receiving membership request with necessary information the CH determines the throughput, delay, SINR, packet loss ratio of the communication link between each SN and CH. The CH then sends that information to the CN for efficient routing table construction as we discussed in section 3.2.

The CN node then sends the routing table to CH. Finally, CH sends the routing information i.e. next hop node; back-up next-hop node and acting node status to each member SNs. CH also sends the soft and hard threshold to each of its member SNs for energy efficient data transmission. The cluster formation procedure is presented in algorithm 1 as intra-cluster routing procedure.

Algorithm 1: Intra-Cluster_Routing ()

1. Command node (CN) selects the cluster heads (CH) randomly to form clusters.
 2. CH broadcast advertisement to the sensor network.
 3. Sensor nodes (SN) transmit membership request to CH node with their position in cluster and energy level.
 4. CH sends all sensor nodes position, energy level, throughput, delay, SINR, packet loss ratio to CN.
 5. CN construct the routing table according to Algorithm 2 and sends the routing table to CH.
 6. CH sends following information to all member nodes: the next-hop node, back-up next-hop node and state of the node with soft and hard threshold values.
 7. Member sensor nodes act according to the defined state and sends data to CH based on the routing table.
 8. CH compress the data and sends to the command
-

3.2 Routing Table Construction

The CN is responsible for efficient routing table construction based on the metrics supplied by the CH regarding the member SNs of its cluster. To construct the route of intra-cluster data dissemination, the CN uses the Greedy method to find out the best hop-by-hop data dissemination path by determining the best next-hop node of each of the member SNs. However, CN also finds the alternative next-hop node for reliable data transmission with higher throughput and lower packet loss. Among the adjacent nodes of any SN the best next-hop node and next-hop alternative nodes are determined according to the link suitability value of equation (1). We presented the procedure of routing table construction in algorithm 2.

Algorithm 2: Routing_Table_Construction ()

1. Find the list of sensor nodes $\{d_1, d_2, \dots, d_m\}$, which are within 1-hop communication range from cluster head
2. For each sensor node N of the cluster C
3. Find the neighbouring sensor nodes $\{s_1, s_2, \dots, s_n\}$ of node N
4. Determine the feasible set of relay nodes by finding common nodes between the list of sensor nodes within 1-hop communication range from cluster head and the neighbouring sensor nodes of node N i.e. $\{a_1, a_2, \dots, a_k\} = \{d_1, d_2, \dots, d_m\} \cap \{s_1, s_2, \dots, s_n\}$
5. Determine the link suitability values $\{L_{N,a1}, L_{N,a2}, \dots, L_{N,ak}\}$ from node N to each of the neighbouring sensor nodes $\{a_1, a_2, \dots, a_k\}$ using equation (1).
6. Find the sensor node a_i with maximum link suitability values $L_{max1} = \max\{L_{N,a1}, L_{N,a2}, \dots, L_{N,ak}\}$ and set the node a_i as the next hop node of N to construct the routing table for node N .

7. Find the sensor node a_j with maximum (i.e. second highest) link values $L_{max2} = \max\{ \{L_{N,a1}, L_{N,a2}, \dots, L_{N,ak}\} \setminus \{L_{N,ai}\} \}$ and set the node a_j as the back-up next hop node of N to construct the routing table for node N.
8. Define the state of the node N from the set of states' $S = \{\text{sensing, aggregating, active, relaying, inactive}\}$ sequentially following random distribution to construct the routing table for node N.

The link suitability value $L_{i,j}$ or payoff function is determined through equation (1), where i is any SN and j 's are the adjacent sensor nodes of node i .

$$L_{i,j} = \alpha_1 * \frac{T_{obs}}{T_{req}} + \alpha_2 * \frac{R_{avg}}{R_{obs}} + \alpha_3 * \frac{PL_{tol}}{PL_{obs}} + \alpha_4 * \frac{SINR_{obs}}{SINR_{std}} + \alpha_5 * \frac{E_{res}}{E_{ini}} + \alpha_6 * \frac{D_{nr}}{D_{rg}} - T_{pnit} - R_{pnit} - PL_{pnit} - SINR_{pnit} - E_{pnit} - D_{pnit} - S_{pnit} \quad (1)$$

Here, the link suitability value is determined through the payoff function, where not only the benefit factors but also the penalty or cost factors are considered. The considered factor are the distance between source and relay node, energy level of the node, throughput, delay, SINR and packet loss ratio of communication link. The penalty functions are formulated as in equation (2) through (7).

$$T_{pnit} = \begin{cases} \frac{T_{req}}{T_{obs}} * \beta_1 ; & \text{if } T_{obs} < T_{req} \\ 0 & ; \text{Otherwise} \end{cases} \quad (2)$$

$$R_{pnit} = \begin{cases} \frac{R_{obs}}{R_{avg}} * \beta_2 ; & \text{if } R_{obs} > R_{avg} \\ 0 & ; \text{Otherwise} \end{cases} \quad (3)$$

$$PL_{pnit} = \begin{cases} \frac{PL_{obs}}{PL_{tol}} * \beta_3 ; & \text{if } PL_{obs} > PL_{tol} \\ 0 & ; \text{Otherwise} \end{cases} \quad (4)$$

$$SINR_{pnit} = \begin{cases} \frac{SINR_{std}}{SINR_{obs}} * \beta_4 ; & \text{if } SINR_{obs} < SINR_{std} \\ 0 & ; \text{Otherwise} \end{cases} \quad (5)$$

$$E_{pnit} = \begin{cases} \frac{E_{ini}}{E_{res}} * \beta_5 ; & \text{if } \frac{E_{res}}{E_{ini}} < E_{thrs} \\ 0 & ; \text{Otherwise} \end{cases} \quad (6)$$

$$D_{pnit} = \begin{cases} \frac{D_{rg}}{D_{nr}} * \beta_6 ; & \text{if } D_{nr} \gg T_{avg} \\ 0 & ; \text{Otherwise} \end{cases} \quad (7)$$

Table 1. Used symbols and their description

<i>Symbol</i>	<i>Description</i>
$L_{i,j}$	Link's suitability value of link between node i and j
T_{obs}	Observed Throughput
T_{req}	Required Throughput
T_{pnlt}	Penalty for not fulfilling the Throughput requirement
R_{obs}	Observed Response time
R_{avg}	Average response time (standard)
R_{pnlt}	Penalty for not fulfilling the Response time requirement
PL_{obs}	Observed Packet Loss ratio
PL_{avg}	Tolerable Packet Loss ratio
PL_{pnlt}	Penalty for higher packet loss than the tolerable range
$SINR_{obs}$	Observed Signal_to_Interference Noise Ratio
$SINR_{std}$	Standard Signal_to_Interference Noise Ratio
$SINR_{pnlt}$	Penalty for higher Noise in signals
E_{ini}	Initial Energy of a sensor node j
E_{res}	Residual energy of the sensor node j
E_{pnlt}	Penalty of selection of a node which has lower residual energy than a threshold energy (E_{thrs})
D_{nr}	Distance from node to relay
D_{rg}	Distance from relay to gateway
D_{pnlt}	Penalty for selecting a node which is very far from the node i ; that is, distance is more than the average transmission range (Tr_{avg})
S_{pnlt}	State switching (changing) cost
$\alpha_1, \alpha_2, \dots, \alpha_6$	Weighting factors of link's gain
$\beta_1, \beta_2, \dots, \beta_6$	Weighting factors of link's penalty

The symbol used in equation (1) through (9) is explained in table 1. The summations of weighting factors are considered as 1, which formulated as in (8) and (9).

$$\alpha_1 + \alpha_2 + \dots + \alpha_6 = 1 \quad (8)$$

$$\beta_1 + \beta_2 + \dots + \beta_6 = 1 \quad (9)$$

For example, to construct the routing table for node D, the CN node calculate the link suitability values $\{L_{D,C}; L_{D,A}; L_{D,B}; L_{D,E}\}$ from node D to its 1 hop neighbouring node $\{C, A,B,E\}$. According to equation (1) through (9) and the measured values of table 1, we found the link suitability values of links as $\{L_{D,C}=0.374; L_{D,A}=0.452; L_{D,B}=0.701; L_{D,E}=-0.214\}$. Among the link suitability values the highest and second highest values are $L_{D,B}=0.701$ and $L_{D,A}=0.452$. So, we set node B as the next hop node for D and we set node A as the next hop alternative node of node D, as shown in figure 4. Using algorithm 2, we can construct the full routing table as shown in figure 4. Link L_{DC} cannot determine as a back-up path because the SN_C have already 2-level path.

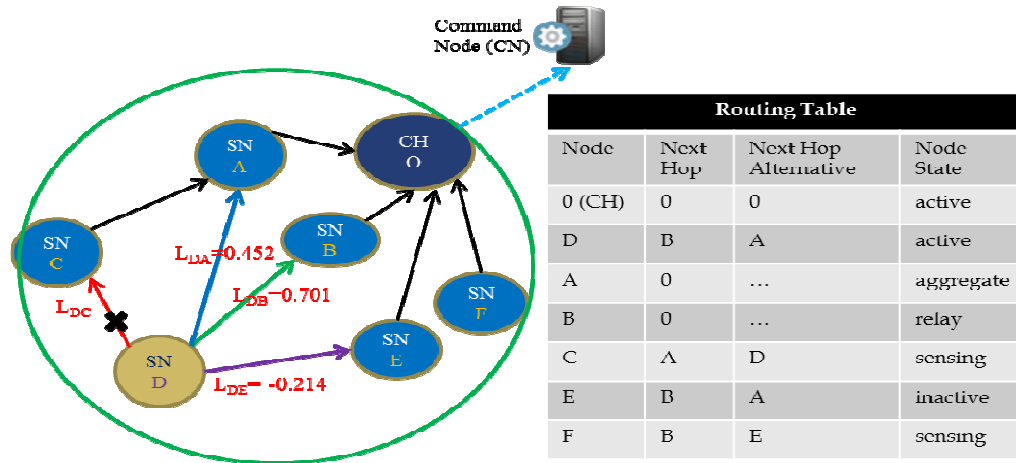


Figure 4. Routing table construction of an example cluster.

Defining states of different node is a practice in sensor networks for energy efficient communication. We consider five different states of sensor nodes for ensuring energy efficiency and the states are sensing, aggregating, active, relaying, and inactive state. Firstly, we define the states of sensor nodes following uniform distribution. Secondly, we maintain the sequential pattern of changing states as sensing \rightarrow aggregating \rightarrow active \rightarrow relaying \rightarrow inactive state. The CH remains active throughout the data communication in a single round. In sensing state, the sensing circuitry of a node remains on and it temporarily stream the sensed data to its buffer. In aggregating state, the sensing and relaying circuits of the node are off and it compares its streamed data with hard and soft threshold, compresses the data and then sends those data towards the gateway in next round when it becomes active. In relaying state, only the communication circuitry remains on to relay the data from other active nodes. In active state, the sensor node can transmit data; it also can sense, aggregate and relay data. Thus the cluster head must be an active node. In inactive state, the node turns off its sensing and communication circuitry and it again becomes alive after predefined waiting time.

4. SIMULATION

4.1 EXPERIMENTAL CONFIGURATION

The performance of the proposed intra-cluster routing approach studied through simulation using a calculation tool. We studied the energy dissipation, lifetime awareness, throughput, average packet delay and connectivity rate to validate our proposed routing algorithm. We also compare our results with the benchmark routing protocol LEACH. However, we studied the performance of EARCBSN and compare the achieved results with our proposed method for the justification of improved performance. The LEACH is just 1-level routing and the EARCBSN is 2-level routing without back-up path.

The simulation scenario is presented in figure 5, where the blue cross mark represents the command nodes position, the black circles are represented as sensor nodes (SNs) and the red circles are represented as the cluster heads (CHs). There are total 200 nodes deployed in 100x100 square meters of area. Total 11% of the sensor nodes are selected as the CH for each round. The simulation parameters and their assumed values are presented in table 2.

Table 2. Simulation parameters and values for performance study

Simulation Parameters	Symbols	Values
Topology	---	2D and flat
Number of nodes	N	200
Simulation Area	W x H	100 x100 square meters
Packet size	b	512 bits
Total number of rounds	R	4500
Transmitter circuitry energy per bit	ETx_circuit	50 nJ/bit
Transmitter amplification energy per bit per square meter	ETx_amplifier	100 pJ/bit/m ²
Receiver circuitry energy per bit	ERx_circuit	50 nJ/bit
Sensing energy per bit	ERx_sensing	50 nJ/bit
Initial energy of each node	E _o	1 joule
Aggregating energy per bit	E _p _aggregation	4.3x10 ⁻³ nanojoules/bit
Channel bandwidth	B	5 Mbps
Round equivalent time in millisecond	T	20 ms

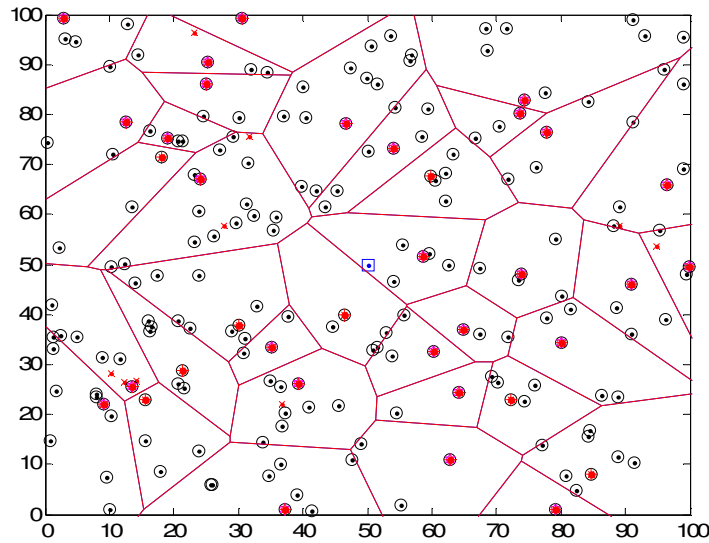


Figure 5: Simulation environment, considering 200 nodes in 100x100 meters

4.2 EXPERIMENTAL RESULTS

Energy efficiency is the first consideration in wireless sensor networks protocol design. The cumulative energy consumption rate in different rounds of data transmission is determined

through simulation study. Figure 6 shows that the energy consumption of LEACH protocol is highest because of the hierarchical and proactive nature of LEACH and also all nodes remain alive in all time. The energy consumption of EARCBSN is also higher than the proposed intra-cluster routing protocol because additional retransmission of packets in case of link failure. The deployment of back-up link and introduction of aggregating state turns the proposed routing method as energy efficient than the existing benchmarks.

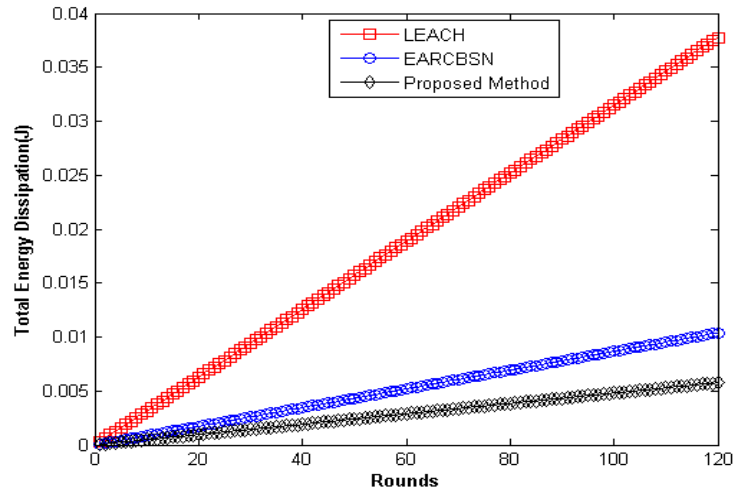


Figure 6: Cumulative Energy Consumptions of Different Routing Protocols

Lifetime awareness is another important metric to measure the performance of wireless sensor network protocol. The protocol with greater lifetime can transmit data in longer time. Figure 7 and 8 shows the lifetime awareness of the studied routing protocol. Figure 7 shows that, in case of LEACH and EARCBSN first node dies at 935th and 1141st round respectively, on the other hand, in case of our proposed approach the first node dies at 1283th round. Earlier collapsing of node makes the network paralyzed.

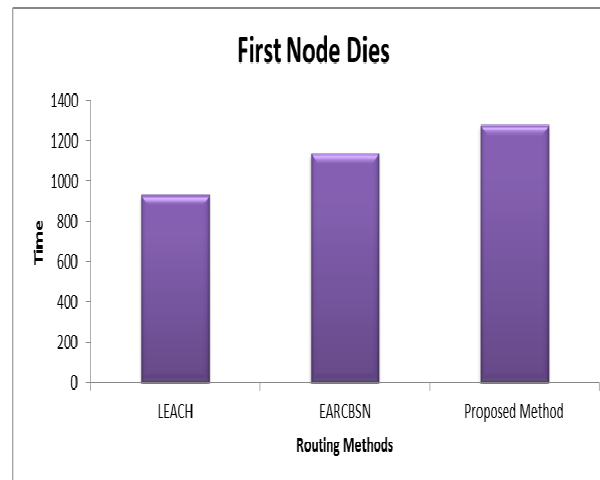


Figure 7: Network partitioning in different routing protocols

The quicker collapsing of nodes guides the sensor networks in an unstable state. Figure 8 shows that, in case of our proposed intra-cluster routing protocol, the network remains alive up to 3956th rounds whereas the LEACH and EARCBSN remains active up to 1308th and 2377nd rounds respectively. The controlled reactive nature of our proposed protocol helps the sensor networks to remain alive in longer time. The balancing of energy consumption of different nodes is controlled by assigning states of nodes in uniform manner. The use of hard and soft threshold also plays vital role in lifetime awareness of our proposed method.

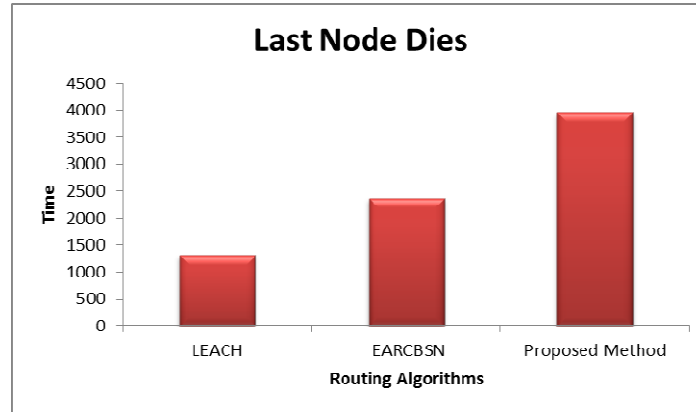


Figure 8. Lifetime of different benchmarked routing protocols

Figure 9 shows the throughput of different routing protocols in different rounds. We analysed the throughput considering the link capacity of 5 Mbps. The hierarchical structure of LEACH hinders the throughputs of LEACH protocol. The higher packet loss and retransmission issue hinders the throughput of EARCBSN. The flat routing topology and the backup links of the proposed routing method increase the network throughput.

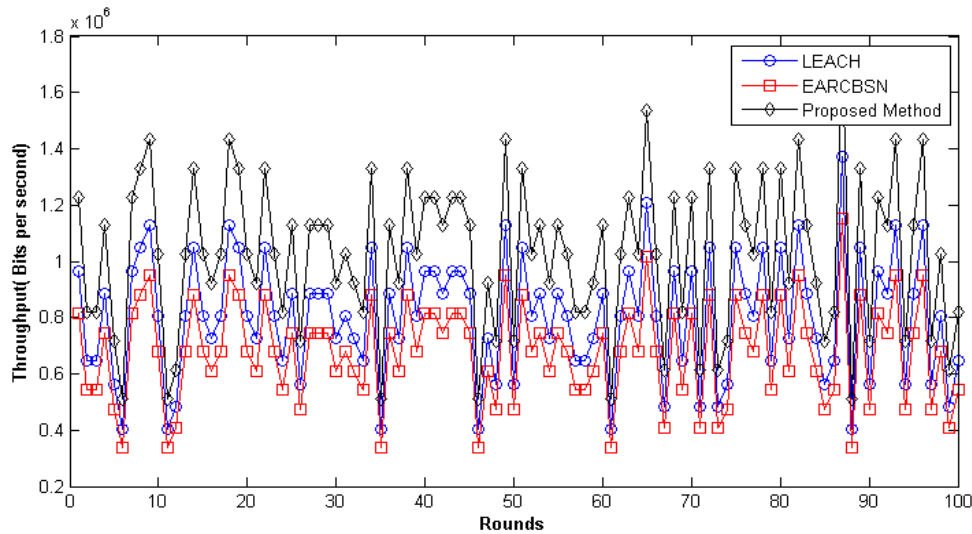


Figure 9. Communication throughput of different routing algorithms

The average delay of per packet data transmission is studied and presented in figure 10. The LEACH used hierarchical clustering architecture (i.e. cluster head level 1, cluster head level 2

etc.), whereas we use flat clustering architecture (i.e. each cluster head is directly connected to command node). For that reason the sensor node can send data directly to command node with lower delay. In the proposed method, we have back-up path to transmit data, so less packet drops are happening in this case, whereas there is no back-up path in LEACH and EARCBSN, so more packet drops are happening, as a result more retransmission is required in case of LEACH and EARCBSN and thus proposed method experiences lower average packet delay than the existing LEACH and EARCBSN.

We also studied the connectivity rate of different routing protocols in figure 11. The connectivity rate is the ratio of number of connected nodes with CH and total number of living nodes on the network. Dis-connectivity may occur due to interference, hidden nodes, signal obstacles and inactive nodes and out of transmission range. As the proposed routing protocol allows some node to be in inactive state, the connectivity rate goes down up to 68.75%. The lowest connectivity rate of LEACH and EARCBSN is higher than the proposed method, but the networks of LEACH and EARCBSN goes down rapidly due to higher energy dissipation. With the cost of connectivity rate the proposed intra-cluster routing algorithm gain energy efficiency and longer lifetime.

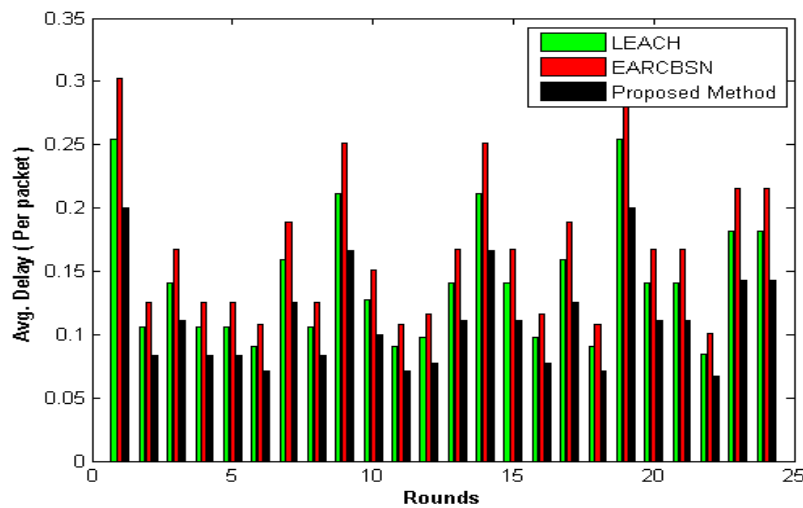


Figure 10. Average delay of different routing protocols

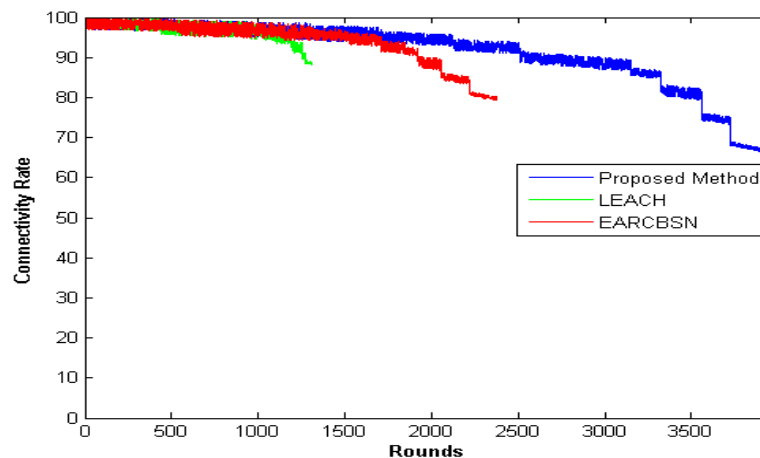


Figure 11. Connectivity rate different routing protocols in different rounds of data transmission

5. CONCLUSIONS

The intra-cluster routing protocol with back-up path is proposed in this research. The simulation results show the energy efficiency and longer lifetime of sensor networks. Although the proposed routing protocol shows lower average connectivity rate, but the back-up path, payoff function and different states of sensor node helps to deliver packets with higher throughput and lower rate of average packet transmission delay. In this proposal, we introduce a new method of link value determination, based on the maximum link value we select links for determining the next hop node of a source node and also determine the next hop alternative to enhance the reliability of sensor networks data communication. We studied energy dissipation, network lifetime, throughput and average delay and compare those with existing EARCBSN method. We found that the proposed method outperforms over the EARCBSN method. We will apply some machine learning and game theoretic approach to design the payoff function, which may enhance the performance of our proposed routing approach.

ACKNOWLEDGEMENTS

This work has been funded by the BK21+ program of the National Research Foundation (NRF) of Korea.

REFERENCES

- [1] Lee, Sung-Ju, and Mario Gerla. "AODV-BR: Backup routing in ad hoc networks." *Wireless Communications and Networking Conference, 2000. WCNC. 2000 IEEE. Vol. 3. IEEE, 2000.*
- [2] Ding, Ding, Liu Fangai, Li Qianqian, and Yang Guangxu. "An Improved Clustering Algorithm Based on Backup Path." *Advances in Information Sciences & Service Sciences* 4, no. 8 (2012).
- [3] Zhang, Jie, Choong Kyo Jeong, Goo Yeon Lee, and Hwa Jong Kim. "Cluster-based multi-path routing algorithm for multi-hop wireless network." *Future Generation Communication and Networking* 1 (2007): 67-75.
- [4] Heinzelman, Wendi Rabiner, Anantha Chandrakasan, and Hari Balakrishnan. "Energy-efficient communication protocol for wireless microsensor networks (LEACH) " *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on. IEEE, 2000.*
- [5] Younis, Mohamed, Moustafa Youssef, and Khaled Arisha. "Energy-aware routing in cluster-based sensor networks." *Modeling, Analysis and Simulation of Computer and Telecommunications Systems, 2002. MASCOTS 2002. Proceedings. 10th IEEE International Symposium on. IEEE, 2002.*
- [6] Pantazis, Nikolaos A., Stefanos A. Nikolidakis, and Dimitrios D. Vergados. "Energy-efficient routing protocols in wireless sensor networks: A survey." *Communications Surveys & Tutorials, IEEE* 15.2 (2013): 551-591.
- [7] Perera, Charith, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. "Context aware computing for the internet of things: A survey." *Communications Surveys & Tutorials, IEEE* 16, no. 1 (2014): 414-454.
- [8] Midha, Surabhi, Ajay K. Sharma, and Geeta Sikka. "A survey on wireless sensor network clustering protocols optimized via game theory." *ACM SIGBED Review* 11.3 (2014): 8-18.
- [9] Alghanmi Ali, and ChongGun Kim. "Energy efficient load balanced routing protocol for wireless sensor networks." *Computer Science* (2014).
- [10] Alam, M. G. R., Cho, E. J., Huh, E. N., & Hong, C. S. (2014, January). Cloud based mental state monitoring system for suicide risk reconnaissance using wearable bio-sensors. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication* (p. 56). ACM.
- [11] Mary Wu, YoungSeok Jung, Chonggun Kim, "The effects of central leader and candidates in Ad Hoc Networks", *Information-An International Interdisciplinary Journal*, Vol. 14, pp. 3601-3609, 2011.

- [12] Mary Wu, InTaek Leem, Jason J. Jung and ChongGun Kim, "A Resource Reuse Method in Cluster Sensor Networks in Ad Hoc Networks," Intelligent Information and Database Systems, Lecture Notes in Computer Science, Volume 7197/2012, 40-50, 2012.
- [13] Ahmed, Mohammad Helal Uddin, Alam Md Golam Rabiul, Kamal Rossi, Hong Choong Seon, and Sungwon Lee. "Smart grid cooperative communication with smart relay." *Journal of Communications and Networks* 14, no. 6 (2012): 640-652.
- [14] Mary Wu, Chonggun Kim, "A cost matrix agent for shortest path routing in ad hoc networks," *Journal of Network and Computer Applications*, 33, 646-652, 2010.
- [15] Mary Wu, SeongGwon Cheon, Chonggun Kim, "A Central Leader Election Method using the Distance Matrix in Ad Hoc Networks", *New Challenges for Intelligent Information and Database Systems*, Vol. 351, pp. 107-116, 2011.
- [16] Alam, Md, Golam Rabiul, Chayan Biswas, Naushin Nower, and Mohammed Shafiul Alam Khan. "A Reliable Semi-Distributed Load Balancing Architecture of Heterogeneous Wireless etworks." *arXiv preprint arXiv:1202.1918* (2012).
- [17] Mary Wu, Byungchul Ahn, ChongGun Kim, "A Channel Reuse Procedure in Clustering Sensor Networks, " *Applied Mechanics and Materials* V.284-287 pp.1981-1985.
- [18] Jaime Lloret, "Underwater Sensor Nodes and Networks", *Sensors* 2013, 13(9), 11782-11796
- [19] Faezeh Arab Hassani, Yoshishige Tsuchiya and Hiroshi Mizuta, "In-Plane Resonant Nano-Electro-Mechanical Sensors: A Comprehensive Study on Design, Fabrication and Characterization Challenges", *Sensors* 2013, 13(7), 9364-9387.
- [20] Fang, Shifeng, Li Da Xu, Yunqiang Zhu, Jiaerheng Ahati, Huan Pei, Jianwu Yan, and Zhihui Liu. "An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things." *IEEE Trans. Industrial Informatics* 10, no. 2 (2014): 1596-1605.
- [21] Chang,, L. Tassiulas, "Energy Conserving Routing in Wireless Ad-Hoc Networks," In *Proc. International Conference on Computer Communications*, Tel-Aviv, Israel, 2000, pp. 22-31.
- [22] Zhang, Jie, Choong Kyo Jeong, Goo Yeon Lee, and Hwa Jong Kim. "Cluster-based multi-path routing algorithm for multi-hop wireless network." *Future Generation Communication and Networking* 1 (2007): 67-75.

RECOGNIZING NAMED ENTITIES IN TURKISH TWEETS

Beyza Eken and A. Cüneyd Tantug

Department of Computer Engineering,
İstanbul Technical University, İstanbul, Turkey

¹beyzaeken@itu.edu.tr

²tantug@itu.edu.tr

ABSTRACT

Named entity recognition (NER) is one of the well-studied sub-branch of natural language processing (NLP). State of the art NER systems give highly accurate results in domain of formal texts. With the expansion of microblog sites and social media, this informal text domain has become a new trend in NLP studies. Recent works has shown, social media texts are hard to process and the performance of the current systems substantially decrease when switched to this domain. We give our experience in improving named entity recognition on informal social media texts for the case of tweets.

KEYWORDS

Named Entity Recognition, Conditional Random Fields, Informal Domain, Tweet, Turkish

1. INTRODUCTION

Named entity recognition (NER) is a natural language processing (NLP) term that refers to the recognition of named entities in natural language. It is a way of extracting information by detecting and classifying named entities in texts. Most studied named entity types are person, location, organization which defined in MUC-6 [1] conference as ENAMEX type. Other mostly studied types are numeric entities like money, percentage as NUMEX type and date, time as TIMEX.

NER could take part in other NLP tasks like machine translation, sentiment analysis, and question-answering.

There have been a lot of studies in NER field in many languages and the state of the art performance has reached to nearly human annotation performance on formal texts [2]. But texts are not always formal like e-mails, microblog texts, social media texts, etc. But off the shelf NLP tools give low accuracy when they are applied to informal texts [3], because they may be ungrammatical and can have spelling mistakes unlike formal texts. As a consequence, the need arises to develop new methods which would work properly for informal structure of texts.

With the expansion of the web, information gathering and sharing via social media has become a rising trend. Twitter is one of the most used microblog site around the world, 500 million tweets are sent per day [4]. Tweets hold great amounts of statistics, they can give important information about a company, person etc. Therefore, at this point NER on tweet domain is holds crucial importance.

The aim of this study is to increase the performance of NER in Turkish tweets. Tweets are short texts that have maximum 140 characters, and most of the time they can contain grammar or spelling mistakes, slang words, smileys and so on. Unfortunately these irregular nature of tweets make it harder to process such data.

Turkish is a highly agglutinative language and it makes Turkish language morphologically rich. Morphological features hold meaningful importance in NLP tasks, they have important information about words. But in informal texts off the shelf morphological analysers do not give sufficient results. So, instead of morphological analysing process we prefer to use first and last four character of word in order to take advantage of morphological features. According to our results when first four characters of the word are used as an alternative to stem of the word, performance changes slightly.

Previous works [5]-[8] have shown that conditional random fields (CRF) method has reached to good performance at NER task. Consequently, in this work CRF have been chosen as the method to build named entity recognition model.

The rest of the paper follows with related works, then describes the method we used, after that gives and explains our results and lastly final section as conclusions.

2. RELATED WORKS

Named entity recognition is a well-studied field in many languages especially in English. First studies started in 1990s [2], now state of the art performance has reached nearly %95.

First NER study specific to Turkish used hidden markov models (HMM) on news data and reached %91.56 performance with person, location, organization types [9]. Bayraktar and Temizel [10] used patterns and word frequency to recognize Turkish person names on financial text domain. Küçük and Yazıcı [11] proposed a rule based system to recognize ENAMEX, TIMEX and NUMEX types, then they improved their system with rote learning algorithm and achieved %90.13 performance on Turkish news data [12]. Tatar and Çiçekli [13] created an automatic rule learning system and they achieved %91.08 performance on Turkish news data. Yeniterzi [6] got %88.94 performance on Turkish news data with CRF using morphological features. Şeker and Eryiğit [7] achieved %92 performance on Turkish news data with CRF using morphological features and gazetteers.

When examining NER for informal domain Özkaya and Diri [8] reached %92.89 performance with ENAMEX types on Turkish e-mails, e-mail domain kind of informal. Çelikkaya et al. [14] normalized tweets and tested on a model trained with Turkish news data and achieved %19 performance, they used CRF with morphological features and gazetteers. Küçük and Steinberger [15] adapted Küçük's rule based NER system [11] to tweet domain and got %61 performance.

Ritter et al. [3] tailored NLP pipeline to tweet domain, and get %51 score on English tweets, they used CRF in part-of-speech tagging, chunking, named entity segmentation parts and they used LabeledLDA [16]. Liu et al. [17] created a semi supervised system using k-nearest neighbors algorithm and CRF, they achieved %80.2 performance on English tweets. Li et al. [18] created an unsupervised system for only segmentation of named entities in English tweets using Wikipedia and Web N-gram corpus. Oliveria et al. [19] created a filter based system for English tweets.

3. DATASETS AND METHOD

We aimed to develop a model that will recognize person, location, organization, date, time, money and percentage named entities in Turkish tweets. Tweets are short texts which can be solecistic. Lack of context, containing spelling errors on purpose or not, slangs, repeating characters to indicate exclamation make hard NER process on tweets. Two root ideas for NER in domain like tweets are to tailor texts to existing NER tools or tailor existing NER tools to fit informal texts.

CRF have been used to build our NER model. CRF are introduced by Lafferty et al. CRF are statistical machine learning techniques which aim is to be applied to sequential data to segment and label. We used CRF++ tool [20] for training and testing system.

We used news data to train a base model just to see our results on news data to make comparison of selected features. Then we used tweets to train a second model which is more feasible for tweet domain.

We used two main data set from two different domain, news data as formal texts and tweets as informal texts. News data set which are collected from Turkish newspapers and labelled by Tür et al. [9]. Tweet data set consisting of two parts, first part is labelled by us for this work and consists of nearly 9K tweets, second part is labelled by Çelikkaya et al. [14] and consists of nearly 5K tweets. Entity counts for all datasets are given in Table 1 and Table 2.

Table 1. News data set entity counts.

	Train	Test	Total
Token	444.475	47.343	491.818
Entity	38.388	3579	36.967
Person	14.492	1598	36.967
Location	10.538	1177	11.715
Organization	8358	804	9162

We divide news data and used %10 for testing and remain for training.

Table 2. Tweets data set entity counts.

	Tweets-1	Tweets-2	Train	Test	Total
Tweet	9.358	5.040	12.471	1.930	14.401
Token	108.743	46.620	137.345	21.300	158.645
Entity	7.838	1.689	5.511	901	6.412
Person	2.744	875	2.099	429	2.528
Location	1.419	277	1.168	172	1.340
Organization	2.935	389	1.733	236	1.969
Date	351	82	261	31	292
Time	86	33	90	21	111
Money	212	29	88	10	98
Percentage	91	4	72	2	74

Tweets-2 column in Table 2 represents entity counts for Çelikkaya et al.'s tweets data set [14], Tweet-1 column represents our tweets data set. We combine two tweets data sets to make balanced training and testing sets, that is to say we take %10 of each tweets data set to comprise final tweets data set, and remaining of each are combined to comprise final training tweets.

We trained our first news model as same way in Şeker and Eryiğit's work [7], we nearly get same results as this work. We apply morphological analyse and disambiguation processes on data after tokenization. Oflazer's tool [21] is used for morphological analyse and Sak's tool [22] for morphological disambiguation. Morphological process is used for to extract stem, inflectional suffixes, part of speech, noun case and proper name case information of tokens, all of these information are used when training the model.

Another encountered writing style for Turkish tweets is that instead of using Turkish characters (ö, ç, ş, ı, ğ, ü) equivalent of English characters (o, c, s, i, g, u) are used. Therefore we asciified all data sets, which means we replaced all Turkish specific characters with equivalent of English characters.

Hence Turkish is an agglutinative language last characters of words are generally suffixes of words so they hold meaningful information about word's morphology. On the other hand, morphological processing tools do not perform well on tweets, so in this second method instead of using morphological features we used first four and last four characters of the tokens as features to train models. This alternative model performs nearly same as first one, so we infer that there is no need to use morphological analysing and disambiguation processes for this work.

Proper name's suffixes should be separated with apostrophe, therefore containing an apostrophe gives important clue about being a named entity, so this is also added as a feature.

Also we applied distance based matching to extract gazetteer features, because of twitter domain peculiarities exact matching can lead to missed out entities. Since tweets contain spelling errors, some named entities can be contracted like "İstnbul" instead of writing correct form of entity which is "İstanbul". Exact matching of input tokens and gazetteers will miss out contracted entities, in order to not miss out these entities we applied distance based matching with Levenshtein distance algorithm [23]. Levenshtein distance algorithm calculate distance between two strings, calculated distance between two strings represents minimum number of edits which are necessary to change one word into the other. For this work we calculate distances between

input token and each token in gazetteer. Zero distances are already named entities, distances closer to zero are candidate named entities. So we give a chance to tokens like “İstanbul” for being a named entity.

4. EVALUATION AND RESULTS

We evaluated our results according to CoNLL metric using CoNLL evaluation script [24], this metric calculates f-measure considering entity type and boundaries. In system output, if both of type and boundaries of a named entity are labelled correctly this entity counted as correct.

We labelled entities in data sets using NER annotation tool from [11] and we represent entities with IOB2 representation style introduced in [25].

Results are on our first model based on this work [7], it trained with news data and named with *N1_model*. We used morphological features, letter case features, start of sentence features and gazetteers to build this model. We tested all our test data on this model and results are in Table 3. We have got nearly same result as in [7] for news test dataset.

Tweets Test Set-1 results in Table 3, 4 and Table 5 are from final tweet test set which is a combination of our tweets and tweets from this [7] work. Tweet Test Set-2 results are represent results of tweets data set from this work [14].

Table 3. Results on first news model (N1_model)

Model	News Test Set	Tweets Test Set-1	Tweets Test Set-2
Surface	82.93	32.30	18.80
Stem	83.36	13.71	14.76
+Surface	84.30	32.43	15.16
+Part of speech	84.85	33.53	15.75
+Noun case	85.59	35.18	17.55
+Proper noun	86.91	21.87	9.58
+Inflections	87.14	22.39	9.79
+Case	90.01	34.66	20.38
+Start of sentence	89.91	34.83	20.27
+Gazetteers	90.38	41.22	25.45
+Distance-based Matching	90.47	41.79	25.62

Second news model named *N2_model* based on some different features instead of morphological features, results in Table 4. Our primary objective is improving tweets data performance for NER but we also trained and tested on news datasets to see and compare results.

Table 4. Results on second news model (N2_model)

Model	News Test Set	Tweets Test Set-1	Tweets Test Set-2
Surface	82.93	32.30	18.80
+First 4 characters	84.07	34.75	19.12
+Last 4 characters	85.34	36.39	20.80
+Apostrophe	86.11	37.43	22.70
+Case	89.41	41.58	24.49
+Start of sentence	89.50	41.57	24.15
+Gazetteers	90.17	46.97	27.90
+Distance-based Matching	90.23	46.57	28.53

When we look at Table 3 and 4 it can be seen we get nearly same results for news test data in both news models. It shows we can capture significant features with second way. Beside that f-measures are improved for tweets on *N2_model*.

Then we trained third model with second way using tweets as training data, which name is *T_model*. We got highest scores for tweets on this model.

Table 5. Results on tweets model (T_model)

Model	Tweets Test Set-1
Surface	49.32
+First 4 characters	56.41
+Last 4 characters	57.00
+Apostrophe	57.98
+Case	61.82
+Start of sentence	61.97
+Gazetteers	64.03
+Distance-based Matching	63.77

5. CONCLUSIONS

We studied on improving performance of NER on Turkish tweets. Although NER is almost a solved problem in formal texts domain, when switch domain to informal texts performance decreases in respectable amount. There are two main way in literature to handle this decrease, tailoring systems to adapt to informal texts or tailoring data to adapt to existing systems. We proposed a NER system for tweets without normalization of tweets.

We improved performance on tweets and get %64 f-measure with some basic features that are first and last 4 characters of the word, capitalization and apostrophe information and gazetteers. We asciified data sets and gazetteers before building our model and apply a little normalization. We employ distance based matching with Levenshtein distance algorithm when extracting gazetteer look up features, we will work on enhance gazetteer look up techniques.

REFERENCES

- [1] R. Grishman & B. Sundheim (1996) "Message Understanding Conference-6: A Brief History", In Proceedings of 16th International Conference on Computational Linguistics, pp. 466-471.
- [2] D. Nadeau & S. Sekine (2007) "A Survey of Named Entity Recognition and Classification", *Linguisticae Investigationes*, 30(1):3-26.
- [3] A. Ritter et al. (2011) "Named Entity Recognition in Tweets: An Experimental Study", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534.
- [4] (2014, Dec 21). <https://about.twitter.com/company>.
- [5] J. R. Finkel et al. (2005) "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363-370.
- [6] R. Yeniterzi (2011) "Exploiting Morphology in Turkish Named Entity Recognition System", In Proceedings of the ACL 2011 Student Session, pp. 105-110.
- [7] G. A. Şeker & G. Eryiğit (2012) "Initial Explorations on using CRFs for Turkish Named Entity Recognition", In Proceedings of the 24th International Conference on Computational Linguistics, pp. 2459-2474.
- [8] S. Özkaya & B. Diri (2011) "Named Entity Recognition by Conditional Random Fields from Turkish Informal Texts" In Proceedings of the IEEE 19th Signal Processing and Communications Applications Conference, pp. 662-665.
- [9] G. Tür et al. (2003) "A Statistical Information Extraction System for Turkish", *Natural Language Engineering*, vol. 9, pp. 181-210.
- [10] O. Bayraktar & T. T. Temizel (2008) "Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach", In 23rd International Symposium on Computer and Information Sciences.
- [11] D. Küçük & A. Yazıcı (2009) "Named Entity Recognition Experiments on Turkish Texts" In Proceedings of the 8th International Conference on Flexible Query Answering Systems, pp. 524-535.
- [12] D. Küçük & A. Yazıcı (2012) "A Hybrid Named Entity Recognizer for Turkish", *Expert Systems With Applications*, vol. 39, pp. 2733-2742.
- [13] S. Tatar & İ. Çiçekli (2011) "Automatic Rule Learning Exploiting Morphological Features for Named Entity Recognition in Turkish", *Journal of Information Sciences*, vol. 37, pp. 137-151.
- [14] G. Çelikkaya et al. (2013) "Named Entity Recognition on Real Data", In Proceedings of the 7th International Conference on Application Information and Communication Technologies, pp. 1-5.
- [15] D. Küçük & R. Steinberger (2014) "Experiments to Improve Named Entity Recognition on Turkish Tweets", In Proceedings of the 5th Workshop on Language Analysis for Social Media, pp. 71-78.
- [16] D. Ramage et al. (2009) "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled corpora", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 248-256.
- [17] X. Liu et al. (2011) "Recognizing Named Entities in Tweets", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 359-367.
- [18] C. Li et al. (2012) "TwNER: Named Entity Recognition in Targeted Twitter Stream", In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 721-730.
- [19] D. Oliveira et al. (2013) "FS-NER A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data", In Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 597-604.
- [20] (2014, Dec 21). <http://crfpp.googlecode.com/>
- [21] K. Oflazer (1994) "Two-Level Description of Turkish Morphology", *Literary and Linguistic Computing*, vol. 9, pp. 137-148.
- [22] H. Sak et al. (2008) "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", 6th International Conference on Natural Language Processing, vol. 5221, pp. 417-427.

- [23] V. Levenshtein (1966) “Binar Codes Capable of Correcting Deletions, Insertions, and Revelsals”, Soviet Physics Doklady, vol. 10, pp. 707-710.
- [24] (2014, Dec 21). <http://www.cnts.ua.ac.be/conll2000/chunking/output.html>
- [25] E. F. Tjong Kim Sang & J. Veenstra (1999) “Representing Texting Chunks”, In Proceedings of the 7th Conference of the European Association for Computational Linguistics, pp. 173-179.

DEVELOPING A FRAMEWORK FOR PREDICTION OF HUMAN PERFORMANCE CAPABILITY USING ENSEMBLE TECHNIQUES

Gaurav Singh Thakur¹, Anubhav Gupta²

¹Cisco Systems, Bangalore email: sai007gaurav@gmail.com,

²Common Floor Technologies, Bangalore email: anubhav992@gmail.com

ABSTRACT

The recruitment of new personnel is one of the most essential business processes which affect the quality of human capital within any company. It is highly essential for the companies to ensure the recruitment of right talent to maintain a competitive edge over the others in the market. However IT companies often face a problem while recruiting new people for their ongoing projects due to lack of a proper framework that defines a criteria for the selection process. In this paper we aim to develop a framework that would allow any project manager to take the right decision for selecting new talent by correlating performance parameters with the other domain-specific attributes of the candidates. Also, another important motivation behind this project is to check the validity of the selection procedure often followed by various big companies in both public and private sectors which focus only on academic scores, GPA/grades of students from colleges and other academic backgrounds. We test if such a decision will produce optimal results in the industry or is there a need for change that offers a more holistic approach to recruitment of new talent in the software companies. The scope of this work extends beyond the IT domain and a similar procedure can be adopted to develop a recruitment framework in other fields as well. Data-mining techniques provide useful information from the historical projects depending on which the hiring-manager can make decisions for recruiting high-quality workforce. This study aims to bridge this hiatus by developing a data-mining framework based on an ensemble-learning technique to refocus on the criteria for personnel selection. The results from this research clearly demonstrated that there is a need to refocus on the selection-criteria for quality objectives.

1. INTRODUCTION

Data mining is the process of extracting useful knowledge from data [8]. It utilizes a combination of a knowledge base, sophisticated analytical skills and domain specific knowledge to uncover many hidden trends and patterns. These patterns and relations can be extracted by using various data-mining algorithms depending on the problem-statement. The application of data mining tools can be extended to diverse fields. Today, the human aspects of software engineering have become one of the critical concerns in IT companies to achieve their business goals. Industries be it in any sector, are now paying attention to selecting the right talent who can perform consistently well throughout all generic framework activities and execute the process properly. Software quality is

greatly dependent on the people and process quality during the development and testing phase. Hence, in this paper we use data mining algorithms to exploit the patterns in the historical data and predict the performance based on project-personnel attributes and thereby enhance the quality of process and the quality of software. Data Mining is the next big revolutionary field that is redefining the industry, be it in terms of technology or research. Here we use mathematical procedures like function-approximation techniques to solve prediction or estimation problems and extract useful trends and patterns from past data to facilitate ourselves to take right decisions with the aim to produce near-optimal results. Classification, allows us to identify association rules. Categorization uses induction algorithm rules to handle categorical outcomes, such as good, average and poor as in this study.

In this paper we address the issue of developing an ideal selection framework for recruiting the right talent which brings us to the basic question of what criteria to follow for the selection procedure. Our aim is to understand the relationship between the various project personal attributes of the candidates and their professional-performance parameter as rated by their managers/supervisors in the industry. Our aim is to find out which are the variables which have the maximum predictive power in estimating the performance capabilities of new recruits. Though there have been many previous studies in this domain, there have been certain issues that still need to be addressed. We try to build upon them and use this research to build a better and more robust model that can be not be applied to different scenarios but also work well on different data sets having varied properties with minimal or no changes. There are a wide range of algorithms in each of these categories, many of which are implemented on WEKA and R [7]. These tools are platforms with GUI and command-line implementations respectively, with a number of machine-learning algorithms for data mining tasks, with a variety of options for regression, classification, data pre-processing, association rules, clustering and visualization. The paper is organized as follows: Section 2 provides the related work and background of data mining Algorithms. Section 3 presents the research methodology to derive at a conclusion. Section 4 discusses about the implementation details while Section 5 depicts the obtained result. Section 6 discusses the summary of this paper and its future scope.

2. RELATED WORK

With increasing complexity of software in the industry and their ever growing demands in multidisciplinary projects, there has been a continuous progress in research works that target the areas of effective project management and Data mining has recently proven itself as one of the most established techniques in this area. Data mining methodologies are developed for several applications including various aspects of software development and we plan to employ this power of algorithms to develop a selection framework. There have been a plethora of studies which incorporate the tools of machine-learning for developing a framework for Prediction of Human-Capability.

The authors in [13] have studied the importance of different variables that come into play during the selection of students - like GPA, Programming Skills, Domain Knowledge Assessment, Reasoning skills, Mental Ability and Mathematical skills, etc using Decision Trees ID3, CART and C4.5. As it turns out, there are many features that must be tested along with GPA to confirm the quality of new personnel. However, the model used in this project has significant scope for improvement and we base our project on the hypothesis similar to this research by Mrs. Sangita Gupta to further improve upon the results obtained using a different approach that involves

Ensemble-learning technique – Random Forest, a bagging based approach to accuracy boosting. Focusing on other works, authors in [4] surveyed different machine learning algorithms for defect prediction in software. Authors of [5] have done a comparative analysis of performance of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Neuro-Fuzzy System for predicting the level of severity of faults in Java-based object oriented software. Data mining results in decision through methods & not through assumptions. Authors in [2] have worked on the improvement of employee selection, by developing a model, using data mining techniques. The specified attributes involved age, gender, marital-status, experience, education, major-subjects and school-tires as potential factors that might affect the performance. As an outcome of their study, it was found that employee-performance is greatly affected by educational-degree, the school-tire, and the work experience. The authors in [3] researched on multiple factors that affect the job performance of employees. They reviewed previous works which study the effect of experience, salary and training, working-conditions and job satisfaction on the performance-parameters. Data-mining thus supports various techniques including Statistical-Analysis, Decision-Trees, Genetic-Algorithm, Bayes-classification, visualization techniques, etc. for analysis and prediction. It further facilitates association, clustering and classification [1]. This research-study involves applying data in WEKA and R tool and derives a classification model for selection criteria. The algorithm we use in this project is Random-Forest, a bagging based accuracy boosting technique which improves upon the result published in [13] which uses Decision Trees namely ID3, C4.5 and CART.

3. RESEARCH METHODOLOGY

The research methodology followed during this study involved the following steps as stated below:

- **Hypothesis:** Project personnel with similar skills set and capabilities will perform similarly.
Project-personal information about the employees is collected from projects which use similar technology and programming language and work on similar platforms. Thus personnel under comparison here have similar capabilities and skill-sets.
- **Data Collection:** Data was collected by using various techniques such as Form-Filling, Brainstorming, obtaining performance information about employees from the Project-leads and managers.
- **Data Preparation:** A basic preliminary research concluded with an almost same set of attributes that must be considered under this problem-statement to obtain a correlation with performance parameters as in [13]. The list of attributes noted are as follows:
 - I. **PS - Programming Skills:** The programming and coding skills of employees were tested using standing coding problems with multiple test cases. A complete score was awarded for each problem only if all the test cases were executed successfully. Employees were tagged as good for scoring above an overall value above 75%, average if they scored between 50% and 75% and the rest were tagged as poor.

- II. **RAS** – Reasoning and Analytical Skills: Similar to programming skills, RAS values were collected through an internal assessment which involved analytical and logical-reasoning questions.
- III. **DSK** - Domain Specific Knowledge: Categorized and normalized similar to PS and RAS. Obtained from a series of internal assessments, DSK refers to both theoretical and field-based knowledge about the domain in which the employees have been working.
- IV. **TE** - Time efficiency: Marked as a simple "Good" or "Bad" option by the project manager/leader of an employee which shows if the employee is considered efficient in completing his/her work within the expected time-period.
- V. **GPA** – Grade Point Average: This is the grade scored by an employee during his/her graduation/ post-graduation courses. It is categorized and normalized to scale the GPAs on the same level based on universities.
- VI. **CS** - Communication Skills: This score for each of the employees was obtained from the project-leaders, their team mates and the respective HR-manager of the employee as well.
- VII. **P** – Performance Parameter: This is our dependent variable for the problem statement which must be correlated to the independent attributes available from the preliminary study of relevant features. The value was acquired by brainstorming with the Team-leads and Managers assigning an overall performance score to the employees in terms of good, average and poor.

This preliminary-study of relevant attributes is followed by the selection of a suitable machine-learning model that must be adopted for the data-mining process. As we have seen, a number of studies have already been done in this domain but the issue has been lack of high-accuracy. Such accuracy issues can be addressed very well if accuracy-boosting techniques are applied. Also given the fact that Knowledge-based Decision Trees have been known to perform decently in such cases, we employ the bagging based Ensemble-learning model Random forest in this scenario to further enhance the performance of old models.

Ensemble learning model - Random Forest

A Classification tree is an input-output model represented by a graphical tree like structure, taking its input (X_i) from an input vector :

$$X = [X_1, X_2, X_3, \dots, X_n],$$

and providing a corresponding output (Y_i) from a set of output possibilities :

$$Y = [Y_1, Y_2, Y_3, \dots, Y_n]$$

These trees may not provide very high accuracies, since they have very high variance values. Randomization based ensemble methods, prove to be a good solution to this flaw. Random-Forest

consists of a collection or ensemble of simple tree predictors, each of which outputs a response when presented with a set of predictor values just as the input vector X . For classification-based problems, this response can be of the forms - class membership or associations, a set of independent predictor values with one of the categories present in the dependent variable. Each tree is created from its own separate bootstrapped sample training set. The *Bootstrap Sampling Method* samples the given training tuples uniformly with replacement i.e. each time a tuple is selected, it is equally likely to be selected again and rendered to the training set. As the number of simple learning models within an ensemble technique increases, the overall variance of the output-value from the actual-value theoretically decreases by $1/(\text{number of individual models})$. However this decrease in variance after a threshold doesn't yield significant improvement and that allows us to decide the number trees we want to create for this random forest technique. It's important to remember that ensemble learning techniques are computationally expensive and hence choosing an optimal value for the number of individual simple predictors within the ensemble technique is a critical task. Individual Decision-trees usually suffer from high-variance, which makes them uncompetitive in terms of accuracy. A highly efficient and simple way to address this issue is to adopt the context of randomization and use them in ensemble-methods.

The Mean-Decrease-in-Accuracy of a variable is evaluated during the calculation-phase of out-of-bag error. As the fall in accuracy of the random-forest increases due to the addition of a single-variable, the more important the particular variable under test is considered and hence variables with a large value for Mean-Decrease-in-Accuracy or Gini are considered as more important for data classification. The Mean-Decrease-in-Gini coefficient is a measure of how each particular variable supplements to the homogeneity of the nodes and terminal-leaves in the resulting Random-forest. Every time one particular variable splits a given node, the Gini coefficient for the children are calculated and compared to that of the original parent node. If the same variable causes multiple splits more than once, then the final difference in the Gini value of the topmost parent node and the bottom-most children nodes is taken as the Mean-decrease-in-Gini value. Using these values, a final graph for *Variable Importance* is plotted, where this graph represents each variable on the vertical y-axis, and their importance-values on the horizontal x-axis. They are ordered in the manner of top-to-bottom as maximum-to-minimum-importance. To measure the accuracy of the classifier, we made use of Sensitivity and Specificity parameters. The following are the meaning of the variables used in the subsequent equations.

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

1. **True-Positive Rate** or Sensitivity is the fraction of training samples predicted correctly by model.

$$TPR = \frac{TP}{TP+FN} \quad (1)$$

where TPR represents the True-positive-Rate and higher this value, the better the model is.

2. **False-Positive Rate** or Specificity i.e. the fraction of training samples predicted incorrectly by model.

$$FPR = \frac{FP}{TN+FP} \quad (2)$$

where FPR represents the False positive Rate and lower this value, the better the model is.

3. **Area under ROC curve (Receiver Operating Characteristics):** is obtained by plotting TPR against FPR. The area under the plotted graph gives a good measure of the accuracy of the classifier. The area can be as high as 1 Sq. unit (maximum accuracy) and as low as 0 (minimum accuracy). Since Area-Under-ROC-curve takes into consideration, both FPR and TPR values, this measure is preferred over the parameters to compare accuracies between models.

4. IMPLEMENTATION DETAILS

Based on the study conducted, data obtained was consolidated and summarized in a tabular form. The algorithm used here is Random Forest, implemented using WEKA and R under "Test options". 10 - Fold Cross validation was applied to supplement the out-of-bag calibration mechanism of Random-Forests. The number of individual predictor-trees was set to 500 in R as no significant reduction in variance was observed beyond this value. Trees were allowed to grow completely without any pre or post pruning. The package used in R for implementing Random Forest is the "randomForest" package which is compatible with versions 4.6 and above. This package is available for download at the official R support website.

For creating bootstrap samples, we used the technique of 632 Bootstrapping, which means that in any bootstrap sample generated, approximately 63.2% of the Dataset will be unique, and the rest would be placed with replacement and duplication. Studies have shown that this bootstrapping technique produces near-optimal results. The plot for Variable Importance has been obtained using R tool and the results obtained are discussed below. The model generated cannot be visualized graphically due to the large number of trees generated, each created from a separate bootstrap sample and each producing its own results. As stated earlier, the final output from the Random Forest model is the average of the results obtained from each of the predictor trees in the forest.

5. RESULTS AND DISCUSSION

Table 1 shows the results obtained by Random forest. Table 2 shows the Variable Importance values for each of the attributes in terms of Mean-Decrease in Gini and Mean-Decrease in Accuracy.

CLASS	TP RATE	FP RATE	Area under ROC curve
Good	0.934	0.037	0.977
Average	0.846	0.037	0.983
Poor	0.929	0.077	0.992

Table 1 : Accuracy Measures for samples classified by output class

Firstly we look at the average area under the ROC curve and as we can see, this area is about 0.984 (average). The area here is close to 1 and for ROC curves, an area 1 refers to the highest

possible accuracy of 100%. Hence, we can see that the Random-Forest Model used is highly accurate and strong conclusions can be drawn from these results. We also found through a comparative study that the model outperformed knowledge-based decision trees and Linear Regression techniques when applied to the same data.

We now take a look at the Variable importance plot using Mean-Decrease in Gini and Mean-Decrease in Accuracy values.

	DSK	RAS	PS	CS	TE	GPA
Mean Decrease Accuracy	35	10	9	0	0	0
Mean Decrease Gini	14	5	4	2	1	2

Table 2 : Variable Importance measures for each Attribute

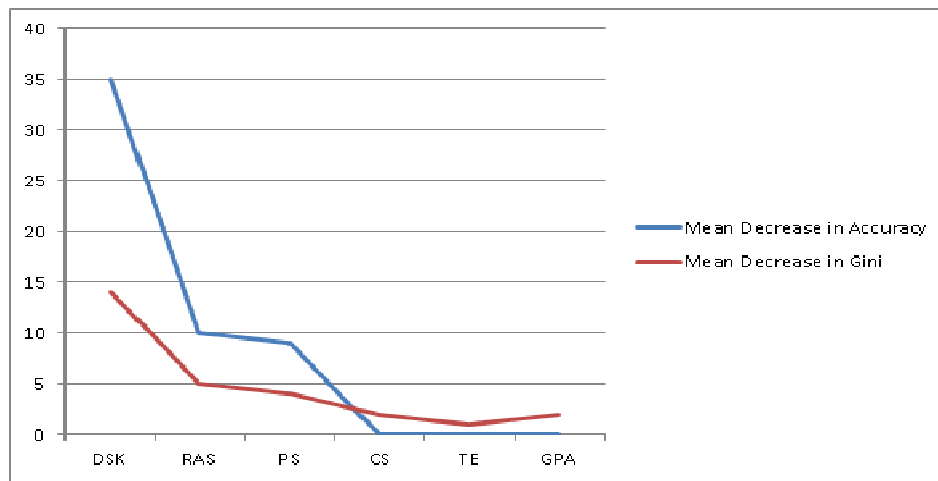


Figure 1 Plot for Variable Importance of different attributes

As we see in these plots, the variables like Domain-Specific knowledge, Reasoning and Analytical skills and Programming Skills are the most important attributes to be considered during recruitment of new personnel, as they have the maximum contribution towards the homogeneity of nodes and classification of data. Another important result obtained in this scenario is the Mean Decrease in Accuracy value for GPA which comes out to be 0. Such low scores for these values stand as a scientific base to challenge the usual recruitment procedure where maximum importance is given to GPA of candidates. What we see here is that, based on the grades and academic scores of students from universities, one cannot predict their performance in the industry. Hence GPA alone is not a very clear reflection of the candidate's capabilities as far as the software industry is concerned. It is needed that the companies separately test the other relevant attributes of the students in order to make a better decision. The software development process involves various intricate steps and complex stages where many other factors and abilities of an individual come into play. Referring to the GPA alone will not yield optimal results and this is the reason, why there has been a changing trend in the recruitment procedure today. Recruiting teams are looking for candidates with a complete package in terms of overall personality, analytical thinking abilities and good inter-personal skills apart from good grades. The importance given to personal interviews lays emphasis on the fact that only after

getting a true idea about the candidate's knowledge in the particular domain through a one-on-one interaction, companies take a final decision regarding the selection procedure.

Hence, using the results from the variable importance plot, we construct an ideal knowledge based decision tree which would be used as a selection criteria and also illustrate a tree pruning mechanism that can be incorporated in appropriate scenarios. Also, we use the Mean-Decrease-in-Gini Parameter for tree-construction to obtain maximum homogeneity of nodes. The ideal-Decision Tree obtained in this case is shown below:

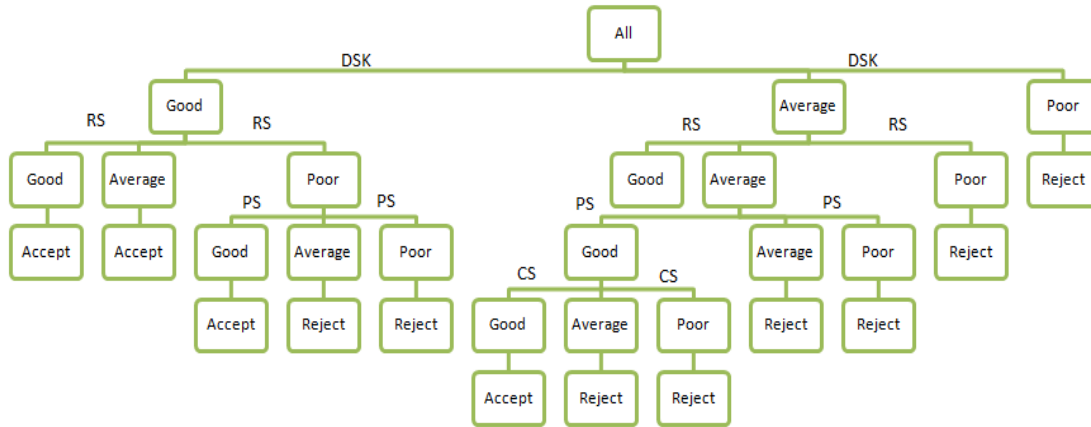


Figure 2 Suggested Ideal Tree To follow While Recruiting

As we see, during the selection process companies must initially clarify the values for each of the variable-parameters that are acceptable to them. For example some companies may accept students with average programming skills but others may only want those who have great programming skills whatever may be the scores for other attributes. Once that's done, they must start classifying students based on the features in the order of their variable importance values. Another important aspect here is to remember that in numerous scenarios, there may be a very large feature set. In such a case, it is not possible for the companies to consider all of them. Hence, to modify the selection tree, a tree pruning mechanism is developed as follows:

1. Set the threshold limit that must be used to prune to the tree, say 'P' percentage of pruning is needed.
2. Evaluate the percentage importance values using the following formula. If for variable X_i the Mean-Decrease-in-Gini score is given as $Imp(X_i)$ then evaluate Percentage $Imp(X_i)$ as

$$Percent_{Imp(X_i)} = \frac{Imp(X_i)}{Imp(X_1) + Imp(X_2) + Imp(X_3) + \dots + Imp(X_n)} * 100$$

3. Now select the variables with minimum $Percent_{Imp(X_i)}$ scores and sum them until this total greater than or equal to 'P'.

$$Percent_{Imp(X_i)min} + Percent_{Imp(X_{i+1})min-1} + Percent_{Imp(X_{i+2})min-2} + \dots + Percent_{Imp(X_{i+n})min-n} \geq P$$

4. All those variables X_i with minimum $Percent_{Imp(X_i)}$ values that were added to the sum $\geq P$ should not be used during the construction of the ideal-tree.

This procedure allows us to cut down on the least important variables to be considered during the selection-procedure using the pre-pruning technique. Such a method would be helpful in scenarios when the number of variables is very high and it's not practical for the companies to look at every attribute. In case of a tie, it totally depends on the discretion of the hiring manager to exclude or include those features from the ideal-selection tree.

Though this model provides high accuracy and allows pruning techniques to adapt to real-time scenarios, there is still scope for further improvement using cost optimization techniques like Gradient Boosting and many more, which would yield better results when dealing with higher dimensional data and inter-dependent variables. In order to develop a selection-framework for other domains, a similar procedure can be followed after incorporating minimal changes to suit them appropriately.

6. CONCLUSION

The primary target of data-mining is to produce near optimal results using the information extracted from patterns and trends hidden in historical data. This brings us to the essential question of choosing the best suitable model that can be applied to any given problem statement. In this scenario, the models of Decision trees have already been tested on a similar dataset with similar attributes and properties in [13]. However, this model suffers from the problem of overfitting and we try to overcome this issue by applying the Random-Forest technique. Also the fact that even when most studies provide a prioritization mechanism for selection procedure, they lack a quantifying measure to represent their importance in the framework.

The Random Forest method is a bagging based accuracy boosting technique, which creates multiple trees out of the bootstrap samples generated from a dataset, hence forming a forest. Here the output from each tree is considered to calculate an overall mean or an average result for the random forest. Since the calibration of the model is done using the out-of-bag samples, the model does not suffer from over-fitting issues which facilitates a superior performance in most of the scenarios when compared to Decision-trees. Apart from these enhancements in accuracy, Random forests also provide supplemental information like variable importance measures, etc. which adds to its value. One of these measures is the Mean Decrease in Gini index and Mean Decrease in Accuracy parameters which gives a holistic view about the contribution of each of the attributes to the final output. We utilize this facility provided by Random-Forest to assign an order of priorities to the features that must be considered during the recruitment of new-personnel.

We clearly see from the results that the GPA/grades of new recruits clearly aren't among the most important selection attributes on which a hiring-decision can be based. Other attributes like Programming Skills, Domain Specific Knowledge and Analytical Skills must be separately tested as they have a significant prediction power in estimating the performance of a person in the software industry. However, only obtaining a number of important parameters using a simple

model do not suffice this need as the degree of accuracy associated with the model is equally important. This critical need for an accurate model would be clearly visible when developing a framework for other domains and working on high-dimensional data and this where Accuracy-boosting techniques like random forests come into play. These algorithms are not only known for their robustness but also perform well in cases of high-dimensional data and sparse data sets. Data-Mining algorithms extract important patterns and have helped in identification of project-members who have a greater probability to perform well. This research not only empowers the managers to refocus on Human-Capability criteria to enhance the development-process of any software-project but also address the issue where due to lack of analytical-methods in human-aspects, IT companies could not select the right talent in the software process and hence failed to achieve the desired-objectives in terms of both, quality and quantity in the time and cost-constraints. The scope of this study can be further extended to various domains and appropriate changes need to be incorporated when developing a selection-framework for each one of them. Such a study would require a preliminary research about finding the relevant features that can be directly correlated to the performance of any given employee in that particular domain. Once this is done, a similar model can be used to develop a feature-ranking order and enlist the most important attributes to be considered to make a hiring-decision using such a framework.

REFERENCES

- [1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques, 2/e", Morgan Kaufmann Publishers, An imprint of Elsevier, 2010.
- [2] C.F. Chien and L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", *Expert Systems and Applications*, vol. 34, 2008, pp. 280-290.
- [3] Hamidah Jantan et al, "Human Talent Prediction in HRM using C4.5 Classification Algorithm", (IJCSE) *International Journal on Computer Science and Engineering* Vol. 02, No. 08, 2010, pp. 2526-2534.
- [4] Suma.V, Pushpavathi T.P, and Ramaswamy. V, "An Approach to Predict Software Project Success by Data Mining Clustering", *International Conference on Data Mining and Computer Engineering (ICDMCE'2012)*, pp. 185-190.
- [5] P. Singh, Comparing the effectiveness of machine learning algorithms for defect prediction, *International Journal of Information Technology and Knowledge Management*, 2009, pp. 481-483.
- [6] J. R. Quinlan, "Introduction of decision tree", *Journal of Machine learning*, 1986, pp. 81-106.
- [7] Witten I. Frank E., and Hall M., "Data Mining: Practical Machine Learning Tools and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2011.
- [8] A. Kusiak, J. A. Kern, K. H. Kernstine, and B. Tseng, "Autonomous decision-making: A data mining approach," *IEEE Trans. Inform. Technol. Biomedicine*, vol. 4, no. 4, pp. 274-284, Aug. 2000.
- [9] A. S. Chang, & Leu, S.S., "Data mining model for identifying project profitability variables," *International Journal of Project Management*, vol. 24, pp. 199-206, 2006.
- [10] T. R. Gopalakrishnan Nair, V. Suma, Pranesh Kumar Tiwari, "Analysis of Test Efficiency during Software Development Process", 2nd Annual International Conference on Software Engineering and Applications (SEA 2011)
- [11] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc, 1992.
- [12] Oded Maimon, Lior Rokach, "The Data Mining and Knowledge Discovery Handbook", Springer publication, 2005.
- [13] Sangita Gupta, Suma V, "Empirical study on selection of team members for software project- A data mining approach", *International Journal of Computer Science and Informatics*, ISSN (PRINT): 2231-5292, Vol 3, no 2, 2013, pp 97-102. Sangita Gupta, Suma V: Prediction of Human Performance Capability during Software Development using Classification. *ICT and Critical Infrastructure*:

AUTHORS

Gaurav Singh Thakur

Gaurav Singh Thakur has completed his B.Tech in 2014 in Information Technology from National Institute of Technology Karnataka, Surathkal and is currently working as a Software Engineer at Cisco Systems, Inc. Bangalore. His technical areas of interest include Machine learning, Networking & Security, Application Development and Algorithms.



Anubhav Gupta

Anubhav Gupta has pursued his Bachelors, at National Institute of Technology Karnataka, Surathkal in the Field of Information Technology (IT) and graduated in 2014. His technical areas of interest include Machine learning, Information Security, Web Development and Algorithms. Currently, he is working as Software Developer Engineer at Commonfloor (MaxHeap Technologies).



INTENTIONAL BLANK

KNOWLEDGE MANAGEMENT IN HIGHER EDUCATION : APPLICABILITY OF LKMC MODEL IN SAUDI UNIVERSITIES

Farzana Shafique

University of Dammam, Saudi Arabia

ABSTRACT

This paper stresses on the need of using Knowledge Management (KM) in the higher education institutions of Saudi Arabia. The paper is based on the literature review and personal experience of the author in the education sector.

The paper aims at highlighting the importance of KM for the educational institutions particularly for developing countries. It also reviews the readiness of Saudi Arabia for KM application by illustrating different development initiatives taken by the Saudi government in different sectors. However, the literature also identifies many barriers on the way.

Keeping the importance of KM for the higher education institutions in view, this paper proposes to adopt the model of Library Knowledge Management Center (LKMC) with needed modifications for Universities of Saudi Arabia. This LKMC model was proposed by Parker, Nitse, and Flowers (2005) for the small business corporate for providing the Knowledge Management (KM) and Competitive Intelligence (CI) services. The paper discusses different components of the LKMC model and their relevance to the education sector.

KEYWORDS

Knowledge Management; Higher Education; LKMC; Kingdom of Saudi Arabia

1. INTRODUCTION

Many of us simply do not think in terms of managing knowledge, but we all do it. Each of us is a personal store of knowledge with training, experiences, and informal networks of friends and colleagues, whom we seek out when we want to solve a problem or explore an opportunity. Essentially, we get things done and succeed by knowing an answer or knowing someone who does (NHS National Library for Health, 2005).

Many researchers believe that knowledge sharing is the vital element in KM (Firestone, 2001). Thus, KM is a process where organizations have formulated ways in the attempt to recognize and archive knowledge assets within the organization that are derived from the employees of various departments or faculties and in some cases, even from other organizations that share the similar

area of interests or specialization (Bouthillier, & Shearer, 2002). In this context, an institution wide approach to KM can largely facilitate the knowledge sharing process; both explicit and tacit, and the subsequent surge benefits

Purpose: The paper aims at identifying the benefits of KM for education sector. It focuses on highlighting the case of developing nations and reviews the current practices of KM in these nations. The paper also analyses the readiness of Saudi Arabia for KM and it proposes adopting the Library Knowledge Management Center (LKMC) model for higher education institutions of Saudi Arabia.

Methodology: The paper is based on review of literature and author's personal teaching and administrative experience in higher education. It discusses different components of LKMC model proposed by Parker, Nitse, and Flowers (2005). For literature search, standard sources of scholarly information were used e.g., Library Literature, World Wide Web search engines and scholarly online databases accessible through University of Dammam portal i.e., ScienceDirect, Emerald, and many more.

Need of KM in Higher Education Institutions

Laal (2011) has stressed on the use of Corporate Sector's Knowledge Management (KM) practices in the higher education institutes. He has reviewed many studies to supplement his claim and has mentioned that KM is a systematic process by which knowledge needed for an organization to succeed is created, captured, shared, and leveraged. Nowadays, the pace of evolution has entered a rapid speed; those who cannot learn, adapt, and change from moment to moment simply won't survive. While discussing the case of developed world, Laal has mentioned that current higher education institutes recognize their valuable intelligences and have adopted their changing role in a society.

Rowley (2000) believes that educational institutes particularly higher education are said to be in the knowledge business since they are involved in knowledge creation, dissemination and learning. Kidwell, Linde, and Johnson (2000) also stressed on the use of KM techniques and technologies in higher education and called it as vital as it is in the corporate sector. They mentioned that educational institutes i.e., colleges and universities have significant opportunities to apply KM practices to support every part of their mission, from education to public service to research. Effective use of knowledge capital in educational institutes can lead to better decision-making capabilities, reduced "product" development cycle time, improved academic and administrative services, and reduced costs. Similarly, Ramanigopal (2012) said that higher education has significant opportunities to apply knowledge management practices to support every part of their mission, from education to learning society to research and development.

Many educational institutions are spending millions of dollars into information technology without considering the effective integration of the same into shared decision-making processes to improve academics, operations, and planning. On the other hand, many of these educational institutions are farther along in developing an "information culture", yet lag behind in their "technology culture". Here, the Knowledge Management (KM) practices can provide great help to the educational institutes to overcome this problem. Because the primary benefit of the KM is that it actively addresses both the "technology culture" and the "information culture" at an institution, and seeks to advance both simultaneously (Petrides, & Nodine, 2003). Omona and

Lubega (2012) proposed conceptual framework for enhancing knowledge management (KM) in higher education in order to advance strategic goals and direction. The key dimensions of their proposed framework were tested using case studies of higher education institutions (HEI) in Uganda to examine relative use and effectiveness of the current existing KM enabling ICT tools and technologies. Their study also aimed at identifying key KM processes and determining critical success factors.

Gopal and Shobha (2012) studied the general understanding of students about Knowledge Management, the opportunities available for their Knowledge Management and KM practices adopted by them in their higher education. They recommended integrating the KM in university teaching and learning process. Steyn (2004) stressed on the importance of KM in the higher education institutions and proposed a model for effective implementation of KM in the higher education institutions. Steyn believed that successful organizations are knowledge creating organizations, which produce, disseminate and embody new knowledge in new products and services

KM in Higher Education: The Case of Developing Countries

The trend of using, KM in educational institutions is also prevailing in the developing world. For example, Abass, Hayat, Shahzad, and Riaz, (2011) in a recent descriptive survey revealed the acceptability of KM practices in public sector of Pakistan. They surveyed two educational institutes, e.g., secondary education board and a higher education institute. During their research, they found positive perceptions of employees towards organizational culture, KM practices, and organizational performance. Their descriptive results showed the positive trend of each of the variable. According to the employees working in different organizations under government sector of Pakistan, the culture of their respective organizations encourages and provides opportunity for the communication of ideas and knowledge; organization encourages and rewards the sharing of knowledge. Moreover, performance of the surveyed organizations was satisfactory in terms of employees' satisfaction, their retention, and operating costs etc.

Yusoff, Mahmood, and Jaafar (2012) explored the relationship of KM implementation and KM enabler among members in a community college of Malaysia. The researchers identified the possibility of developing a socioeconomic center at the community colleges with strong links to the government, coordinated by a management, which actively supports the technology and knowledge transfer and provides communities with facilities and services. This attracts mainly local communities who expect benefits and synergies from these colleges. The researchers believed that their research had just opened a space to improve the KM. In future researchers would recognize the main activities in community college that can be defined clearly as part of KM and then they would further propose a framework for those activities that can be used practically by the staff of community college. Similarly, Chumjit (2012) explored the application of knowledge management (KM) in the higher education institutions of Thailand. Based on a qualitative research design, Chujit studied four autonomous universities of Thailand and reported that the four universities have tried to create new knowledge (both tacit and explicit knowledge). New methods for improving teaching, research, administration, and strategic planning have been created and KM has been successfully applied within various sections and departments.

KM Initiatives in Saudi Arabia

The Knowledge Management is an emerging concept in Saudi Arabia. Many new KM initiatives have been taken by the Government of Saudi Arabia. The remarkable example is of “Knowledge Economic City” (KEC) project at Madinah. The KEC is aimed to serve Saudi Arabia’s economic diversification strategy and reviving Madinah’s role as a center for the Islamic knowledge and a global knowledge and culture center. Through this project, the Saudi government aims to enhance the quality of life and economic prosperity in the region. It would create many investment and development opportunities across all the sectors through commercial, residential, educational, and hospitality projects (Knowledge Economic City, 2012).

Yaghi and Zamzami, 2014) stressed on the importance and need of knowledge management in the higher education institutions of Saudi Arabia. They highlighted the major constraints on the way and provided a framework for applying the concept of KM in higher education institutions of Saudi Arabia. They suggested that the higher management in these educational institutions should adopt the strategic thinking of knowledge management. They also stressed on the need of knowledge sharing among employees/workers of these institutions, so that the tacit/implicit knowledge can be explicit. Al-Hussain (2011) probed the Barriers to Knowledge Management in Saudi Arabia through an intensive research. The author’s research and analysis led to the identification of barriers related to organization (19 barriers), technology (24 barriers), leadership (32 barriers), and learning (22 barriers).

Despite all these barriers and hurdles, government policy documents clearly indicate their awareness of these issues and draws framework for overcoming the same. For example, the Ministry of Education (2004) documents identifies the weaknesses in the education system and stresses on the need of knowledge intensive education system. Alsereihy, Alyoubi, and El-Emary (2012) reviewed some case studies on KM implementation in Saudi Arabia. The reviewed studies were related to different ongoing initiatives/projects in Saudi Arabia; such as schools for girls, public sector firms, oil, and chemical factory and construction firms. They stressed that the initiatives would be successful only when more people have participated in the initiative and they have exchanged their views thus creating knowledge. They also stressed on the need of using Social media for knowledge sharing purpose. Similarly, Abokhodiar (2013) proposed a model for the implementation of KM at the Women’s Branch of the Institute of Public Administration in the Kingdom of Saudi Arabia. She mentioned that several perspectives urged the need of KM implementation at WIPA. According to her, the most important one was the current status of the workforce and the generation gap between faculty members. This could result in a loss of a portion of the accumulated knowledge and experiences of the intellectual capital of WIPA. Her proposed model aimed to provide an integrated strategy of KM implementation at WIPA by 2018.

Library Knowledge Management Center (LKMC) Model for Higher Education Institutions of the Saudi Arabia

Parker, Nitse, and Flowers (2005) foresaw a dynamic and influential role of libraries and information centers for providing KM and CI services to the other sectors. Competitive Intelligence (CI): Organizations use the CI process to gather information, to add value to it through analysis, and to report the findings to managers to solve a wide variety of problems or satisfy requests for information.

Although they stressed on the applicability of LKMC (Library Knowledge Management Center) model for small business sector, however, their proposal has equally great potential of applicability for not-for-profit organizations particularly education sector. Libraries are the integral part of education sector; on the other hand, libraries are struggling for their own survival in this digital world and arena of escalating fiscal resources for libraries. In this perspective, the role of libraries as LKMC will not only benefit the libraries but it would also be a great contribution to education sector.

Components of a Library Knowledge Management Center (LKMC)

After a comprehensive review of related literature, Parker, Nitse, Flowers (2005) believe that several steps must be undertaken for libraries to act as KM centers, such as developing domain ontologies to help categorize resources for specific clients; thorough understanding of the domain; Domain ontology provides a specification of a shared conceptualization to be used for formulating knowledge-level theories about a domain (Domingue and Motta, 1999). Thus, they have to develop the general and then specific domain terms. The effective use of needs identification tools like key intelligence topics (KITs) is also required. The KITs process can help to identify and define critical intelligence needs (Herring, 1999).

Similarly, the Multi-Class Interest Profile (M-CLIP) can also be developed in-place or parallel to KITs; the M-CLIP provides a strategically aligned framework based on the various types of information needs in order to insure that key items within each domain are accounted for. The KITs and M-CLIP techniques are generally used in conjunction with Competitive Intelligence (CI) etc. (Parker & Nitse, 2001).

Moreover, Parker and Nitse (2001) suggested the hiring of a specialist trained in knowledge engineering, a specialist trained in knowledge engineering who can greatly assist the specification of key concepts for the domain ontology. The use of natural language processing techniques for determining the contents of each digital document; collecting and managing internal and external sources or even subscribing databases from third party vendors or other informal online sources are also suggested as important elements of LKMC model. They further ascertained on the management of internally generated knowledge, which refers to that knowledge within the minds of their employees. In order to handle such knowledge, the system should provide an interface to allow users to store information that will be sharable with other users of the system through customized and interactive search interfaces. Furthermore, the emphasis should also be on preserving and further utilizing the users' search feedback while searching in library catalog. The final suggestion was the utilization of Semantic Web for semantically linking and then retrieving or accessing the library's resources. The Semantic Web promises to give well-defined meaning to the web by incorporating into web documents well-defined semantics. Agents should be able to determine the semantic linkages between web resources by following links from web pages to topic-specific ontologies.

They further claimed that Results can be delivered on a push or pull basis to provide ongoing competitive (and other) intelligence. For bearing the expanses of the center and for its long term sustainability, they further suggested that a subscription fee can be charged from the clients for providing the user specific services of LKMC. The alternative approaches such as web portals for individuals or communities to access the internet and to conduct research on topics of interest to them can also be utilized for making LKMC more effective (Sadeh and Walker, 2003). However,

for coping with the new developments after the proposed model of Parker and Niste (2005), one should focus on providing KM and Environment Intelligence (EI) services to the end users. According to McGee and Sawyer, (2003), the environment intelligence is an encompassing concept covering disciplines such as business intelligence, competitor intelligence, competitive intelligence, and social intelligence. The environment intelligence is the collection of information about events and changes happening in the external environment of an organization by using legal and ethical information gathering channels and techniques. The external environment here refers to relevant social and physical factors outside the typical boundaries of an organization which may affect its performance and future survival.

2. CONCLUSIONS AND RECOMMENDATIONS

The review of the literature made in this article has supported the idea of using knowledge management in the higher education institutions. It shows that the now developing countries are also becoming aware of knowledge management. Many new initiatives have been taken in this regard. Saudi Arabia is also emerging as an active KM enabler from the developing world. Many new KM projects particularly the initiation of “Knowledge Economic City” (KEC) project at Madinah shows the Saudi Arabia’s readiness for knowledge society. In this context, the paper proposes implementing the LKMC model with necessary modifications in the universities of Saudi Arabia. It is believed that if implemented truly, it would boost up the higher education of Saudi Arabia. It is suggested that the system/center proposed by Parker and Niste (2005) should look at competitive environment in broader perspective. Besides, the Knowledge Management (KM) services, it should also be able to provide Environment Intelligence (EI= Competitive, Competitor, and Social Intelligence) services to the educational administrators.

REFERENCES

- [1] Abass, F., Hayat, M., Shahzad, A., & Riaz, A. (2011). Analysis of Knowledge Management in the Public Sector of Pakistan. *European Journal of Social Sciences*, 19, (4), 471-478.
- [2] Abokhodiar, E. S. (2013). Knowledge management implementation at the Women’s Branch of the Institute of Public Administration in Saudi Arabia: A Proposed Model. *Excellence in Higher Education*, 4: 119-128
- [3] Bouthillier, F. & Shearer, K. (2002). Understanding knowledge management and information management: the need for an empirical perspective. *Information Research*, 8(1).
- [4] Chumjit, S. (2012). Knowledge management in higher education in Thailand. ProQuest, UMI Dissertations Publishing.
- [5] Domingue, J. & Motta, E. (1999). A knowledge-based news server supporting ontology-driven story enrichment and knowledge retrieval. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling, and Management (EKAW 1999)*, Dagstuhl Castle, Germany, Springer-Verlag, Berlin, pp. 103-20.
- [6] Firestone, J. M. (2001). Key Issues in Knowledge Management. *Knowledge and Innovation. Journal of the KMCI*, 1(3), 8-38.
- [7] Herring, J. P. (1999). Key intelligence topics: a process to identify and define intelligence needs. *Competitive Intelligence Review*, 10(2), 4-14.
- [8] Kidwell, J. J., Linde, K. M. V., & Johnson, S. L. (2000). Applying corporate knowledge management practices in higher education. *Educause Quarterly*, 4, 28-33.
- [9] Laal, M. (2011). Knowledge management in higher education. *Procedia Computer Science*, 3, 544–549.

- [10] McGee, J. E. & Sawyer, O. O. (2003). Uncertainty and information search activities: A study of owner-managers of small high-technology manufacturing firms. *Journal of Small Business Management*, 41(4), 385-401.
- [11] NHS National Library for Health. (2005). What is knowledge management? *ABC of Knowledge Management*; 1-68.
- [12] Parker, K. R. & Nitse, P. S. (2001). Improving competitive intelligence gathering for knowledge management systems. In *Proceedings of the 2001 International Symposium on Information Systems and Engineering – ISE'2001-Workshop: Knowledge Management Systems: Concepts, Technologies and Applications*, Las Vegas, Nevada.
- [13] Parker, K. R., Nitse, P. S., Flowers, K. A. (2005). Libraries as knowledge management centers. *Library Management*, 26(4/5), 176-189.
- [14] Petrides, L. A. & Nodine, T. R. (2003). Knowledge management in education: Defining the landscape. Retrieved October 25, 2014, from: <http://iskme.path.net/kmeducation.pdf>
- [15] Rowley, J. (2000). Is Higher Education Ready for Knowledge Management? *The International Journal of Educational Management*, 14(7), 325-333.
- [16] Sadeh, T. & Walker, J. (2003). Library portals: toward the Semantic Web, *New Library World*, 104(1184/1185), 11-19.
- [17] Yusoff, M. Y. M., Mahmood, A. K., & Jaafar, J. (2012). A Study of KM process and KM enabler in a Malaysian Community College. *Journal of Knowledge Management Practice*, 13(1). Retrieved October 24, 2014, from: <http://www.tlinc.com/article297.htm>
- [18] Knowledge Economic City. (2012). The New Gateway to Madinah. Retrieved December 12, 2014, from: <http://www.madinahkec.com/en/project/kec>
- [19] Yaghi, K., & Zamzami, O. A. (2014). Obstacles of Implementing Knowledge Management in the High Education Institutes - Saudi Arabia (Analytical study). *International Multilingual Academic Journal*, 1(1).
- [20] Al-Hussain, A. Z., (2011). Barriers to Knowledge Management in Saudi Arabia. Unpublished Dissertation, The George Washington University, Proquest. 213 pages; 3481091. Ministry of Education. (2004). The Development of Education, 47th Session of the International Conference on Education, September 8–11, 2004, Geneva.
- [21] Alsereihy, H. A., Alyoubi, B. A., & El-Emary, I. M. M. (2012). Effectiveness of Knowledge Management Strategies on Business Organizations in KSA: Critical Reviewing Study. *Middle-East Journal of Scientific Research*, 12 (2): 223-233.
- [22] Omona, W., & Lubega, J. T. (2012). Enhancing Knowledge Management Using ICT in Higher Education: An Empirical Assessment, *Journal of Knowledge Management Practice*, 13(3).
- [23] Gopal, V., & Shobha, K. (2012). Knowledge management in higher education. *Asian Journal of Research in Social Sciences and Humanities*, 2(8). Retrieved December 12, 2014, from: <http://www.indianjournals.com/ijor.aspx?target=ijor:ajrssh&volume=2&issue=8&article=006>
- [24] Ramanigopal, C. (2012). Knowledge management strategies in higher education. *International Journal of Advanced Research in Management (Ijarm)*, 3(1), pp. 20-29.
- [25] Steyn, G.M. (2004). Harnessing the power of knowledge in higher education, *Educational Development*, 124(4), 615-630.

INTENTIONAL BLANK

AN APPROXIMATE POSSIBILISTIC GRAPHICAL MODEL FOR COMPUTING OPTIMISTIC QUALITATIVE DECISION

BOUTOUHAMI Khaoula and KHELLAF Faiza

Recherche en Informatique Intelligente et Mathématiques Appliquées.
Université des Sciences et de la Technologie Houari Boumediene.
Algiers, Algeria

boutouhami_khaoula@yahoo.fr, hanedfaiza@yahoo.fr

ABSTRACT

Min-based qualitative possibilistic networks are one of the effective tools for a compact representation of decision problems under uncertainty. The exact approaches for computing decision based on possibilistic networks are limited by the size of the possibility distributions. Generally, these approaches are based on possibilistic propagation algorithms. An important step in the computation of the decision is the transformation of the DAG into a secondary structure, known as the junction trees. This transformation is known to be costly and represents a difficult problem. We propose in this paper a new approximate approach for the computation of decision under uncertainty within possibilistic networks. The computing of the optimal optimistic decision no longer goes through the junction tree construction step. Instead, it is performed by calculating the degree of normalization in the moral graph resulting from the merging of the possibilistic network codifying knowledge of the agent and that codifying its preferences.

KEYWORDS

Possibilistic decision theory, Min-based possibilistic networks, Moral graph, optimistic criteria.

1. INTRODUCTION

Decision making under uncertainty plays an important role in Artificial Intelligence (AI) applications. Several decision making tools have been developed to assist decision makers in their tasks: simulation techniques, dynamic programming, logical decision models and graphical decision models. This paper focuses on graphical decision models which provide efficient decision tools by allowing a compact representation of decision problems under uncertainty [14]. A decision problem is a choice between a list of possible alternatives taking into account the knowledge of an agent (knowledge is sometimes tainted with uncertainties) as well as his/her preferences. The results of his/her decision are expressed by a set of utilities. The qualitative possibilistic decision model allows a progressive expression of preferences as well as knowledge of the decision-maker. This model offers two qualitative criteria of utilities for the approach of decision under uncertainty: the pessimistic decision criterion and the optimistic decision criterion. Interest in the issue of the calculation of qualitative decision continues to grow and many approaches and models have been proposed [7][12].

In addition to the calculation of decision, the aim of this method is to improve other methods and overcome their limits as regards the presentation form, the calculation time as well as the ease of understanding. In our work we focus on graphical decision models that provide effective tools for decision problems under uncertainty using a compact representation. Several evaluation methods have been proposed to select the optimal decision. Among these methods, there is an exact approach based on possibilistic networks spread. This approach requires a transformation of an original graph into a secondary structure called the junction tree [11] which is used then in various calculations. In this work, our goal is to propose a new approximate approach to compute the optimal optimistic decision. Our approach is based on the moral graph associated with the result of merging the networks representing the agent's beliefs and preferences. This approach has a polynomial complexity [1]. Indeed, it avoids the transformation of the initial graph into a junction tree which is known to be an intractable problem (NP-hard). Using the approximate approach provides very close answers to the exact marginal distributions [1].

The reminder of the paper is organized as follows. The next section briefly recalls the fundamental concepts of possibility theory and min-based possibilistic networks. The main results of merging min-based possibilistic networks are also briefly presented in this section. Section 3 describes the new approach and its use in calculating the optimal optimistic decision and section 4 concludes the paper.

2. BACKGROUND

2.1 Basic concepts of possibility theory

This section gives a brief refresher on possibility theory which is issued from fuzzy sets theory [16], [13] and represents a main approach in dealing with uncertainty. Let $\mathcal{V} = \{A_1, A_2, \dots, A_n\}$ be a set of variables. We denote by $D_A = \{a_1, \dots, a_n\}$ the domain associated with the variable A . a_i denotes any instance of A . The universe of discourse is denoted by $\Omega = \times_{A_i \in \mathcal{V}} D_{A_i}$, which is the Cartesian product of all variable domains in \mathcal{V} . Each element $\omega \in \Omega$ is called an interpretation which represents a possible state of the world. It is denoted by $\omega = (a_1, \dots, a_n)$ or $\omega = (a_1 \wedge a_2 \wedge \dots \wedge a_n)$. Where $\{a_i \mid 1 \leq i \leq n\}$ are the instances of the variable A_i . ϕ, ψ denote propositional formulas (corresponding to events, i.e., subsets of Ω) constituted from the variables in \mathcal{V} .

2.1.1 Possibility distribution

The basic element in possibility theory is the notion of possibility distribution π which is a mapping from Ω to the scale $[0, 1]$. This distribution encodes available knowledge on real world: $\pi(\omega) = 1$ means that ω is completely possible and $\pi(\omega) = 0$ means that it is impossible to ω to be the real world. A possibility distribution π is said to be α -normalized, if its normalization degree $h(\pi)$ is equal to α , namely:

$$h(\pi) = \max \pi(\omega) = \alpha \quad (1)$$

If $\alpha = 1$, then π is said to be normalized.

Given a possibility distribution π on the universe of discourse Ω , two dual measures are defined for each event $\phi \subseteq \Omega$:

- *Possibility measure*: this measure evaluates to what extent ϕ is consistent with our knowledge. It is given by:

$$\Pi(\phi) = \{\pi(\omega) : \omega \in \phi\} \quad (2)$$

- *Necessity measure*: it is the dual of the possibility measure. The necessity measure evaluates at which level ϕ is certainly implied by our knowledge. It is given by:

$$N(\phi) = 1 - \Pi(\bar{\phi}) \quad (3)$$

2.1.2 Possibilistic conditioning

The possibilistic conditioning consists in the revision of our initial knowledge, encoded by a possibility distribution Π , after the arrival of a new certain information $\phi \subseteq \Omega$. The initial distribution Π is then replaced by another one, denoted $\Pi' = \Pi(\cdot | \phi)$. The two interpretations of the possibilistic scale (qualitative and quantitative) induce two definitions of possibilistic conditioning [4]. In this paper, we focus on min-based conditioning (qualitative one) defined by:

$$\Pi(\omega | \phi) = \begin{cases} 1 & \text{if } \pi(\omega) = \Pi(\phi) \text{ and } \omega \models \phi \\ \pi(\omega) & \text{if } \pi(\omega) < \Pi(\phi) \text{ and } \omega \models \phi \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

2.2 Min-based possibilistic network

2.2.1 Preliminaries

There are two ways of knowledge representation: a logical representation and a graphical representation. In this paper we are interested to the graphical representation. It is qualitative network. A possibilistic network is an adaptation of the probabilistic (Bayesian) network, in the sense where we use the same graphic structure which is the direct acyclic graph (DAG)

A min-based possibilistic network [10] over a set of variables V denoted by $\Pi G_{min} = (G, \pi_{min})$ is characterized by:

- A graphical component: which is represented by a Directed Acyclic Graph (DAG) where nodes correspond to variables and arcs represent dependence relations between variables.
- Numerical components: these components quantify different links in the DAG by using local possibility distributions for each node A in the context of its parents denoted by U_A . More precisely:
 - For every root node $A(U_A = \emptyset)$, uncertainty is represented by the a priori possibility degree $\pi(a)$, for each instance $a \in D_A$, such that $\max \pi(a) = 1$.
 - For the rest of the nodes $A(U_A \neq \emptyset)$, uncertainty is represented by the conditional possibility degree $\pi(a | U_A)$, for each instance $a \in D_A$, and $U_A \in D_A$, such that $\max \pi(a | U_A) = 1$, for any U_A .

The a priori and the conditional possibility degrees induce a unique joint possibility distribution defined by:

$$\pi_G(A_1, \dots, A_n) = \Pi_{min}(A_i | U_{Ai}) \quad (5)$$

2.2.2 Fusion of min-based possibilistic networks

Merging uncertain information [6] is important to exploit complementarities between sources. It provides thus a global and complete point of view. In this paper, we are interested in conjunctive mode which makes sense if all sources are considered as equally and fully reliable. One of the basic conjunctive operators is the minimum operation (min). Given two min-based possibilistic networks $\Pi G_{min} = (G, \pi_G)$ and $\Pi G'_{min} = (G', \pi_{G'})$, the result of merging ΠG and $\Pi G'$ is the possibilistic network $\Pi G_{\oplus} = (G_{\oplus}, \pi_{\oplus})$ [15], such that:

$$\forall \omega, \pi_{\oplus}(\omega) = \min(\pi_G(\omega), \pi_{G'}(\omega)) \quad (6)$$

The syntactic counterpart of the fusion of two possibility distributions, associated to two possibilistic networks, using the min operator is a new min-based possibilistic network whose definition depends on the union of the two initial ones. In [15], the authors propose two principal classes for merging min-based possibilistic networks:

- **Fusion of two possibilistic networks ΠG and $\Pi G'$ having the same network structure.** The resulting network ΠG_{\oplus} retains the same structure: $G_{\oplus} = G = G'$. The possibility degrees are computed as follows: for each variable A , $\pi_{\oplus}(A | U_A) = \min(\pi_G(A | U_A), \pi_{G'}(A | U_A))$.
- **Fusion of two possibilistic networks ΠG and $\Pi G'$ with different structures.** Two cases are distinguishable:
 - The union of graphs is acyclic. In this case, the union of the two graphs is retained as the result of the fusion. The set of its variables is the union of the sets of variables belonging to ΠG and $\Pi G'$. For each variable A , its parents are both ΠG and $\Pi G'$.
 - The union of graphs is cyclic. Further variables are added to eliminate cycles. The new conditional distributions of the new variables ensure equivalence between the new and the old variables.

For more details on the fusion of possibilistic networks see [15].

3. QUALITATIVE POSSIBILISTIC DECISION

In a problem of decision under uncertainty, knowledge of the decision-maker is generally not very informative. In other words, the agent does not know the real state of the world, but he knows only that this state belongs to a finite set of possible states. A decision system is defined by a finite set of states $S = \{s_1, s_2, \dots, s_n\}$, a finite set of consequences X , a set of decisions noted $D = \{d_1, d_2, \dots, d_m\}$, and a set of preferences among the consequences. Each decision $d_i: S \rightarrow X$ is a function that associates to every possible state of the world a consequence. The preferences among the consequences are encoded by the utility function $v: X \rightarrow U$ where U is a preferably ordinal scale.

The theory of possibility allows one to express the uncertain knowledge on different states of the world by using a possibility distribution. Indeed, it allows one to represent uncertain knowledge by distinguishing what is plausible to what is less plausible. It provides also a suitable mean to represent preferences on the consequences of decisions in order to distinguish the desirable consequences from the less desirable ones [9].

The uncertainty on the possible states of the world is represented by a normalized possibility distribution π that associates to a set of state variables a value in the interval $[0, 1]$. Likewise, the preferences of the agent are represented by a different possibility distribution μ that associates to a set of consequences a value in an ordinal scale U , represented by the interval $[0,1]$ [8]. We assume that the uncertainties and preferences are commensurable [5].

In the context of decision theory under uncertainty proposed by Savage, uncertainty of the agent is modeled by a probability distribution π on the set of possible states of the world and its preferences by a utility function μ with real values on the set X of the possible consequences of his/her actions.

In contrast, in the possibilistic framework, knowledge of the agent is modeled by a normalized possibilistic distribution π which is a function from states to a simply ordered scale L of plausibility: for a world ω , $\pi(\omega) \in L$: represents the degree of likelihood that ω is the real state of the world. If we consider that the information possessed by the agent on the decision problem is purely ordinal, it is reasonable to think that not only his/her knowledge can be expressed by a possibilistic distribution but also his/her preferences [3][12]. A distribution of possibilities can be then seen as a utility [12] function.

Let μ be the possibility distribution representing the agent's preferences. μ takes its values in a simply orderly scale in $[0, 1]$. As in Savage theory, an action is represented by a function d that associates to a world an element of X [12]. The utility of an action (decision) d in a state ω and whose consequence is $d(\omega) \in X$ can be evaluated by combining the possibility degrees $\pi(\omega)$ and the utilities $\mu(d(\omega))$ in an appropriate manner for all the possible states of world [12].

Two evaluation criteria have been proposed to achieve such combinations assuming some form of commensurability between the scales of plausibility and utility [2] [12]:

- **Pessimistic criterion (Minimax).** Criterion of a pessimistic decision maker: the chosen decision is that having the largest minimum utility :

$$U^*(d) = \min_{\omega \in \Omega} \max(\pi_{Kd}(\omega), \mu(\omega)) \quad (7)$$

- **Optimistic criterion (Maximin):** Criterion of an optimistic decision maker: the chosen decision is that having the largest maximum utility :

$$U^*(d) = \max_{\omega \in \Omega} \min(\pi_{Kd}(\omega), \mu(\omega)) \quad (8)$$

In this work, we are interested in the optimistic criterion for the calculation of the decision.

Example 1: Let us consider the problem of deciding whether we should or not take an umbrella, knowing that it would rain. The two min-based possibilistic networks representing knowledge and preferences of the agent are denoted ΠK_{min} and ΠP_{min} respectively. Before presenting the possibilistic graphs, let us first present the set of nodes used in the networks and their meanings.

- R: It's raining.
- W: The grass is wet.
- UM: Take the umbrella.
- C: Cloudy atmosphere.

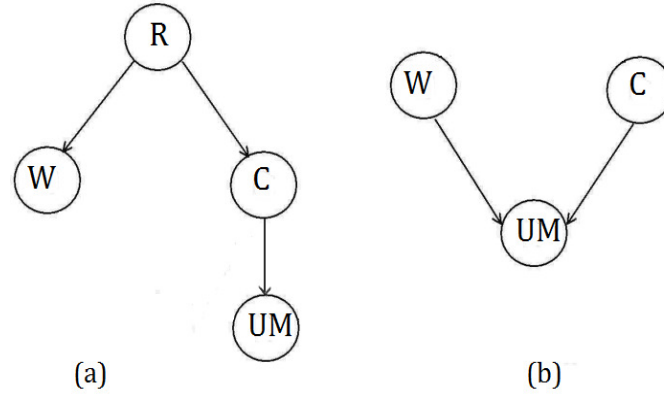


Figure1. The possibilistic networks of knowledge and preference of an agent

- **Agent's knowledge:** described by the min-based possibilistic network $\Pi K_{min} = (G_K, \pi_K)$, where the graphical component G_K is given by Figure 1 (a). It contains one possible states of the world R, one decision variable UM and two consequences {W, C}. The initial possibility distributions associated with ΠK_{min} are given by Tables 1 and 2. We suppose that the variables are binary.

R	$\pi_K(R)$
r1	0.9
r2	1.0

Table 1. Initial possibility distributions relative to ΠK_{min}

W	R	$\pi_K(W R)$	C	R	$\pi_K(C R)$	UM	C	$\pi_K(UM C)$
w1	r1	0.4	c1	r1	1.0	um1	c1	1.0
w1	r2	1.0	c1	r2	0.2	um1	c2	1.0
w2	r1	1.0	c2	r1	0.3	um2	c1	1.0
w2	r2	0.0	c2	r2	1.0	um2	c2	1.0
			bio1	of2	0.8			
			bio2	of1	0			
			bio2	of2	1			

Table 2. Initial possibility distributions relative to ΠK_{min}

- **Agent's preferences:** expressed by the min-based possibilistic network $\Pi P_{min} = (G_P, \mu)$, where the graphical component G_P is given by Figure 1 (b). It contains one decision variable UM and two consequences {W, C}. The initial possibility distributions associated with ΠP_{min} are given by Tables 3 and 4.

W	$\mu(W)$	UM	$\mu(UM)$
w1	1.0	um1	1.0
w2	1.0	um2	1.0

Table 3. Initial possibility distributions relative to ΠP_{min}

UM	W	R	μ (UM W R)	UM	W	R	μ (UM W R)
um1	w1	r1	1.0	um2	w1	r1	1.0
um1	w1	r2	1.0	um2	w1	r2	0.0
um1	w2	r1	1.0	um2	w2	r1	0.8
um1	w2	r2	1.0	um2	w2	r2	1.0

Table 4: Initial possibility distributions relative to ΠP_{min}

4. ON THE COMPUTATION OF OPTIMAL OPTIMISTIC DECISIONS BASED ON MIN-BASED FUSION

This section presents the computation of qualitative possibilistic decision which is regarded as a problem of merging data from two possibility distributions: the first represents agent's beliefs and the second represents the qualitative utility. Knowledge and preferences of the agent are both represented by two separated min-based possibilistic networks, namely $\Pi K_{min} = (G_K, \pi_K)$ and $\Pi P_{min} = (G_P, \mu)$, respectively. In what follows, we propose a directed method for computing optimal optimistic decisions based on the fusion of π_K and μ (or ΠK_{min} and ΠP_{min}). Each decision d induces a possibility distribution π_{Kd} defined as follows:

$$\pi_{Kd}(\omega) = \min(\pi_K(\omega), \pi_d(\omega)) \quad (9)$$

We recall that making a decision comes down to choosing a subset d of the decision set D which maximizes the optimistic qualitative utility by:

$$U^*(d) = \max_{\omega \in \Omega} \min(\pi_{Kd}(\omega), \mu(\omega)) \quad (10)$$

Where,

$$\pi_d(\omega) = \begin{cases} 1 & \text{if } \omega \models \phi_i \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

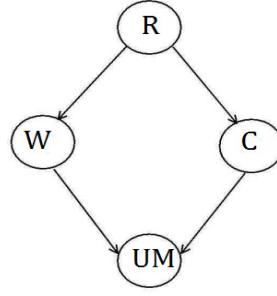
Using equation (11), the optimistic utility decision $U^*(d)$ becomes:

$$U^*(d) = \max_{\omega \in \Omega} \min(\min(\pi_{Kd}(\omega), \mu(\omega)), \pi_d(\omega)) \quad (12)$$

Using technical merging of two min-based possibilistic networks, this Equation (12) down to:

$$U^*(d) = \max_{\omega \in \Omega} \min(\pi_{\oplus}, \pi_d(\omega)) \quad (13)$$

Example 2: The two DAGs (G_K and G_P) given in Example 1, Figure 1 have a different structures. Their union is acycles, the result of merging ΠK_{min} and ΠP_{min} is the min-based possibilistic network $\Pi G_{\oplus} = (G_{\oplus}, \pi_{\oplus})$ where G_{\oplus} , is given in Figure 2.

Figure2. The DAG G_\oplus

The initial possibility distributions are given by Tables 5 and 6.

R	$\pi_K(R)$	W	R	$\pi_K(W R)$	C	R	$\pi_K(C R)$
r1	0.9	w1	r1	0.4	c1	r1	1.0
r2	1.0	w1	r2	1.0	c1	r2	0.2
		w2	r1	1.0	c2	r1	0.3
		w2	r2	0.0	c2	r2	1.0

Table 5. Initial possibility distributions relative to ΠG_\oplus

UM	W	R	$\mu(UM W R)$	UM	W	R	$\mu(UM W R)$
um1	w1	r1	1.0	um2	w1	r1	1.0
um1	w1	r2	0.0	um2	w1	r2	1.0
um1	w2	r1	0.8	um2	w2	r1	1.0
um1	w2	r2	1.0	um2	w2	r2	1.0
				bio2	of1		0
				bio2	of2		1

Table 6. Initial possibility distributions relative to ΠG_\oplus .

4.1 Computing optimal decisions using moral graph

Computing the optimistic optimal decisions amounts to find the normalization degree of the moral graph resulting from the merging of the two possibilistic networks codifying knowledge of the agent and its preferences respectively without going through the junction tree. Note that the construction of the moral graph is done only once and has a polynomial complexity. However, the stabilization procedure, multiple stabilization procedure and initialization (see below) (which are all three polynomials) are repeated for each decision d^* .

1) Building the moral graph.

The construction of the possibilistic moral graph, noted \mathcal{MG} , from the initial graph is done as follows:

- For each variable A_i , form a cluster $C_i = A_i \cup U_A$
- For each edge connecting two nodes A_i and A_j : form an undirected edge in the moral graph between the cluster C_i and the cluster C_j labeled with a separator S_{ij} corresponding to their intersection.

2) Initialization.

For a given decision d , once the moral graph is built, we proceed to its quantification by taking into account the decision d as follows:

- For each cluster C_i , (resp. S_{ij}) $\pi_{C_i}^I \leftarrow 1$. (resp. $S_{ij} \leftarrow 1$)
- For each variable A_i , choose a cluster C_i containing $A_i \cup U_A$
 $\pi_{C_i} \leftarrow \min(\pi_{C_i}, \pi_{\oplus}(A_i|U_A))$.
- Encode the evidence $D = d_i$ as likelihood $\Lambda_D(d)$:
-

$$\Lambda_D(d): \begin{cases} 1 & D \text{ is instantiated as } d \\ 0 & D \text{ is instantiated as a value } d' \neq d \end{cases} \quad (14)$$

- Identify a cluster C_i containing D : $\pi_{C_i}^I \leftarrow \min(\pi_{C_i}^I, \Lambda_D)$.

Proposition 1: Let $\Pi K_{min} = (G_K, \pi_K)$, be a min-based possibilistic network representing agent's beliefs and $\Pi P_{min} = (G_P, \mu)$ be a min-based possibilistic network representing agent's preferences. Let $\Pi G_{\oplus} = (G_{\oplus}, \pi_{\oplus})$ be the result of merging ΠK_{min} and ΠP_{min} using the min operator. Let \mathcal{MG} , be the moral graph corresponding to ΠG_{\oplus} generated using the above initialization procedure. Then,

$$U^*(d) = \max_{\omega \in \Omega} (\mathcal{MG}(\omega)) \quad (15)$$

Where $U^*(d)$ is given in Equation 13.

3) Simple Stability Procedure.

The simple stabilization procedure ensures that the potential of each clique is in agreement with that of its neighbors. This procedure is applied through a mechanism of passage of messages between different cliques. Indeed, each separator collects information from its corresponding cliques in order to distribute it then to each of them in order to update them.

The potentials of any adjacent clusters C_i and C_j (with separator S_{ij}) are updated as follows:

- Collect evidence (Update separator) :

$$S_{ij}^{t+1} \leftarrow \min(\max_{C_i/S_{ij}} \pi_{C_i}^t, \max_{C_j/S_{ij}} \pi_{C_j}^t) \quad (16)$$

- Distribute evidence (Update clusters) :

$$\pi_{C_i}^{t+1} \leftarrow \min(\pi_{C_i}^t, S_{ij}^{t+1}) \quad (17)$$

$$\pi_{C_j}^{t+1} \leftarrow \min(\pi_{C_j}^t, S_{ij}^{t+1}) \quad (18)$$

This procedure is defined as follows:

Definition 1: Let C_i and C_j be two adjacent clusters in a moral graph \mathcal{MG} , and let S_{ij} be their separator. The separator S_{ij} is said to be stable if:

$$\max_{C_i/S_{ij}} \pi_{C_i}^I = \max_{C_j/S_{ij}} \pi_{C_j}^I \quad (19)$$

Where $\max_{C_i/S_{ij}} \pi_{C_i}$ is the marginal distribution of S_{ij} defined from $\pi_{C_i}^I$ (resp. $\pi_{C_i}^I$).

A moral graph \mathcal{MG} is stable if all its separators are stable.

Proposition 2: Let \mathcal{MG} be a stabilized moral graph, let $\pi_{\mathcal{MG}}$ be the joint distribution encoded by MG after the initialization procedure. Then,

$$\forall C_i, \max \pi_{\mathcal{MG}}^S \geq \alpha \quad (20)$$

Where, α is the maximum value in all clusters.

4) Multiple Stability Procedure.

[1] Proved that the simple stabilization procedure does not always guarantee accurate marginal. One needs to stabilize each clique with respect to all of its adjacent cliques but this can turn out to be very costly in terms of calculation if the number of cliques is important. For that, [1] has proposed to follow several steps in stabilizing the possibilistic moral graph over subsets of its adjacent cliques. Authors of [1] have proposed several progressive stabilization procedures based on n parents, n children, n parents children and n neighbors by varying the value of n from 2 up to the cardinality of the considered subset. To illustrate the multiple stabilization procedure, we consider the case of two parent's stabilization. The principle of this procedure is to ensure for each clique, with at least two parents, its stabilization over each pair of parents. Once stability has been reached, the calculation of qualitative utility over a decision d will be obtained as follows:

Proposition 3. Let $\Pi K_{min} = (G_K, \pi_K)$ be a min-based possibilistic network representing agent's beliefs and $\Pi P_{min} = (G_P, \mu)$ a min-based possibilistic network representing agent's preferences. ΠG_{\oplus} is the result of merging of ΠK_{min} and ΠP_{min} by using the min operator. Let \mathcal{MG} be the moral graph of ΠG_{\oplus} . The computation of optimistic decisions returns to calculate the normalization degree of MG:

$$U^*(d) = \max_{C_i} (\pi_{C_i}) \quad (21)$$

4.2 Algorithm

The computation of the optimal optimistic decisions is obtained using the following algorithm.

Algorithm : Computation of optimal optimistic decision

Data: $\Pi K_{min} = (G_K, \pi_K)$: Knowledge possibilistic network

$\Pi P_{min} = (G_P, \mu)$: Preferences possibilistic network

$D = \{D_1, \dots, D_n\}$: Set of decisions,

Result: decisions μ^*

Begin:

$\Pi G_{\oplus} = (G_{\oplus}, \pi_{\oplus})$ Fusion of ΠK_{min} and ΠP_{min}

$\mathcal{MG} = \text{MoralGraph}(\Pi G_{\oplus})$;

$\mu^* \leftarrow 0$

$i \leftarrow 1$

Norm $\leftarrow 0$

Decision $\leftarrow \emptyset$

For $i = 1 \dots n$ do

$INIT(\mathcal{MG}, D_i)$

Norm $\leftarrow SSP(\mathcal{MG}, D_i)$

Norm $\leftarrow MSP(\mathcal{MG}, D_i)$

 IF **Norm** $> \mu^*$

```

Then Decision  $\leftarrow D_i$ 
     $\mu^* \leftarrow \text{Norm}$ 
Else
    IF Norm =  $\mu^*$ 
        Then Decision  $\leftarrow \text{Decision} \cup D_i$ 
    EndIf
EndIf
Endfor
Return <Decision>
End

```

Example 3: Let us continue with Example 2. We need to compute the optimal optimistic decision $UM = \{um1; um2\}$. First, we start by constructing the moral graph (see Figure3) associated with the graph G_{\oplus} (Figure 2) representing the fusion of ΠK_{min} and ΠP_{min} . The resulted moral graph contains four cluster $C1 = \{R\}$, $C2 = \{R, W\}$, $C3 = \{R, C\}$ and $C4 = \{R, W, UM\}$ and their separator $S12 = \{R\}$, $S13 = \{R\}$, $S23 = \{R\}$, $S24 = \{w\}$ and $S34 = \{c\}$.

Then, for each decision value in $UM = \{um1; um2\}$, we must run the algorithm in order to compute the normalization degree associated with the moral graph.

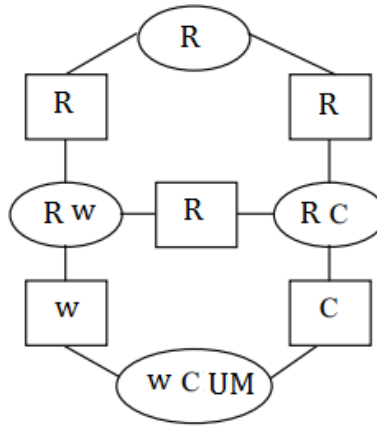


Figure.3. Moral Graph MG of th DAG in Figure 2

ω	π_{MG}	ω	π_{MG}
r1 w1 c1 um1	0.3	r2 w1 c1 um1	0.0
r1 w1 c1 um2	0.3	r2 w1 c1 um2	0.0
r1 w1 c2 um1	0.0	r2 w1 c2 um1	0.0
r1 w1 c2 um2	1.0	r2 w1 c2 um2	0.0
r1 w2 c1 um1	0.3	r2 w2 c1 um1	0.8
r1 w2 c1 um2	0.3	r2 w2 c1 um2	0.9
r1 w2 c2 um1	0.4	r2 w2 c2 um1	0.2
r1 w2 c2 um2	0.4	r2 w2 c2 um2	0.2

Table 7. Joint distribution π_{MG}

Step 1: $UM = um1$

In this case, the fact that $UM = um1$ is encoded interns of likelihood a s follows:

$$\Lambda_{UM}(um1): \begin{cases} 1 & UM \text{ is instanciated as } um1 \\ 0 & UM \text{ is instanciated as } um2 \end{cases}$$

The table 8 represents the joint distribution encoded by \mathcal{MG} after the initialization procedure us.

ω	$\pi_{\mathcal{MG}}$	ω	$\pi_{\mathcal{MG}}$
r1 w1 c1 um1	0.3	r2 w1 c1 um1	0.0
r1 w1 c1 um2	0.0	r2 w1 c1 um2	0.0
r1 w1 c2 um1	0.0	r2 w1 c2 um1	0.0
r1 w1 c2 um2	0.0	r2 w1 c2 um2	0.0
r1 w2 c1 um1	0.3	r2 w2 c1 um1	0.8
r1 w2 c1 um2	0.0	r2 w2 c1 um2	0.0
r1 w2 c2 um1	0.4	r2 w2 c2 um1	0.2
r1 w2 c2 um2	0.0	r2 w2 c2 um2	0.0

Table 8. Joint distributions $\pi_{\mathcal{MG}}$ after the initialization procedure

Once the moral graph is quantified, then the simple stabilization procedure allows us to compute the normalization degree of the moral graph which corresponds to the normalization degree of any cluster. Using this procedure, we obtain:

R	π_{C1}	R W	π_{C2}	R C	π_{C3}	W C UM	π_{C4}	W C UM	π_{C4}
r1	0.9	r1 w1	0.4	r1 w1	0.9	w1 c1 um1	0.9	w2 c1 um1	0.8
r2	0.9	r1 w2	0.9	r1 w2	0.3	w1 c1 um2	0.0	w2 c1 um2	0.0
		r2 w1	0.9	r2 w1	0.2	w1 c2 um1	0.0	w2 c2 um1	0.9
		r2 w2	0.0	r2 w2	0.9	w1 c2 um2	0.0	w2 c2 um2	0.0

Table 9. Normalized potentials with UM=um1

$$\max_{\pi_{C1}} = \max_{\pi_{C2}} = \max_{\pi_{C3}} = \max_{\pi_{C4}} = 0.9$$

From the table 8 we can check that: $h(\pi_{\mathcal{MG}}) = 0.8 \neq 0.9$, which means that the moral graph is not consistent. So we must to re-stabilize the moral graph using the multiple. stability procedure. Using this procedure, we obtain:

$$\max_{\pi_{C1}} = \max_{\pi_{C2}} = \max_{\pi_{C3}} = \max_{\pi_{C4}} = 0.8$$

The normalization degree of the moral graph is: $U^*(um1) = 0.8$

Step 2: UM= um2

We repeat the same procedure described in the previous step, with:

$$\Lambda_{UM}(um2): \begin{cases} 1 & UM \text{ is instanciated as } um2 \\ 0 & UM \text{ is instanciated as } um1 \end{cases}$$

Then, we get:

$$U^*(um2) = \max_{\pi_{C1}} = \max_{\pi_{C2}} = \max_{\pi_{C3}} = \max_{\pi_{C4}} = 1.0$$

Thus, we can conclude that the optimal decision is **UM = um2** with the maximal qualitative utility which equals 1.0

5. CONCLUSION

In this paper, we proposed a new approximate approach for the computation of the qualitative possibilistic optimal optimistic decision in a graphical context. Our approach first merges possibilistic networks associated with uncertain knowledge and possibilistic networks associated with agent's preferences. We then showed that computing optimistic decisions comes down to computing a normalization degree of the moral graph associated to the result graph of merging agent's beliefs and preferences networks.

This approach allows one to avoid the transformation of the initial graph into a junction tree which is known as a difficult problem. This approach is interesting when accurate approaches fail, i.e., when the generation of the local possibility distributions by the standard algorithm is impossible or takes a too long response time. In such case, our approach provides answers that are very close to the exact marginal distributions.

REFERENCES

- [1] N. Ben Amour. Qualitative Possibilistic Graphical models From Independence to propagation algorithm. PhD thesis, Université d'Artois, 2002.
- [2] D. Dubois, J. Lang, and H. Prade. Possibilistic logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, (D. Gabbay et al., eds, 3, Oxford University Press :pages 439-513, 1994.
- [3] D. Dubois, D. Le Berre, H. Prade, and R. Sabbadin. Using possibilistic logic for modeling qualitative decision: Atms based algorithms. In *Fundamenta Informaticae*, 37 :1-30, 1999.
- [4] D. Dubois and H. Prade. (with the collaboration of H. Farreny, R. Martin-Clouaire and C. Testemale). *Possibility Theory - An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York., 1988
- [5] D. Dubois and H. Prade. Possibility theory and data fusion in poorly informed environments. In *Control Engineering Practice*, volume 2(5), pages 811-823, 1994.
- [6] D. Dubois and H. Prade. Possibility theory as a basis for qualitative decision theory. In *14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, Montréal, pages 1924-1930, 1995.
- [7] D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. In *Handbook of Defeasible Reasoning and Uncertainty Management Systems*. (D. Gabbay, Ph. Smets, eds.), Vol. 1: *Quantified Representations of Uncertainty and Imprecision*, (Ph. Smets, ed.) Kluwer, Dordrecht: 169-226, 1998.
- [8] F. Haned-Khellaf S. Benferhat and I. Zeddigha. Computing optimal optimistic decisions using min-based possibilistic networks. In *North American Fuzzy Information Processing Society, Berkeley NAFIPS 2012, JUIN 2012*.
- [9] L. Garcia and R. Sabbadin. Diagrammes d'influence possibilistes. *Revue d'Intelligence Artificielle*, 21(4): 521-554, 2007.
- [10] J. Gebhardt and R. Kruse. Background and perspectives of possibilistic graphical models. In *4th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECS-QARU'97)*, LNAI 2143, pages 108-121, 1997.
- [11] A. Mokhtari S. Benferhat, F. Haned-Khellaf and I. Zeddigha. A possibilistic graphical model for handling decision problems under uncertainty. In *8th conference of the European Society for Fuzzy Logic and Technology, EUSFLAT-2013*, September 2013.
- [12] R. Sabbadin. Une approche logique de la résolution de problèmes de décision sous incertitude basée sur les atms. In *Actes du 11ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'98)*, Clermont-Ferrand, pages 391- 400, 20-22 janvier 1998.
- [13] P.P Shenoy. Valuation based systems: A framework for managing uncertainty in expert systems. In *Fuzzy Logic for the Management of Uncertainty*, pages 83-104. L. A. Zadeh and J. Kacprzyk, Eds. John Wiley and Sons, New York, NY, 1992.

- [14] P.P Shenoy. A comparison of graphical techniques for decision analysis. In European Journal of Operational Research, volume 78, pages 1-21, 1994.
- [15] F.Titoune. Fusion de réseaux causaux possibilistes. PhD thesis, Université d'Artois, 2009.
- [16] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1 :3-28, 1978.

REAL TIME CLUSTERING OF TIME SERIES USING TRIANGULAR POTENTIALS

Aldo Pacchiano¹ and Oliver J. Williams²

¹Massachusetts Institute of Technology
aldopacchiano@gmail.com

²Markham Rae LLP, London, UK. (The views and opinions expressed herein are those of the author and do not necessarily reflect the views of Markham Rae LLP.)
oliver.williams@markhamrae.com

ABSTRACT

Motivated by the problem of computing investment portfolio weightings we investigate various methods of clustering as alternatives to traditional mean-variance approaches. Such methods can have significant benefits from a practical point of view since they remove the need to invert a sample covariance matrix, which can suffer from estimation error and will almost certainly be non-stationary. The general idea is to find groups of assets which share similar return characteristics over time and treat each group as a single composite asset. We then apply inverse volatility weightings to these new composite assets. In the course of our investigation we devise a method of clustering based on triangular potentials and we present associated theoretical results as well as various examples based on synthetic data.

KEYWORDS

Clustering, Expected Utility, Graphical Models, k-Clique Problem

1. INTRODUCTION

A common problem in finance is the question of how best to construct a diversified portfolio of investments. This problem is ubiquitous in fund management, banking and insurance and has led to an extensive evolving literature, both theoretical and empirical. From an informal mathematical perspective the central challenge is to devise a method for determining weightings for a set of random variables such that *ex post* realisations of the weighted sum optimise some objective function *on average*. The objective function most typically used in financial economics is a concave *utility* function, hence from an *ex ante* perspective the portfolio construction problem is a matter of optimising so-called *expected utility*. Koller and Friedman provide a detailed discussion of utility functions and decision theory in the general machine learning context [1].

The theoretical literature analyses many alternative weighting strategies which can be distinguished based on such criteria as: (a) the investor's time horizon (e.g. does utility depend on realisations on a single time horizon in a 'one-shot' scenario or does uncertainty resolve over multiple time periods, affording the investor opportunities to alter portfolio composition dynamically?), (b) the nature of the information available to investors regarding the distribution of future returns (this may be extremely limited or highly-structured for mathematical expediency), and (c) the investor's

particular utility function (where, for instance, it can be shown that curvature can be interpreted as representing the investor's risk-preferences [2]).

One of the most prominent theoretical results is the concept of *mean-variance efficiency* which has its roots in the work of Markowitz [3]: the idea is that in a one period model (under certain restrictive assumptions) if investors seek to maximize return and minimise portfolio variance, the optimal *ex ante* weighting vector w is given by

$$w = \frac{1}{\lambda} \Omega^{-1} (\mu - r\mathbf{1}) \quad (1)$$

where Ω is the covariance matrix of future returns, μ is the mean vector of expected returns, λ is a risk-aversion parameter and r is the risk-free rate of return [4]. A key aspect of this formula is the dependency on the inverse of the covariance matrix which is never known with certainty and will in practice be a forecast in its own right (and the same will be true for μ and quite possibly r). When deploying this formula in real-world investment, practitioners are divided over how to account for parameter uncertainty, with a number of alternative approaches in common usage (including ignoring uncertainty entirely).

Unfortunately it is widely recognised that the exact weightings in (1) have a sensitivity to covariance assumptions which is unacceptably high; in other words small changes in covariance assumptions can lead to large changes in prescribed weightings. Further significant concerns are raised by the fact that long time series are required to generate acceptable estimates for a large covariance matrix but financial returns series are notoriously non-stationary – it is therefore easy for an analyst to fall into the trap of thinking that they are applying prudent statistical methods when in reality their input data may be stale or entirely inappropriate. The forecasting of expected returns is also regarded as an exceptionally difficult task.

In these circumstances one strand of literature considers simpler weighting schemes which are predicated on relatively few assumptions; one prominent example, popular with practitioners, is the self-explanatory *equally-weighted* (or $\frac{1}{n}$) approach [5]. This method requires no explicit forecasts of correlation or returns and it can be shown that this is equivalent to mean-variance methods if the correlation between all possible pairs of investments is equal, along with all means and variances. Although this may be far from the truth it may be more innocuous to assume this than to suffer potentially negative effects of erroneous statistical forecasts and there is a body of empirical literature which demonstrates the efficiency of the approach [6]. Refinements to the basic method can include weighting each asset by the inverse of the forecast standard deviation of its returns (known as *volatility*) which allows some heterogeneity to be incorporated.

Nevertheless it is intuitively obvious that such a simple method presents potential dangers of its own, and is particularly inappropriate when the universe of alternative investments contains subgroups of two or more investments which are highly correlated with each other. Suppose, for instance, a portfolio of investments in world stock market indices which includes several alternative indices for the United States (e.g. Dow Jones, S&P 500, Russell 2000) but only single indices for other markets (e.g. the CAC-40 for France, FTSE-100 for UK, etc.). In this setting the $\frac{1}{n}$ approach may (arguably) significantly overweight US equities in comparison to each foreign market and in general regional weightings will be more dependent on the cardinality of available indices than any economic properties of the markets. In a systematic investment process it is clearly

impractical to have analysts manually sift through investments to ensure an appropriate ‘balance’ (which defeats the object of a weighting algorithm) and indeed potential diversification benefits argue in favour of including a broad range of investments anyway.

The contribution of this paper is to explore potential weighting methods based on clustering, such that highly ‘similar’ investments can be identified, grouped together and treated (for weighting purposes) as if they are a single ‘composite’ investment. By contrast, investments which exhibit relatively little similarity to each other are treated individually in their own right. Our focus here is on a process for identifying clusters rather than evaluation of *ex post* investment performance, which we leave for a separate analysis, and in fact we draw attention to the applicability of our methods to fields beyond finance where clustering may be required, e.g. well-known problems in biology, medicine and computer science. We also present an intriguing theoretical result arising from our work, which emphasises limitations of certain clustering techniques and may help to guide other researchers in their search for suitable methods.

The paper is organised as follows: in Section 2 we formally specify the problem at hand, in Section 3 we demonstrate spectral clustering as a preliminary benchmark approach and in Section 4 we explore an alternative method based on a graphical model where we propose a specific estimation technique involving *triangular potentials* and provide illustrative examples. Section 5 briefly considers extension to a more dynamic setting (via a Hidden Markov Model) and Section 6 concludes.

2. PROBLEM SPECIFICATION

Definition 1 Let $n \in \mathbb{N}$, define $[n] = \{1, \Lambda, n\}$ the set of natural numbers from 1 to n .

Let $\{t_1\}, \Lambda, \{t_n\}$ be n time series, where $t_i = \{t_{i_1}, \Lambda, t_{i_m}\}$ for $m, n \in \mathbb{N}$.

Definition 2 *Clustering.*

A clustering of $\{t_1\}, \Lambda, \{t_n\}$ is an equivalence relation \sim over $[n] = \{1, \Lambda, n\}$ such that:

1. Reflexivity: If $i \sim j$ then $j \sim i$.
2. Transitivity: If $i \sim j$ and $j \sim k$ then $k \sim i$.

Definition 3 *Time dependent clustering.*

We say $i \overset{k}{\sim} j$ if i and j are clustered at time k .

Our aim is to find a sequence $\{\sim\}_{k=1}^m$, i.e. we allow the nature of the clustering relation to evolve over time.

We denote the *distance* between series at time k as $d^k(t_i, t_j)$ for all i, j and the similarity at time k defined as $s^k(t_i, t_j)$. The functions d^k, s^k are specified by the user of the algorithm and may be chosen based on prior domain-specific knowledge, or perhaps by a more systematic process of searching across alternative specifications guided by out-of-sample performance.

Definition 4 *Distance Matrix.*

Given a family of time-dependent distance functions $\{d^k(\cdot, \cdot)\}_{k=1}^m$, we define a family of distance matrices as $D_{i,j}^k = \{d^k(t_i, t_j)\}$.

Definition 5 *Similarity Matrix.*

Given a family of time-dependent similarity functions $\{s^k(\cdot, \cdot)\}_{k=1}^m$, we define a family of similarity matrices as $S_{i,j}^k = \{s^k(t_i, t_j)\}$.

Definition 6 *Similarization function.*

We say $z: \mathbb{R}_+ \rightarrow [0, 1]$ is a similarization function if for any distance function $d: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}_+$, $z \circ d$ is a valid similarity function.

In what follows we restrict our attention to reflexive and non-negative distance and similarity functions and thus to symmetric similarity and distance matrices. We will also use the variable n to represent the number of data points observed at each time step when the clustering algorithm will be applied.

3. SPECTRAL CLUSTERING

Here we introduce the Spectral Clustering algorithm, which is suitable for data where the cluster structure does not change over time. Later in the paper we will compare the performance of our proposed approach with this benchmark method.

Definition 7 *The Laplacian matrix L of a similarity matrix S is defined as follows:*

$$L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

where $D_{ii} = \sum_j S_{ij}$.

The most basic spectral clustering algorithm for bipartition of data is the Shi Malik bipartition algorithm which we describe below.

3.1. Shi Malik algorithm

Given n items and a similarity matrix $S \in \mathbb{R}^{n \times n}$, the Shi Malik algorithm bipartitions the data into two sets (B_1, B_2) with $B_1 \cup B_2 = [n]$ and $B_1 \cap B_2 = \emptyset$ based on the eigenvector v corresponding to the second smallest eigenvalue of the normalized Laplacian matrix L of S .

Algorithm 1 *The Shi Malik bipartition algorithm:*

1. Compute the Laplacian from a similarity matrix.
2. Compute the second smallest eigenvalue and its corresponding eigenvector v .
3. Compute the median m of its corresponding eigenvector.
4. All points whose component in v is greater than m are allocated to B_1 , the remaining points are allocated to B_2 .

Unfortunately the Shi Malik algorithm is not a dynamic procedure, i.e. it is not intended to identify an underlying cluster structure which is time-varying. However various clustering approaches are available which specifically seek to address this and we outline one such approach next.

3.2. A generalized spectral clustering approach

The following algorithm is an extension of the Shi Malik algorithm that can handle two or more clusters. It can be found at [7]. Given n items and a similarity matrix $S \in \mathbb{R}^{n \times n}$ the goal of Dynamic Spectral Clustering is to find a clustering \sim of $[n]$.

Algorithm 2 *Dynamic Spectral Clustering*

1. Compute the Laplacian of the similarity matrix.
2. Compute the Laplacian's eigenvalues and eigenvectors
3. Let c be a desired number of clusters.
4. Find the eigenvectors of the corresponding eigenvalues found on the previous step. Let the corresponding $n \times c$ matrix be called V .
5. Rotate V , by multiplying it with an appropriate rotation matrix R so each of the corresponding rows of $Z = VR$ have (ideally) only one nonzero entry. In reality the resulting matrix we will use the largest (in absolute value) entry of the matrix. R is a rotation matrix in $\mathbb{R}^{c \times c}$.
6. The cluster to which point i is assigned is $\arg \max_{j \in \{1, \Lambda, c\}} |Z_{i,j}|$.

In order to find an appropriate rotation matrix R , there is a theorem that guarantees that any rotation matrix $R \in \mathbb{R}^{c \times c}$ can be written as a product $G_1 \cdot \Lambda \cdot G_k$ where $k = \frac{c(c-1)}{2}$ and each G_i equals a Givens rotation matrix.

Givens rotation matrices $G(i, j, \theta)$ are parameterized as follows:

$$G(i, j, \theta) = \begin{bmatrix} 1 & \Lambda & 0 & \Lambda & 0 & \Lambda & 0 \\ M & O & M & & M & & M \\ 0 & \Lambda & \cos \theta & \Lambda & -\sin \theta & \Lambda & 0 \\ M & & M & O & M & & M \\ 0 & \Lambda & \sin \theta & \Lambda & \cos \theta & \Lambda & 0 \\ M & & M & & M & O & M \\ 0 & \Lambda & 0 & \Lambda & 0 & \Lambda & 1 \end{bmatrix}$$

Hence for each G_i there is an associated angle θ_i and we represent these k angles by the vector $\Theta \in \mathbb{R}^{c(c-1)/2}$. In order to find the optimal Θ for a given number of clusters c , we use gradient descent on the following objective function:

$$\min_{\Theta} J = \sum_{i=1}^n \sum_{j=1}^c \left(\frac{Z_{ij}}{M_i} \right)^2$$

subject to the constraint

$$Z_{n \times c} = V_{n \times c} R(\Theta)_{c \times c}.$$

Following [7] we set $M_i = \max_j |Z_{ij}|$.

As suggested by [7], the optimal number of clusters can be obtained by choosing the value of c that maximizes a scoring function given by

$$q(c, n) = 1 - \left(\frac{J}{n} - 1 \right).$$

3.2.1. A dynamic clustering algorithm

Given a family of time dependent similarity functions $\{s^k(\cdot, \cdot)\}_{k=1}^m$ defining a family of similarity matrices $S_{i,j}^k = \{s^k(t_i, t_j)\}$, an optimal time-varying clustering structure \sim^k can be estimated by applying Algorithm 2 at time k using input similarity matrix $S_{i,j}^k$. Hence for time series data we propose the following algorithm:

Algorithm 3 Let $\{t_1\}, \Lambda, \{t_n\}$ be n time series. Where $t_i = \{t_{i_1}, \Lambda, t_{i_m}\}$ for $m, n \in \mathbb{N}$. Let $w \in \mathbb{N}$ be a window parameter, $d(\cdot, \cdot) : \mathbb{R}^w \times \mathbb{R}^w \rightarrow \mathbb{R}_+$ be a distance function and $z : \mathbb{R}_+ \rightarrow [0, 1]$ be a similarization function.

1. Let D_{ij}^m be the distance matrix having

$$D_{i,j}^m = d([t_{i_{m-w+1}}, \Lambda, t_{i_m}], [t_{j_{m-w+1}}, \Lambda, t_{j_m}])$$
 for every pair $i, j \in [n]$.
2. Let $S_{i,j}^m$ be the similarity matrix having $S_{i,j}^m = z(D_{i,j}^m)$ for every pair $i, j \in [n]$.
3. Let \sim^m be the clustering resulting from running Algorithm 2 with input similarity matrix $S_{i,j}^m$.
4. Output clustering \sim^m .

Extensions of this approach include considering a geometric decay factor in the distance computation, alternative distance functions and different similarization functions. We tried various combinations as shown in Table 1 but found no significant improvement on the stability of the resulting clusters.

We did not consider a scenario where the distance or similarization functions change through time although there may be certain applications where this might be appropriate.

Table 1: Alternative distance and similarity functions. The second similarity function is a generalization of the first.

Distances	L^1 norm	L^2 norm
Similarities	$\exp\left(-\left\ \frac{x_1}{\ x_1\ } - \frac{x_2}{\ x_2\ }\right\ \right)$	$\exp\left(-c \cdot \left\ \frac{x_1}{\ x_1\ } - \frac{x_2}{\ x_2\ }\right\ \right)$ set to zero when it achieves values less than λ for different combinations of c, λ

3.3. Overview

We present the performance of this algorithm in Figure 1. Some of the observed characteristics of this method are the following:

- The resulting clustering values are notably sensitive to the similarity function used in the model.
 - The clustering structure estimated by this method tends to be relatively unstable over time.
- Although in some applications this may be plausible, in the context of financial time series we have a strong prior belief that clusters typically arise due to common factors relating to economic fundamentals (e.g. similar commodities, currency pairs belonging to close trading partners, etc.) which would tend to change very slowly relative to the frequency of market data.

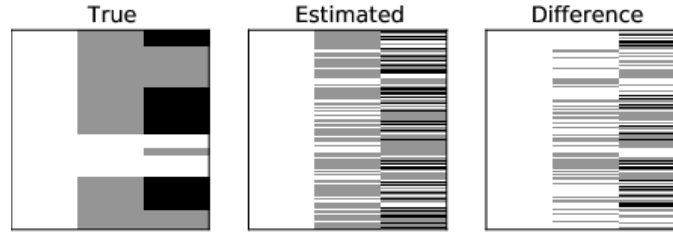


Figure 1: Performance of the spectral clustering algorithm on 5,000 periods of synthetic data with $n = 3$: at each time step we generate random standard normal variates which are common to each cluster, then for each of the 3 returns we add independent Gaussian noise with a relatively small variance. The members of each cluster therefore have a large portion of randomness in common, but each observation also includes its own independent noise. The cluster structure is randomly changed over time and represented by coloured bars in each row, i.e. all columns with the same colour belong in the same cluster.

4. GRAPHICAL MODEL APPROACH

Instead of representing clusterings as a binary matrix $Z(i, j)$ such that $Z(i, j) = 1$ if $i \in$ cluster j as the authors of [8] do, we approach the problem in a different way. Consider a symmetric ($C_{i,j} = C_{j,i}$) family of Bernoulli random variables, $\mathbf{C} = \{C_{i,j}\}_{i,j \in [n]}$ such that:

$$C_{i,j} = 1 \text{ if } i, j \text{ are in the same cluster, or } 0 \text{ otherwise.}$$

We wish to learn a distribution over the ensemble $\mathbf{C} = \{C_{i,j}\}$. The model we will use in this paper is the following

$$\mathbf{C} \rightarrow S$$

where S is a similarity matrix; in other words, we consider that the observed similarity between a pair of points will come from one of two distributions, depending on whether or not the two points belong to the same cluster.

In what follows it will be useful to think of the matrix $\{C_{i,j}\}$ as an adjacency matrix. The resulting graph $G = (V, E)$ where $V = [n]$ and $E = \{(i, j) \mid C_{i,j} = 1\}$, has an edge between every two nodes that are in the same cluster. Learning a distribution over $\{C_{i,j}\}$ can be thought of as learning a distribution over the set of undirected graphs (V, E) with $V = [n]$.

The goal of this section is to compute the following posterior:

$$P(\mathbf{C} | S).$$

The algorithms we present here output $\arg \max_{\mathbf{C} \in \mathcal{C}} P(\mathbf{C} | S)$, the MAP estimator for the posterior. A short algebraic manipulation (Bayes Theorem) yields:

$$P(\mathbf{C} | S) = \frac{P(S | \mathbf{C}) \cdot P(\mathbf{C})}{P(S)}.$$

Since S is fixed:

$$\arg \max_{\mathbf{C} \in \mathcal{C}} P(\mathbf{C} | S) = \arg \max_{\mathbf{C} \in \mathcal{C}} P(S | \mathbf{C}) P(\mathbf{C}).$$

In the following two sections we present different models for inference on the ensemble \mathbf{C} , their performance and their relationship to clusterings.

The training data will be:

1. A set of similarity matrices $\{S_{i,j}^k\}_{i=1}^m$.
2. The set of corresponding clusterings $\{\sim\}^k$ produced via a clustering algorithm such as the ones described earlier in Section 3.

4.1. Exponential model

As a starting point we propose the following model for the ensemble \mathbf{C} , in which we impose conditional independence assumptions between observed similarities. We therefore assume the following factorization:

$$P(\mathbf{C} | S) = \frac{1}{Z(S)} \Phi(\mathbf{C}, S) = \frac{1}{Z(S)} \prod \Psi_{i,j}^1(C_{i,j}, S_{i,j}) \Psi^2(C_{i,j})$$

$$Z(S) = \sum_{\mathbf{C} \in \mathcal{C}} \prod \Psi_{i,j}^1(C_{i,j}, S_{i,j}) \Psi^2(C_{i,j})$$

In this model we assume $\Psi_{i,j}^1(C_{i,j}, S_{i,j}) = P(S_{i,j} | C_{i,j})$, and $\Psi^2(C_{i,j}) = P(C_{i,j})$. This is equivalent to assuming full pairwise independence of the variables $C_{i,j}$ and the conditionals $P(S_{i,j} | C_{i,j})$.

For implementational purposes we assume $C_{i,j} \rightarrow S_{i,j}$ are exponentially distributed and the $C_{i,j}$ are Bernoulli random variables.

4.1.1. Training

$\{\sim\}^k$ can be translated into a training sequence of ensemble values $\{C^k\}$ via the transformation

$C_{i,j}^k = 1$ if $i \sim^k j$. Because of the independence assumptions underlying this model, the ML estimate for the posterior distribution of the ensemble can be computed by obtaining the ML estimate for each of the distributions $P(S_{i,j} | C_{i,j})$ and $P(C_{i,j})$. The ML estimate for the rate parameter of $P(S_{i,j} | C_{i,j})$ equals the inverse of the sample mean, and the ML estimate for the mean of $P(C_{i,j})$ equals the sample frequency of $C_{i,j} = 1$. More formally:

Observation 1 Define $\lambda_{i,j} = \frac{1}{m} \sum_k S_{i,j}^k$. And let $p_{i,j} = \frac{1}{m} \sum_k C_{i,j}^k$.

The ML estimator of the parameters for the posterior distribution $P(C | S_{i,j}) = \frac{1}{P(S)} P(S_{i,j} | C) P(C)$ has $P(S_{i,j} | C_{i,j}) \sim \exp(\lambda_{i,j})$ and $P(C_{i,j} = 1) = p_{i,j}$.

4.1.2. Prediction

Prediction under this model is performed by finding the MAP assignment for the ensemble C^* and turning it into a clustering. C^* is obtained by maximizing each likelihood $P(S_{i,j} | C_{i,j}) P(C_{i,j})$ independently:

$$C_{i,j}^* = \arg \max_{C_{i,j} \in \{0,1\}} P(S_{i,j} | C_{i,j}) P(C_{i,j}).$$

For the ensemble assignment C^* we output a clustering composed of a cluster for each connected component of the graph corresponding to C^* . Results are presented in Figure 2.

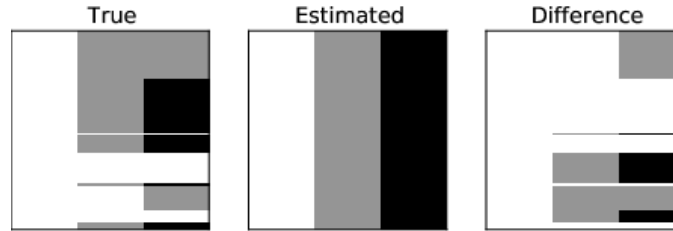


Figure 2: Performance of the exponential model on 5,000 periods of synthetic data with $n = 3$.

The prediction algorithm is linear.

4.1.3. Limitations

Consider the following joint posterior distribution over clusterings of $\{1,2,3\}$.

$$p(\cdot) = \begin{cases} 0.1 & \text{if } \cdot = (1,2,3) \\ 0.41 & \text{if } \cdot = (1,2),(3) \\ 0.41 & \text{if } \cdot = (1,3),(2) \\ 0 & \text{if } \cdot = (2,3),(1) \\ 0.17 & \text{if } \cdot = (1),(2),(3) \end{cases}$$

The marginals $p((1,2)), p((2,3)) > 0.5$. The current algorithm will output $(1,2,3)$.

4.2. Triangular Potentials

The main limitation of the approach described in the previous section is that there is potential for spurious large clusters to emerge solely from the independent optimization of the potentials. If the marginal probability $p_{i,j}$ is large, it is likely that the MAP of the ensemble \mathbf{C} will have $C_{i,j} = 1$ regardless of the values of any of the other similarities $S_{k,m}$ or clustering assignments $C_{k,m}$. It is also possible for the algorithm to suggest cluster shapes which are intuitively implausible (and do not conform to prior notions of cluster structure which may be appropriate to a particular domain); we illustrate this in Figure 3.

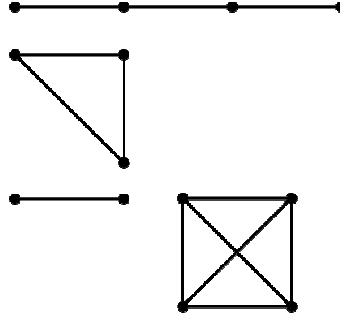


Figure 3: Alternative cluster structures: (top to bottom) the first implausible configuration is ruled-out by the use of *triangular potentials*, however the second and third configurations are possible (which is our deliberate intention).

We therefore proceed to address these issues by a modification to the basic model as described by the following observations:

Observation 2 \mathbf{C} is a valid clustering \sim if, for all triplets of distinct numbers $i, j, k \in [n]$,
 $C_{i,j} = C_{j,k} = 1 \Rightarrow C_{k,i} = 1$.

Observation 3 \mathbf{C} is a valid clustering if the graph whose adjacency matrix equals \mathbf{C} is composed of a disjoint union of cliques.

In this section we assume the following factorization:

$$P(\mathbf{C} | S) = \frac{1}{Z(S)} \Phi'(\mathbf{C}, S)$$

$$= \frac{1}{Z(S)} \left(\prod_{i,j} \Psi_{i,j}^1(C_{i,j}, S_{i,j}) \Psi_{i,j}^2(C_{i,j}) \right) \prod_{i,j,k} \Psi_{i,j,k}^3(C_{i,j}, C_{i,k}, C_{j,k}) \quad (2)$$

$$Z(S) = \sum_{C \in \mathcal{C}} \left(\prod_{i,j} \Psi_{i,j}^1(C_{i,j}, S_{i,j}) \Psi_{i,j}^2(C_{i,j}) \right) \prod_{i,j,k} \Psi_{i,j,k}^3(C_{i,j}, C_{i,k}, C_{j,k})$$

where

$$\Psi_{i,j,k}^3(C_{i,j}, C_{i,k}, C_{j,k}) = \begin{cases} 0 & \text{if } C_{i,j} = C_{i,k} = 1, C_{j,k} = 0 \\ 0 & \text{if } C_{i,j} = C_{j,k} = 1, C_{i,k} = 0 \\ 0 & \text{if } C_{i,k} = C_{j,k} = 1, C_{i,j} = 0 \\ 1 & \text{otherwise} \end{cases}$$

This has the effect of turning $\Phi(\mathbf{C}, S)$ into a potential function $\Phi'(\mathbf{C}, S)$ such that all the assignments of the joint distribution of the ensemble \mathcal{C} with a nonzero probability are valid clusterings.

4.2.1. Training algorithm

We use the same construction for the univariate and bivariate potentials as the one used in the previous section. The distribution over clusterings will vary because the triangular potentials restrict the mass of the distribution to the space of valid clusterings. It is of course also possible to add other potentials relating different sets of clustering variables although we leave that direction for future research.

4.2.2. Prediction algorithms

This model can be thought of as an undirected graphical model with variables $\{C_{i,j}\}$ for $i < j$ and $i, j \in [n]$ and edges $C_{i,j}, C_{i,k}$, $C_{i,j}, C_{j,k}$, and $C_{j,k}, C_{i,j}$ for all $i < j < k$. If the variable $C_{i,j}$ is identified with the point (i, j) , then there is an edge between every two variables on the same vertical line and between every two variables on the same horizontal line.

We tackle the problem of obtaining the MAP assignment over clusterings under this model using either the Elimination Algorithm or MCMC. To obtain an estimate for the MAP assignment using MCMC we sample from the posterior and output the clustering arrangement which appears most often. The MCMC chain construction is described in the next section.

By construction there is a clique of size $n-1$ along the horizontal line $(1, i)$ for $i = 2, \dots, n$. As a consequence, the elimination algorithm has an exponential running time over this graphical model. Similarly, there are no easy theoretical guarantees for the performance of the MCMC method. In particular, it is possible for the probability mass over the optimal assignment to be so small that there are no concentration inequalities to guarantee that the proposed algorithm will output the MAP with high probability in polynomial time.

In the following section we show this behavior is not only a result of the graphical model formulation or our proposed algorithm but an intrinsic limitation of the model itself.

4.3. Results and Limitations

We next apply the classic sumproduct algorithm or the MAP elimination algorithm to find the best clustering, with results shown in Figure 4, however the drawbacks are that this solution becomes intractable as the number of products becomes large. The elimination algorithm could be worst case 2^{n^2} which becomes intractable quite fast.

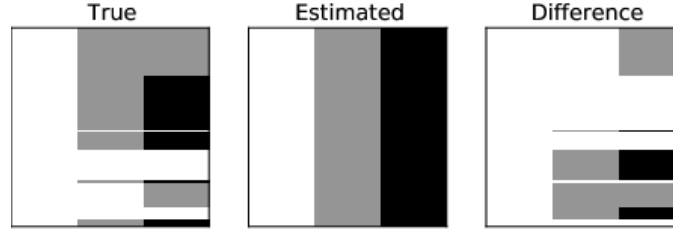


Figure 4: Performance of the graphical model with triangular potentials; $n = 3$.

4.3.1. Theoretical limitations

Let $\hat{p}_{i,j} \in [0,1]$ be an ensemble of probabilities with $i \neq j$ such that $\hat{p}_{i,j} = \hat{p}_{j,i}$ and $i, j \in [n]$. Define a distribution over simple graphs via

$$P(G) = \left(\prod_{(i,j) \in E} \hat{p}_{i,j} \right) \left(\prod_{(i,j) \notin E} (1 - \hat{p}_{i,j}) \right) \forall G = (V, E), V = [n].$$

Let $\hat{P}(G) = P(G \mid G \text{ is a disjoint union of cliques})$.

It is easy to see that finding the MAP assignment for the distribution defined via Equation (2) is equivalent to finding the MAP assignment for $\hat{P}(G)$ with:

$$\hat{p}_{i,j} = \frac{P(S_{i,j} \mid C_{i,j} = 1)P(C_{i,j} = 1)}{P(S_{i,j} \mid C_{i,j} = 1)P(C_{i,j} = 1) + P(S_{i,j} \mid C_{i,j} = 0)P(C_{i,j} = 0)}$$

Since it is conceivable that any arrangement of the values $\hat{p}_{i,j}$ can result from the training data, the two problems are equivalent.

In what follows we talk interchangeably of the MAP assignment $\{\hat{q}_{i,j}^*\}$ of $\hat{P}(G)$ ($\hat{q}_{i,j}^* \in \{\hat{p}_{i,j}, 1 - \hat{p}_{i,j}\}$) and the graph $G^* = (V^*, E^*)$ defined by $V^* = [n]$ and $E^* = \{\{i, j\} \mid \hat{q}_{i,j}^* = \hat{p}_{i,j}\}$. The complement of G^* contains all those pairs $\{i, j\}$ for which $\hat{q}_{i,j}^* = 1 - \hat{p}_{i,j}$.

Theorem 1 *If there is a polynomial time algorithm for finding the MAP assignment over $\hat{P}(G)$ then $P = NP$.*

Proof. Let $A(\{\hat{p}_{i,j}\})$ be an algorithm for finding the MAP over the distribution $\hat{P}(G)$ as defined by $\{\hat{p}_{i,j}\}$. We show A can be used to construct an algorithm for solving the k -clique problem. The k -clique problem is the problem of deciding whether a graph $G = (V, E)$ has a clique of size k where both G and k are inputs to be specified. If A was polynomial, the algorithm we propose for k -clique would run in polynomial time. Because k -clique is NP complete we conclude the existence of A would imply $P = NP$. The following algorithm solves k -clique:

Algorithm 4 *Inputs:* $\langle G, k \rangle$.

Let $N > \max\{2 \cdot |E| + 2, 2|V|\}$ and $q > \frac{1}{2}$.

1. Construct $G' = (V', E')$ with $V' = V \cup V_N$ and $E' = E \cup \{\{v_1, v_2\} \mid v_1 \in V, v_2 \in V_N\} \cup \{\{v_1, v_2\} \mid v_1, v_2 \in V_N, v_1 \neq v_2\}$. The edges of G' equal all edges in G , plus all possible edges between V and V_N and all possible edges among elements of V_N .

2. For all pairs $i, j \in [|V'|]$ define:

$$\hat{p}_{i,j} = \begin{cases} 0 & \text{if } i, j \notin E \\ q & \text{otherwise} \end{cases}$$

3. Let \hat{E}^* be the output edges in the MAP assignment from $A(\{\hat{p}_{i,j}\})$.

4. If $|\text{MaxClique}(\hat{E}^* \cap E)| \geq k$ output 1, else output 0. This step runs in polynomial time because every connected component of $\hat{E}^* \cap E$ is a clique graph.

The probability of the MAP assignment equals

$$\left(\prod_{(i,j) \in \hat{E}^*} \hat{p}_{i,j} \right) \left(\prod_{(i,j) \notin \hat{E}^*} (1 - \hat{p}_{i,j}) \right)$$

which can be written as a product $P_N^* \cdot P_{N \times V}^* P_V^*$ of the product of the chosen probabilities of pairs belonging to $V_N \times V_N$, a cross component of probabilities from $V_N \times V$ and a component of probabilities from $V \times V$. By construction, the edges in $V \times V$ but not in E are not chosen. The MAP restricted to V_N and V is a disjoint union of cliques. Because $P_V^* \leq q^{|E|} < q^{N/2-1}$ we can conclude:

1. The MAP assignment restricted to V_N must be a complete graph: Suppose the MAP restricted to V_N had more than one component, say K_1, Λ, K_r , $|K_1| \geq \Lambda \geq |K_r|$ with K_1^1, Λ, K_r^1 their (possibly empty) corresponding clique intersections in V . It can be shown via the rearrangement inequality that the MAP must have $|K_1^1| \geq \Lambda \geq |K_r^1|$. Let MAP1 be the assignment obtained via joining K_1, Λ, K_r into K_N (the complete graph on V_N) and

reconnecting all to K_1^1 . If $r \geq 2$, a simple counting argument shows that $\text{edges}(K_N) - \sum_{i=1}^r \text{edges}(K_i) \geq \frac{N}{2}$. The latter, and $|K_1^1| \geq \Lambda \geq |K_r^1|$ imply that $P(\text{MAP1}) > P(\text{MAP})$, a contradiction.

2. The $V_N \times V$ edges must connect V_N with one of the largest cliques of G .

The correctness of the algorithm follows. The algorithm above runs in polynomial time, provided \mathbf{A} is in P.

5. EXTENSIONS

5.1. HMM

Because the training procedure we propose is done over fully annotated data, more sophisticated and time-dependent models can be explored. We propose a generalization of the previous models via an HMM.

In this model, each hidden state is a clustering and the transition probabilities are obtained from the sampled frequencies of the transitions in the training phase. When the hidden states of the training data are known, the ML estimate of the transition probabilities of an HMM equals the transitions sample frequencies.

The results of applying this method are shown in Figure 5, where it is apparent that relatively good performance is achieved.

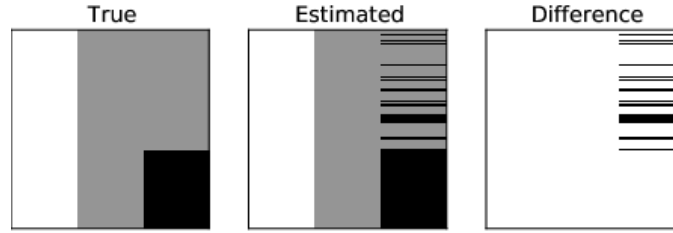


Figure 5: Performance of the HMM clustering algorithm on 2,000 periods of synthetic data with $n = 3$.

The version implemented here is hard-coded for only 3 series and therefore only 5 possible clustering states. The length of the chain can be adjusted as desired.

5.2. Coagulation Fragmentation

The underlying chain for the MCMC sampler uses a fragmentation coagulation process to walk over clusterings. At each step, the chain either selects a random cluster, and divides it into two, or selects two random clusters and joins them together. The acceptance/rejection probabilities can be computed with respect to any coagulation fragmentation process. In our implementation, we pick either a uniform random cluster and a random bipartition of it (fragmentation), or a uniform random pair of clusters (coagulation). We believe the mixing time of this process should be fast as it is related to a coagulation fragmentation process known as the random transposition walk. Diaconis and Shahshahani provided a polynomial upper bound for this walk's mixing time [9].

5.2.1. Alternative model

We believe a worthwhile alternative to the ideas described above is to represent the clustering evolution as an HMM on fragmentation-coagulation parameters: the simplest model having only two parameters (p, q) , one controlling the probability of fragmentation and the other controlling the probability of coagulation. If the number of fragmentation-coagulation parameters is small, inference could be tractable.

6. CONCLUSIONS

Our intention in this paper has been to show how various clustering methods can be applied to datasets which arise in financial markets. We have documented the process by which we analysed the problem and considered a method for determining clusters using *triangular potentials*. This latter method can be computationally intensive and we have provided some preliminary theoretical results concerning its limitations. However, notwithstanding these considerations, we have found promising empirical results from applying the method to simulated datasets and we look forward to extending this to real-world data in due course.

In future work we aim to extend the idea to a setting where we place a non-uniform prior on clusterings, e.g. if expert knowledge suggests that a group of investments are likely to share similar return characteristics then we can configure potentials such that appropriate weighted links are established among these products.

There is also considerable scope to investigate efficiency improvements to the MCMC estimation process, based on the particular structure of potentials in this context.

REFERENCES

- [1] Koller, D. & Friedman, N. (2009) Probabilistic Graphical Models, MIT Press.
- [2] Mas-Colell, A., Whinston, M.D. & Green, J.R. (1995) Microeconomic Theory, Oxford University Press.
- [3] Markowitz, H. (1952) 'Portfolio Selection', Journal of Finance, Vol. 7, No. 1.
- [4] Ingersoll Jr., J.E. (1987) Theory of Financial Decision Making, Rowman and Littlefield.
- [5] Benartzi, S. & Thaler, R.H. (2001) 'Naive diversification strategies in defined contribution saving plans', American Economic Review, 91(1), pp. 79-98.
- [6] De Miguel, V., Garlappi, L. & Uppal, R. (2009) 'Optimal versus naive diversification: How inefficient is the portfolio strategy?', Review of Financial Studies, 22.
- [7] LaViers, A., Rahmani, A. & Egerstedt, M. (2010) 'Dynamic Spectral Clustering', Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems – MTNS 2010, July.
- [8] Chi, Y., Song, X., Zhou, D., Hino, K. & Tseng, B.L. (2007) 'Evolutionary Spectral Clustering by Incorporating Temporal Smoothness', Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York.
- [9] Diaconis, P. & Shahshahani, M. (1981) 'Generating a Random Permutation with Random Transpositions', Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57(2), pp 159-179.

AUTHORS

Aldo Pacchiano is a Bachelor of Science in Computer Science and Mathematics from MIT, with a masters degree in Mathematics from Cambridge University and a recently obtained Masters of Engineering degree from MIT; his interests span the areas of Machine Learning, Computational Biology and he has a special interest in quantitative finance.



Oliver Williams is a quantitative investment specialist with interests including financial economics, asset pricing and systematic trading; he has co-authored a number of papers in this area and his career has been spent in investment banking and asset management. He holds an MA in Computer Science and Management Studies, MPhil and PhD in Financial Economics from Cambridge University.



A NOVEL APPROACH BASED ON TOPIC MODELING FOR CLONE GROUP MAPPING

Ruixia Zhang, Liping Zhang, Huan Wang and Zhuo Chen

Computer and information engineering college,
Inner Mongolia normal university, Hohhot, China
zhangruixia923@163.com

ABSTRACT

Clone group mapping has a very important significance in the evolution of code clone. The topic modeling techniques were applied into code clone firstly and a new clone group mapping method was proposed. By using topic modeling techniques to transform the mapping problem of high-dimensional code space into a low-dimensional topic space, the goal of clone group mapping was indirectly reached by mapping clone group topics. Experiments on four open source software show that the recall and precision are up to 0.99, thus the method can effectively and accurately reach the goal of clone group mapping.

KEYWORDS

code clone; software evolution; topic; topic modeling; clone group mapping

1. INTRODUCTION

The activities of the programmers including copy, paste and modify result in lots of code clone in the software systems. A code clone is a code portion in source files that is identical or similar to another[1]. It is suspected that many large systems contain approximately 9%-17% clone code, sometimes as high as even 50%[2].

After a decade of active research, it is evident that code clones have both a positive [3] and a negative [4] impact in the maintenance and evolution of software systems. For example, copying source code without defect can reduce the potential risk of writing new code, save development time and cost; code clones can cause additional maintenance effort. Changes to one segment of code may need to be propagated to several others, incurring unnecessary maintenance costs.

Code clone is inevitable in software development, and in order to exploit the advantages of clones while lowering their negative impact, it is important to understand the evolution of clones and manage them accordingly. Therefore, in order to meet the demands of clone evolution, a clone mapping method is put forward. Clone group mapping reflect how an clone group evolve from a previous version to the current version, is the core technology in the evolution of code clone across versions.

The topic modeling techniques is applied to code clone firstly and a new clone group mapping method is proposed. Topic modeling technology can make full use of the source text and structure information to transform the mapping problem of high-dimensional code space into a

low-dimensional topic space, and thus indirectly achieve the clone group mapping purposes by mapping clone group topics.

2. RELATED WORK

Software development and maintenance in practice follow a dynamic process. With the growth of the program source, code clones also experience evolution from version to version. what change the Clone group have happened from one version to next need to be made judgments by clone group mapping.

To map clone group across consecutive versions of a program, mainly five different approaches have been found in the literature.

- **Based on text[5]:** It separates clone detection from each version, and then similarity based heuristic mapping of clones in pairs of subsequent versions. Text similarity are often computed by the Longest Common Subsequence(LCS)or Edit Distance(Levenshtein Distance, LD) algorithm that have quadratic runtime, which lead to inefficient clone mapping. The method is susceptible to large change in clone.
- **Based on version management tools (CVS or SVN)[6]:** Clones detected from the first version are mapped to consecutive versions based on change logs obtained from source code repositories. It is faster than the above technique, but can miss the clones introduced after the first version.
- **Based on incremental clone detection algorithm[7]:** Clones are mapped during the incremental clone detection that used source code changes between revisions. It can reduce the redundant computation and save time .So it is faster than the above two techniques, but cannot operate on the clone detection results obtained from traditional non-incremental tools.
- **Based on functions[8]:** It separates clone detection from each version, functions are mapped across subsequent versions, then clones are mapped based on the mapped functions. To some extent, it improves the efficiency and accuracy of the mapping, but it is susceptible to similar overloaded/overridden functions for its over-reliance on function information.
- **Based on CRD(Clone Region Descriptor)[9]:** Clone code is represented by CRD, then clones are mapped based on CRD between versions. It is not easily influenced by position of the code clone. Mutations or big difference between versions can reduce the mapping validity greatly.

This paper presents a new clone group mapping method based on topic modeling technology, unlike the mapping method based on text, the basic collection of the mapping is the clone group topics, not intermediate representation of clone code(e.g., token and AST). Topic has a large granularity and the higher level of abstraction. However, the difference of topics between different clone groups in the same version is very large and the difference of topics between same clone groups in the different version is very little, which make the clone group mapping method based on topic modeling practicable and effective.

3. APPROACH

In this section we present a new clone group mapping approach based on topic modeling for tracking clone groups across different versions.

3.1 Overview of Topic Model

Topic models are generative probabilistic models, originally used in the area of natural language processing for representing text documents. LDA (Latent Dirichlet Allocation) has recently been applied to a variety of domains, due to its attractive features. First, LDA enables a low-dimensional representation of text, which (i) uncovers latent semantic relationships and (ii) allows faster analysis on text [10]. Second, LDA is unsupervised, meaning no labeled training data is required for it to automatically discover topics in a corpus. And finally, LDA has proven to be fast and scalable to millions of documents or more [11]. For these reasons, in this paper we use LDA as our topic model.

In the LDA model, LDA is statistical models that infer latent topics to describe a corpus of text documents [12]. Topics are collections of words that co-occur frequently in the corpus. For example, a topic discovered from a newspaper collection might contain the words {cash bank money finance loan}, representing the “finance industry” concept; another might contain {fish river stream water bank}, representing the “river” concept. So, documents can be represented by the topics within them. Topic modeling techniques transform the text into topic space.

Recently, researchers found topics to be effective tools for structuring various software artifacts, such as source code, requirements documents, and bug reports. Kuhn [13] made the first attempt to apply topic modeling technique to source code, and tried to discover the functional topics. W. Thomas[14]performed a detailed investigation of the usefulness of topic evolution models for analyzing software evolution, they found that topic models were an effective technique for automatically discovering and summarizing software change activities. Asuncion[15] used the topic modeling techniques to study software traceability. Tian [16]used LDA to Automatically classify software in the software repository. Gethersd[17]developed the IDE plug-in, they combined topic modeling results and the existing software development tools to help developers apply topic modeling results. Liu Chao[18] applied topic model to retrieval traceability links between source code and Chinese documentation. Xie Bing[19] from Beijing university proposed a function recognition approach based on LDA and code static analysis technology o better support the activity of code reuse. In addition, topic model was also used to study class cohesion [20]and bug location[21].

3.2 Mapping Clone Group Based on Topic Modeling

3.2.1 Framework of The Algorithm.

Typically a clone detection tool reports results as a collection of clone groups where each clone group has two or more clone fragments. The paper uses the LDA topic modeling technology to map clone group. It mainly works in the following three steps: (1) extracting the topics from clone group, (2) calculating the similarity between topics,(3) mapping clone group topics. Let $CG^n = \{cg_1^n, cg_2^n, \dots, cg_s^n\}$ be the reported clone groups in V_n , $T^n = \{t_1^n, t_2^n, \dots, t_s^n\}$ refers to the clone group topics extracted from the clone group CG^n where t_i^n was extracted from cg_i^n , $1 \leq i \leq n$. Figure 1 shows the framework of the algorithm.

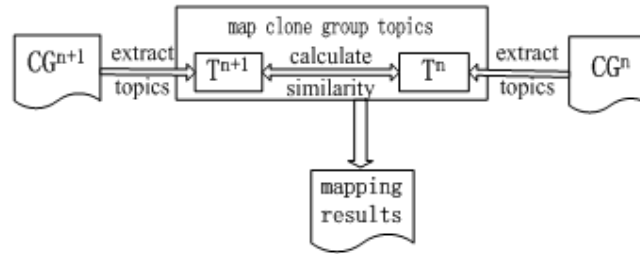


Figure 1. The frame of mapping algorithm

To track clone groups over two different versions V_{n+1} and V_n , we compare every clone group in version V_{n+1} to every clone group in version V_n . Topic modeling technology is used to extract the topics from each clone group in the version V_n and V_{n+1} respectively. At this point, since topic is the only representation of corresponding clone group, the problem of mapping clone group between two versions of a program is reduced to the mapping of clone group topics between two versions. Then clone group topics are mapped by comparing similarity between topics in the version V_{n+1} and V_n . If the topic t_i^{n+1} of a clone group cg_i^{n+1} in version v_{n+1} matches to the topic t_j^n of a clone group cg_j^n in the version V_n , we know that the clone group cg_i^{n+1} in V_{n+1} and the clone group cg_j^n in V_n are the same. Due to the transitivity of the relation of equivalence, we can then conclude that clone group cg_i^{n+1} is related to clone group cg_j^n . The algorithm is as follows:

Clone Group Mapping algorithm

- $\forall cg_i^{n+1} \in CG^{n+1}, CG^{n+1}$ in V_{n+1}
 - extract t_i^{n+1} from cg_i^{n+1}
 - $\forall cg_j^n \in CG^n, CG^n$ in V_n
 - extract t_j^n from cg_j^n
 - calculate similarity between t_j^n and t_i^{n+1} , and store similarity value in the array unit $sim[j]$
 - suppose $sim[k] = \max\{sim[j]\}$:
 IF $sim[k] \geq \delta$, cg_i^{n+1} is mapped back to cg_j^n , namely $cg_i^{n+1} \rightarrow cg_j^n$;
 THEN $cg_i^{n+1} \rightarrow null$.
 - Return all mapping results
-

3.2.2 Extract Clone Group Topics

Under the standard programming style, software is suitable for extracting the topics using the LDA model. The paper uses MALLET topic modeling toolkit to extract the topics. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. It contains efficient, sampling-based implementations of LDA.

Preprocess the source code. Each topic is collections of words that co-occur frequently in the clone group, and is the only representation of corresponding clone group. The topic contains a large number of stop words which play a small role in characterization of clone group information. So, we remove stop words to reduce noise before extracting the topics. Stop words mainly include the following three categories : 1) programming language keywords, such as "for", "return", and "class", etc. 2) Programming related words, such as "main", "arg", and "util", etc. 3) common English language stop words, such as "the", "it", and "on", etc.

Choose the number of topic. In order to make the topic accurately represent the information of clone group, the proper number of topic is a key to influence the accuracy of clone group mapping. For any given corpus, there is no provably optimal choice for the number of topics. The choice is a trade-off between coarser topics and finer-grained topics. setting the number of topics to extremely small values results in topics that contain multiple concepts, while setting the number of topics to extremely large values results in topics that are too fine to be meaningful and only reveal the idiosyncrasies of the clone group.

In the paper, through experimental analysis, it is best for setting the number of topics to one. In the same clone group, clone code is a code portion that is identical or similar to another. The whole clone group is multiple copies of the same clone whose syntactic or semantic function is same. In other words, a clone group can be represented by a topic. The topics extracted from clone group by MALLET are shown in Figure 2.

```
<topics>
- <topic titles="tmplist, false, tmpdoc, tdocument, bfwin, save, backend, modified,
data, documentlist" totalTokens="62" alpha="1.001" id="0">
  <word count="12" weight="0.1935483870967742">tmplist</word>
  <word count="8" weight="0.12903225806451613">false</word>
  <word count="8" weight="0.12903225806451613">tmpdoc</word>
  <word count="4" weight="0.06451612903225806">list</word>
  <word count="4" weight="0.06451612903225806">tdocument</word>
  <word count="4" weight="0.06451612903225806">bfwin</word>
  <word count="4" weight="0.06451612903225806">save</word>
  <word count="2" weight="0.03225806451612903">backend</word>
  <word count="2" weight="0.03225806451612903">doc</word>
  <word count="2" weight="0.03225806451612903">modified</word>
  <word count="2" weight="0.03225806451612903">data</word>
  <word count="2" weight="0.03225806451612903">documentlist</word>
  <word count="2" weight="0.03225806451612903">glist</word>
  <word count="2" weight="0.03225806451612903">tbwin</word>
  <word count="1" weight="0.016129032258064516">widget</word>
  <word count="1" weight="0.016129032258064516">gtkwidget</word>
  <word count="1" weight="0.016129032258064516">cb</word>
  <word count="1" weight="0.016129032258064516">file</word>
</topic>
</topics>
```

Figure 2. The topics extracted by MALLET

3.2.3 Mapping Clone Group Topics

Clone group mapping is determined by the degree of similarity between clone groups in the different versions. In the paper, clone group mapping is determined indirectly by similarity between clone group topics from different versions. If the similarity between the topic t_i^n and topic t_i^{n+1} is highest, and the similarity values is not less than certain threshold (δ). In that way, we can conclude that the topic t_i^{n+1} is mapped back to the topic t_j^n , namely the clone group cg_i^{n+1} is mapped back to the clone group cg_j^n . In the paper, Similarity threshold δ is set to 0.8. That's because the similarity value between t_i^{n+1} and t_j^n vary from 0.8 to 1 when cg_i^{n+1} is mapped back to the clone group cg_j^n , and the similarity value between t_i^{n+1} and t_j^n is less than 0.8 when cg_j^n is not origin of clone group cg_i^{n+1} .

In the paper, the mapping is carried out from the version V_{n+1} to V_n . That's because the number of clone group is generally on the rise in the process of software evolution. If the mapping is carried out from the version V_n to V_{n+1} , new clone groups are failed to map. On the contrary, disappeared clone groups are failed to map. However, we are more interested in clone code near to the current version in the study of clone evolution. That is to say, compared with disappeared clone group, we are more interested in new clone group. So, the mapping is carried out from the version V_{n+1} to V_n .

4. CASE STUDY

4.1 Systems Under Study

Due to the difference in size of software system, number of clone group in each version ranging from dozens to thousands ,in view of the limitations of manually inspection, so we perform case study on the source code of four small and medium-sized, open source software systems which is written in different programming languages. The detail of software is shown in Table 1.

Table 1. The detail of software

software	Bluefish	MALLET	ArgoUML	PostgreSQL
Implementation language	C	JAVA	JAVA	C
Average size	23MB	31MB	35MB	92MB
Number of the selected version	2	3	3	4
Number of Clone group (on average)	20	145	299	506

In the paper, NiCad is used to detect clone code. NiCad , a clone detector, can detect Type-1 , Type-2 and Type-3 clones written in multiple programming language (C、JAVA、C#) and have a high precision rate and recall rate. In the Linux platform, NiCad is used to detect Type 1, Type 2 and Type 3 clones of the software. Then we transfer clone group files to the Windows platform to map the clone group across versions.

4.2 Evaluation Measures

To evaluate the feasibility and validity of the approach, we use Precision and Recall as Evaluation Measures to manually inspect the results of the approach based on topic modeling . Precision and Recall are defined as follows:

Precision: Of all the clone group mappings discovered, how many are correct?
We calculate the precision of the experimental result as

$$\text{Precision} = \frac{\text{the number of correct mapping}}{\text{the number of correct and incorrect mapping}}$$

Recall: Of all the actual clone group mappings, how many were discovered?
We calculate the recall of the experimental result as

$$\text{Recall} = \frac{\text{the number of correct mapping}}{\text{the number of actual clone group mappings}}$$

4.3 Results

Take bluefish for example, mapping results of clone groups between bluefish 2.2.4 and bluefish 2.2.3 are shown in Figure 3.

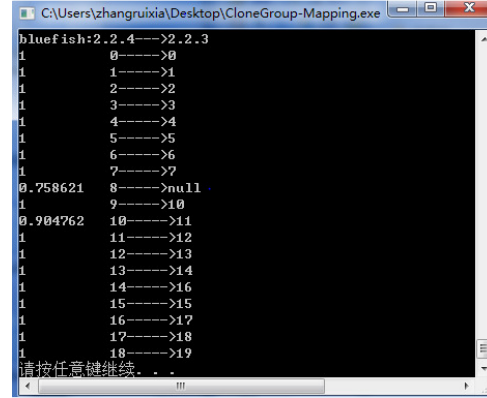


Figure 3. Mapping results of clone groups between Bluefish 2.2.4 and 2.2.3

The second and third column of the figure show clone group number of the corresponding version in Bluefish. There are 19(From 0 to 18) clone groups in Bluefish 2.2.4. There are 20(From 0 to 19) clone groups in Bluefish 2.2.3. The arrows indicate the corresponding clone group is traced to its origin clone group. For example, the 14th clone group of Bluefish 2.2.4 is mapped back to the 16th clone group of Bluefish 2.2.3. But the 14th clone group of Bluefish 2.2.4 is not traced to its origin clone group, which indicate that it is a new clone group, probably the great changes have taken place in its origin during the software evolution from Bluefish 2.2.3 to Bluefish 2.2.4, which similarity value between them is less than the threshold δ . We note that 8th and 9th clone group of Bluefish 2.2.3 do not appear in the list, probably they are removed or take great change during software evolution. The first column of the figure show the largest similarity values of clone group topics between Bluefish 2.2.4 and Bluefish 2.2.3. If the value is not less than δ (0.8), There is a mapping relationship between them. It can be seen in the Figure 3 that most of the similarity value are as high as 1, namely most of the clone groups do not change during software evolution. Few of the similarity values are not 1, which indicate that clone codes have experienced some degree of change, such as the clone group is deleted, a few of clone fragments are added or removed.

Clone group mapping is carried on consecutive versions of other software, and manually inspect the Precision and Recall of the mapping results. The results can be seen in Table 2 and Table 3. The Precision and Recall of the approach are as high as 0.99, which reveal the validity and feasibility of clone group mapping approach based on topic modeling. The runtime of clone group mapping across versions is acceptable. Since number of clone groups in some software is large, in view of the limitations of manually inspection, we conduct experiments on only 12 versions of the above 4 software. But the results are enough to reveal the feasibility of the approach.

Table 2. The experimental results of the approach

Software and versions Evaluation Measures	Bluefish		PostgreSQL		PostgreSQL		PostgreSQL	
	2.2.4	2.2.3	9.1.5	9.1.4	9.1.4	9.1.3	9.1.3	9.1.2
Precision	1		0.996		0.996		0.994	
Recall	0.95		1		1		1	

Table 3. The experimental results of the approach

Software and versions Evaluation Measures	ArgoUML		ArgoUML		MALLET		MALLET	
	0.27.3	0.27.2	0.27.2	0.27.1	2.0.7	2.0.6	2.0.6	2.0.5
Precision	1		0.996		1		0.992	
Recall	0.996		0.993		0.982		0.992	

5. DISCUSSION AND THREATS TO VALIDITY

5.1 Limitations of Similarity Threshold

In the paper, similarity threshold between clone group topics across versions is determined based on the experience knowledge, and different software use the same similarity threshold, which have an impact on the results. Firstly, similarity threshold based on the experience knowledge can't reflect mapping efficiency of the algorithm. Secondly, the same threshold is used to different software that they exist remarkable differences in programming language, programming style and the degree of change between versions, which will reduce the validity of the mapping algorithm.

5.2 Limitations of Clone Detector

The clone detector provides the basis data for clone group mapping, so clone group mapping approach directly is affected by clone detector. It is critical for clone group mapping to choose an accurate clone detector.

5.3 The Differences between Versions

It is discovered by the experimental results that the smaller differences between versions is, the higher accuracy the approach has. If clone group have happened so significant changes during software evolution that similarity value between two versions exceed the permitted threshold, which clone group that could have been traced to its origin clone group is failure to mapping. That is to say, Mutation or big difference between versions can reduce the accuracy of the mapping. Therefore, It contributes to improvement of accuracy of clone group mapping that using revision of software rather than release.

6. CONCLUSIONS

The activities of the programmers including copy, paste and modify result in lots of code clone in the software systems. However, Clone group mapping has a very important significance in the evolution of code clones. The clone group mapping approach based on topic modeling is proposed in the paper. By using topic modeling techniques to transform the mapping problem of high-dimensional code space into a low-dimensional topic space, the goal of clone group mapping was indirectly reached by mapping clone group topics. Experiments on 12 versions of 4 open source software show that the recall and precision are up to 0.99, thus the approach can effectively and accurately reach the goal of clone group mapping.

REFERENCES

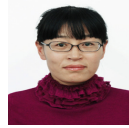
- [1] Bettenburg N, Shang W, Ibrahim W, et al. An Empirical Study on Inconsistent Changes to Code Clones at Release Level[C]//Proc. of the 2009 16th Working Conference on Reverse Engineering. IEEE Press, pp. 85-94, 2009.
- [2] Zibran M F, Roy C K. The Road to Software Clone Management: A Survey[R], Technical Report 2012-03, The University of Saskatchewan, Canada, 2012, pp. 1-66.
- [3] M. Kim, V. Sazawal, D. Notkin, and G. C. Murphy, "An Empirical Study of Code Clone Genealogies," Proc. ESEC-FSE, 2005, pp. 187–196.
- [4] F. Rahman, C. Bird, P. Devanbu, "Clones: What is that Smell?," Proc. MSR, 2010, pp. 72–81.
- [5] Bakota T, Ferenc R, Gyimothy T. Clone smells in Software evolution[C]//IEEE International Conference on Software Maintenance. Washington DC: IEEE Computer Society, 2007:24-33.
- [6] Barbour L, Khomh F, Zou Y. Late propagation in software clones[C]//Proceedings of the 27th IEEE International Conference on Software Maintenance. Washington DC:IEEE Computer Society, 2011: 273-282.
- [7] Gode N, Koschke R. Incremental Clone Detection[C]//Proceedings of the 2009 European Conference on Software Maintenance and Reengineering. Washington DC:IEEE Computer Society, 2009: 219-228.
- [8] Saha R K, Roy C K, Schneider K A. An automatic framework for extracting and classifying near-miss clone genealogies[C]//Software Maintenance (ICSM), 2011 27th IEEE International Conference on. IEEE, 2011: 293-302.
- [9] Duala-Ekoko E, Robillard M P. Tracking Code Clones in Evolving Software[C]//Proceedings of the 29th international conference on Software Engineering. Washington DC:IEEE Computer Society, 2007:158-167.
- [10] C.X. Zhai, Statistical language models for information retrieval, Synthesis Lectures on Human Language Technologies 1 (1) (2008) 1–141.
- [11] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, M. Welling, Fast collapsed Gibbs sampling for latent Dirichlet allocation, in: Proceeding of the 14th International Conference on Knowledge Discovery and Data Mining, 2008, pp. 569–577.
- [12] D.M. Blei, J.D. Lafferty, Topic models, in: Text Mining: Classification, Clustering, and Applications, Chapman & Hall, London, UK, 2009, pp. 71–94.
- [13] Kuhn A, Ducasse S, Girba T. Semantic clustering: Identifying topics in source code. Information and Software Technology, 2007, 49(3):230–243
- [14] Thomas S W, Adams B, Hassan A E, et al. Studying software evolution using topic models[J]. Science of Computer Programming, 2012
- [15] Asuncion H, Asuncion A, Taylor R. Software traceability with topic modeling.32nd ACM/IEEE International Conference on Software Engineering (ICSE). 2010:95–104
- [16] Tian K, Reville M, Poshyvanyk D. Using Latent Dirichlet Allocation for automatic categorization of software. 6th IEEE International Working Conference on Mining Software Repositories (MSR). 2009:163–166
- [17] Gethers M, Savage T, Di Penta M, et al. CodeTopics: Which topic am I coding now? 33rd International Conference on Software Engineering (ICSE). 2011:1034–1036
- [18] HAN Xiaodong ,WANG Xiaobo, LIU Chao.Retrieval method for traceability links between source code and Chinese documentation[J]. Journal of Hefei University of Technology: Natural Science, 2010 , 33 (2) : 188-192.
- [19] JIN Jing, LI Meng, HUA Zhebang, SONG Huaida, ZHAO Junfeng, XIE Bing. Code function recognition approach based on LDA and static analysis[J]. Computer Engineering and Applications,2013(15).
- [20] Liu Y, Poshyvanyk D, Ferenc R, et al. Modeling class cohesion as mixtures of latent topics[C]//Software Maintenance, 2009. ICSM 2009. IEEE International Conference on. IEEE, 2009: 233-242
- [21] Lukins S, Kraft N, Etzkorn L. Bug localization using latent Dirichlet allocation. Information and Software Technology, 2010, 52(9):972–990.

AUTHORS

Ruixia Zhang, born in 1989, master, student at Inner Mongolia normal university. Her current research interests include software engineering, code analysis.



Liping Zhang, born in 1974, master, professor at Inner Mongolia normal university. Her current research interests include software engineering, code analysis.



Huan Wang, born in 1991, master, student at Inner Mongolia normal university. His current research interests include software engineering, code analysis.



Zhuo Chen, born in 1989, master, student at Inner Mongolia normal university. His current research interests include software engineering, code analysis.



AUTHOR INDEX

- Abeer Elkorany* 21
Abobakr Ahmed Bagais 01
Adilah Hanin Zahri N 73
Akram Salah 21
Aldo Pacchiano 197
Alexandra L. Uitdenbogerd 43
Anubhav Gupta 163
- Beyza Eken* 155
BOUTOUHAMI Khaoula 183
- Chae-tae Im* 11
ChongGun Kim 141
Cüneyd Tantug A 155
Cynthia S. Mlambo 33
- Farzana Shafique* 175
Fatma Algali 93
Fulufhelo V Nelwamondo 59
Fulufhelo V. Nelwamondo 33
Fumiyo Fukumoto 73
- Gaurav Singh Thakur* 163
- hcbae* 11
Huan Wang 213
Hyeoncheol Zin 141
- Imtiaz Hussain Khan* 01
- JongHun Jung* 11
- Kamal Mansoor Jambi* 01
Khaled Aldossari 101
Khalid Majrashi 43
KHELLAF Faiza 183
- Liping Zhang* 213
- Mabroka Maeref* 93
Margaret Hamilton 43
Mary Wu 141
Matsyoshi Suguru 73
Meshack B. Shabalala 33
Mohammad Derawi 111
- Muazzam Ahmed Siddiqui* 01
- Oliver J. Williams* 197
Ong Bi Lynn 73
- Ruixia Zhang* 213
- Saad Alkahtani* 123
Samah Meghawry 21
Sisanda Makinana 59
Soojin Yoon 11
- Tarek Elghazaly* 21
Tendani Malumedzha 59
Turki Abdullah 141
- Wei Liu* 123
William J. Teahan 123
- Zhuo Chen* 213