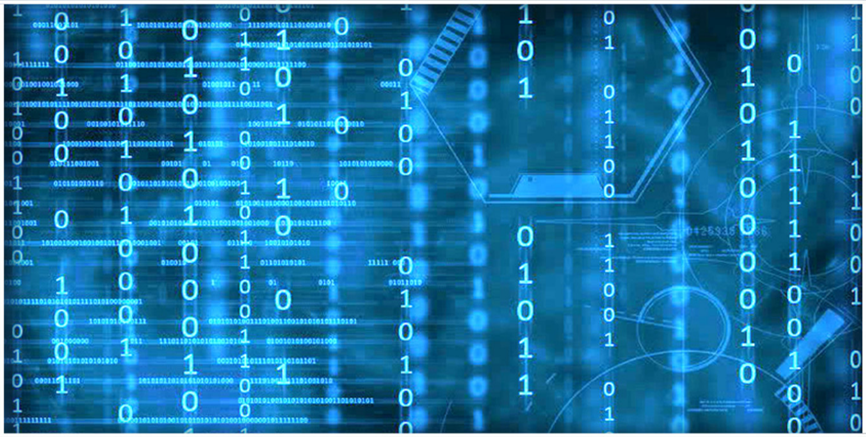
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 4 Issue 11

November 2013



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the November 2013 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 4 Issue 11 November 2013
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Implicit Sensitive Text Summarization based on Data Conveyed by Connectives

Henda Chorfi Ouertani
Information Technology Department
College of Computer and Information Sciences, King Saud University
Riyadh,Saudi Arabia

***Abstract*—So far and trying to reach human capabilities, research in automatic summarization has been based on hypothesis that are both enabling and limiting. Some of these limitations are: how to take into account and reflect (in the generated summary) the implicit information conveyed in the text, the author intention, the reader intention, the context influence, the general world knowledge…. Thus, if we want machines to mimic human abilities, then they will need access to this same large variety of knowledge. The implicit is affecting the orientation and the argumentation of the text and consequently its summary. Most of Text Summarizers (TS) are processing as compressing the initial data and they necessarily suffer from information loss. TS are focusing on features of the text only, not on what the author intended or why the reader is reading the text. In this paper, we address this problem and we present a system focusing on acquiring knowledge that is implicit. We principally spotlight the implicit information conveyed by the argumentative connectives such as: but, even, yet …. and their effect on the summary.**

Keywords—Automatic summarization; implicit data; topoi; topos; argumentation

I. Introduction

Nowadays, text summarization has become widely used on the internet. Users of text summarization are countless. They can be simple internet surfers searching for different news, e-learners looking for specific educational materials or scientists exploring particular publications… Text summarization can help those users identify, in a short time (by reducing a large amount of information to a summary), which documents are most relevant to their needs. But, there is widespread agreement that summarization that reduces a large volume of information to a summary preserving only the most essential items, is a very hard process. Indeed, the human summarization is the process that given a document one tries to *understand, interpret, abstract* it and finally *generate* a new document as its summary [1].

So far and trying to reach human capabilities, research in automatic summarization has been based on hypothesis that are both enabling and limiting. Some of these limitations are: how to take into account and reflect (in the generated summary) the implicit information conveyed in the text, the author intention, the reader intention, the context influence, the general world knowledge ... Thus, If we want machines to mimic human abilities, then they will need access to this same large variety of knowledge [2].

Most of Text Summarizers (TS) are processing as compressing the initial data and they necessarily suffer from information loss. TS are focusing on features of the text only, not on what the author intended or why the reader is reading the text. Thus a TS system must identify important parts and preserve them. In this paper, we will focus on acquiring knowledge that is implicit in the data and how to preserve it when generating the summary. The system we present generate argumentative text based on the implicit stored data conveyed by the “argumentative connectives” such as nevertheless, therefore, but, little, a little... When those connectives appear in sentences, they impose constraints on the argumentative movement. This movement is based on gradual rules of inference denoted by “topoi” [3]

The paper is organized as follows: in section 2, we give an overview of the state of the art on text summarization. Section 3 reports on the theory of Argumentation Within Language (AWL) on which is based our implicit extractor. In section 4, we describe our system architecture. In conclusion, we summarize the contributions of this paper and introduce future research directions.

II. Text Summarization

A. Types of summarizers

Text summarization is now an established field of natural language processing, attracting many researchers and developers. We can distinguish two types of summarizers based on the volume of text to be summarized:

- Single Document Summarization (SDS): If summarization is performed for a single text document then it is called as the single document text summarization
- Mutli Document Summarization (MDS) : If the summary is to be created for multiple text documents then it is called as the multi document text summarization

B. Summarization techniques

Techniques may vary depending on the summarization type. When considering the Single Document Summarization, we can cite the most important techniques:

- Sentences extracting: This technique relies on trivial features of sentences, such as word frequency, presence of keywords, and sentence position, or a combination of such features [4], [5].

- Identification of the relevant information: permitting to generate a textual summary from the facts that need to be included [6], [7].

However, when dealing with Multi-document summarization, we can talk about

- Extractive summarization: this technique involves assigning scores to some units (e.g. sentences, paragraphs) of the documents and extracting those with highest scores [8].

- Abstractive summarization: this technique usually needs information fusion, sentence compression and reformulation [4].

III. How connectives are affecting sentence orientation

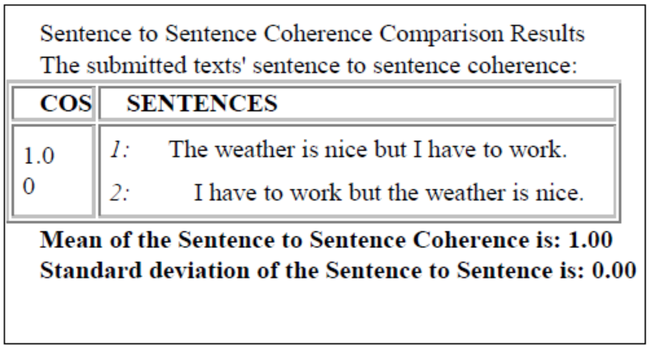
A. *Introduction*

In order to show the importance of the connective on the orientation of the sentence and on its general meaning, we used LSA tool (http://lsa.colorado.edu/) to compare two apparently same sentences. LSA is a theory and a method for extracting and representing the contextual usage meaning of words by statistical computation. LSA measures of similarity are considered highly correlated with human meaning similarities among words and texts. Moreover, it successfully imitates human word selection and category judgments [9].

Example 1:
Let us consider the two following sentences:

1) *The weather is beautiful but I have to work*
2) *I have to work but the weather is beautiful*

With LSA the two sentences will be represented with the same semantic vectors (fig. 1.) because for LSA the words like I, to, but … are ignored and the word order is not take into account.



Sentence to Sentence Coherence Comparison Results
The submitted texts' sentence to sentence coherence:

COS	SENTENCES
1.0 0	*1:* The weather is nice but I have to work. *2:* I have to work but the weather is nice.

Mean of the Sentence to Sentence Coherence is: 1.00
Standard deviation of the Sentence to Sentence is: 0.00

Fig. 1. Comparison of two sentences similarity, Comparison from http://lsa.colorado.edu/

But we agree that the two sentences argue to two different conclusions. So, it is definitely the impact of ignoring the connective *but.*

B. *Argumentation Within Language Theory*

The Argumentation Within Language Theory (AWL) [10] has been concerned with the analysis of the "argumentative articulators" such as nevertheless, therefore, but, little, a little... When those articulators appear in utterances, they impose on constraints on the argumentative movement. This movement is based on gradual rules of inference denoted by "topoi". According to [11] and [12], a topos is an argumentative rule shared by a given community (which need have no more members than the speaker and the hearer). Topoi are the guarantors of the passage from the argument to the conclusion. Topoi are used to license the move from an argument to a conclusion.

A topos (singular of topoi) is:

- Presented as general: in the sense that the speaker implicates that the topos holds for other situations. It is not particular for the situation where it is used.

- Presented as shared: in the sense that the speaker considers that the topos is accepted at least by the audience.

- Gradual.

The canonical form of the topos includes two argumentative scales: the argument (antecedent) and the conclusion (consequent).

Each scale is marked on "plus" or on "minus" from which the next topical forms are concluded:

// + P , + Q//,
// - P , - Q//,
// + P , - Q// and
// - P , + Q//.

If we believe // + P , + Q//, we necessarily believe // - P , - Q// and in the same way for (//+ P , - Q// ; // - P , + Q//)
To illustrate the presentation above, let us consider the utterance

(1) The weather is beautiful but I have to work.

The antecedent uses a topos such as //plus weather is beautiful, plus we want to go out//, the conclusion uses a topos such as //plus I have a work to do, minus I go out//. The use of "but" in the utterance influences its argumentative orientation and the all utterance orientation will be the orientation of the conclusion.

Let us now consider together the two sentences of example1: According to the AWL, the two sentences have opposite argumentative orientations.

Indeed, for the sentence 1, if the antecedent uses topos like //+ beautiful weather, + outing// and the conclusion uses topos like //+ work, - outing// then the presence of "but" imposes that the sentence have the argumentative orientation of the conclusion i.e. "- outing".

However, for the sentence 2, and with the same reasoning, its argumentative orientation is "+ outing"

To end this illustration, we note the importance of "but", in the sense that it imposes the argumentative orientation of the sentence. This importance of connectives was already revealed by different works on Natural Language Process such as in [13] “interclausal connectives carry meaning, they connect textual meanings at both local and global levels and they mark discourse continuity and discontinuity both in the text and as inferred by the reader”

Connectives can shape the actual meaning of the text, they can also serve as efficient markers for instructions in the communicative process established between writer and reader.

After this short outline on the theory of the Argumentation Within Language, in the next section we give a description of the architecture of an Argumentative Single Document Summarizer (ASDS).

IV. SYSTEM ARCHITECTURE

This section gives an overview of the ASDS architecture and describes the functions of its various components. The global architecture is represented in Figure 1. It is composed of three layers of software : the Data pre-processor, the constraints generator and the summary generator.

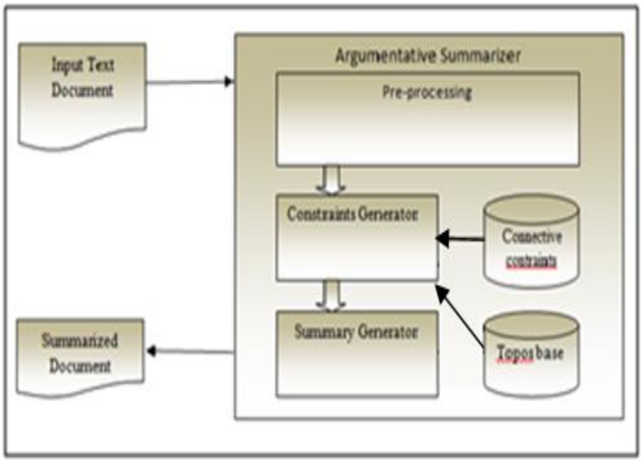


Fig. 2. ASDS Architecture

The pre-processing layer aims at extracting connective elements. ASDS uses GATE [14] a natural language processing system.

The generator constraints layer Generate constraints based on the connectives constraints and the topos base. It permits to annotate the relevant sentences in the text. In our work we consider the sentence as the basic extraction unit. The connective constraints determine the type of argumentative relation between the argument and the conclusion - whether an argument-conclusion relation or argument-anti-argument relation- The topos base is used to link arguments to conclusions. This base allows the comparison of two arguments across scales (since a topos is gradual as discussed above).

We notice that the proposed summarization is focused on single document texts where argumentation takes an important place. The summary generator aims to filter sentence according to the constraints predetermined by the constraints generator. The algorithm below gives the different steps of summary generation :

- Identify all sentences S={Si} of the document d.
- Calculate sentences score with respect to their importance for the overall understanding of the text. This ranking is based on key words and connectives.

Sentences with connectives are weighted contrary to other sentences.

Key words are determined by their frequency in the document.

A Word-Sentence matrix is generated, where the column represents the sentences and the row represents the words. Words with maximum frequency are considered as key words.

Calculate the score for each sentence using a formula using the key words weight and connectives weight :

$$Score(Si) = Cw*Ww$$

	S1	S2	…	….	Sn
W1					
W2					
..					
…					
Wn					
Ww					
Cw					
Score					

Where Cw is the weight of connectives and Ww is the weight of key words.

- Rank the sentences in the decreasing order of calculated scores.
- Apply connectives constraints on sentences including connectives to generate conclusions.
- Top ranked sentences and generated conclusions are combined in sequence as document summary.

V. FUTURE WORK

In the present work, we showed the role of connectives in argumentative texts when dealing with the orientation of the whole text. The analysis of these connectives indicates the existence of specific values intentionally assigned to them by the writer named topoi. As future work, we plan to investigate the topoi base. Many works need to be conducted especially how this base will be initialized and how it will be updated. We would like to continue the implementation of ASDS to apply our approach. Moreover, choosing argumentative texts to be used as input to our system needs further investigation.

VI. CONCLUSION

In this paper we showed the role of connectives in argumentative texts when dealing with the orientation of the whole text. The analysis of these connectives indicates the existence of specific values intentionally assigned to them by the writer. For example *But* was shown to be functioning in sentence to impose constraints on the conclusion intended by the writer. Some recent trends of investigation support different roles for these connectives in the construction of summaries of argumentative texts. In this context, we present the architecture of ASDS, an Argumentative Single Document Summarizer. ASDS is based on topoi which are gradual rules of inference. Topoi are the guarantors of the passage from the argument to the conclusion.

Acknowledgment

This research project was supported by a grant from the Research Center of the Center for Female Scientific and Medical Colleges in King Saud University.

References

[1] Georges Gardarin, Huaizhong Kou, KarineZeitouni, XiaofengMeng, Haiyan Wang: SEWISE: An Ontology-based Web Information Search Engine. NLDB 2003: 106-119, 2003.

[2] Benjamin D. Van Durme, Extracting Implicit Knowledge from Text, ProQuest, UMI Dissertation Publishing, 2011

[3] S.Bruxelles, O.Ducrot, P.Y. Raccah, Argumentation and the lexical topical fields, Journal of Pragmatics 24, 99-114, 1995

[4] Paice, Chris and Paul Jones. The identification of important concepts in highly structured technical papers. In Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR'93), 1993.

[5] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995.

[6] D. Radev and K. McKeown. Generating natural language summaries from multiple on-line sources.*Computational Linguistics*, 24(3):469-500, September 1998

[7] C.Y. Lin and E. Hovy.The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18thInternational Conference on Computational Linguistics (COLING2000),* Saarbrcken, Germany, July 31- August 4, 2000

[8] Rada Mihalcea and Paul Tarau, "An Algorithm for Language Independent Single and Multiple Document Summarization", *In Proceedings of theInternational Joint Conference on Natural Language Processing (IJCNLP),* Korea, 2005

[9] Landauer, T.k., Foltz, P. w., Laham d. An introduction to Latent Semantic Analysis, Discourse processes, 25, 259-284, 1998.

[10] Anscombre, J.-C. & O. Ducrot, *L'argumentation dans la langue*. Brussels: Pierre Mardaga, 1983

[11] Moeschler, J. & A. Reboul, *Dictionnaire encyclopédique de pragmatique*. Paris: Seuil, 1994

[12] Nyan, T., Metalinguistic Operators with Reference to French.Bern: Peter Lang, 1998.

[13] SEGAL, E. M., J. F. DUCHAN, and P. SCOTT, "The Role of Interclausal Connectives in Narrative Structuring: Evidence from Adults' Interpretations of Simple Stories." *Discourse Processes* 14: 27-54, 1991.

[14] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL-02*, 2002

[15] http://lsa.colorado.edu/

The Development of Learning Resource for the Saleng Community under the Bridge of Zone 1 Entitled “How to Repair Electrical Appliances”

Pornpapatsorn Princhankol, Kuntida Thamwipat, Paveena Thambunharn, Wuttipong Phansatarn
Faculty of Industrial Education and Technology
King Mongkut’s University of Technology Thonburi
Bangkok, Thailand

***Abstract*—This research was aimed to develop a learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” in order to determine the quality of learning community and to examine the satisfaction of the sampling group which was chosen using specified sampling method. The tools used in this study consisted of 1) learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances”, including video clips, posters, leaflets, and exhibitions; 2) the quality assessment forms; and 3) the satisfaction evaluation form. According to the quality assessment of the learning resource for the Saleng community under the bridge of zone 1 by the panel of 6 experts, the mean score for the quality in terms of contents was 4.83 with S.D. of 0.29, or at a very good level. As for the quality in terms of media, the mean score was 4.47 with S.D. of .63, or at a good level. Afterwards, the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” was used by the sampling group of 30 persons. The satisfaction towards the learning resource was 4.71 with S.D. of .64, or at the highest level. It could be concluded that the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” can be used as a practical instructional tool for the community.**

Keywords—learning resource; Saleng community under the bridge of zone 1; repairing electrical appliances

I. Introduction

The communities under the bridge in many parts of Bangkok are places where many homeless lived. Bangkok Metropolitan Administration and National Housing Agency have collaborated to provide over 700 families of these people with 3 plots of land which are not far from their previous places. These are the community under the bridge at Pracha-Utit 76, Toong-kru District (Zone 1), the community under the bridge at Poonsarp, Saimai District (Zone 2) and the community under the bridge at Onnuch, Prawate District (Zone 3). The community under the bridge of zone 1 is located at Pracha-Utit Road Soi 76, Toong-kru District, Bangkok which is about 10 kilometres away from King Mongkut’s University of Technology Thonburi. This 13-rai* (*1 rai is equal to 1,600 square metres) plot of land houses 182 families at the moment. There are public areas such as sports field, playground, and pre-school development centre to hold meetings and activities among family members. The majority of people or over 70% of them work as itinerant junk buyers, in other words, they buy and collect unwanted or faulty electrical appliances, litter, empty plastic bottles, paper and the like and then they classify and sell them later. The majority of people are poor and their educational level was not high. They use saleng or three-wheeled pedal cart as their vehicle and as such, their community is sometimes called “Saleng Community” which is one of many disadvantaged communities in Thailand.

In 2011, King Mongkut’s University of Technology Thonburi (KMUTT) conducted the research study entitled “Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok” [1] with National Institute of Development Administration and Bangkok Metropolitan Administration in order to examine and analyse the current situations and requirements of the community. The results from this research included 15 developmental policy plans which had been amended by the community commission and the community people.

Recently, from August 2012 to January 2013, the researchers have participated in the project entitled “Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers”, which was funded by Siam Commercial Bank PLC in order to improve the living quality of those people. In this project, teachers from electrical and automobile departments as well as recycling specialists have taught the community members so that they could apply the knowledge to their profession. The research outcome shows that the community expressed their satisfaction at a high level and such research project was the first runner up of the national SCB Challenge Community Project Award granted by Siam Commercial Bank PLC [2]. The committee recognises the importance of community development with the aim that the community will develop continuously and that the demands of the community members under the bridge of zone 1 will be met as a way to enhance and extend the above-mentioned research project entitled “Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok”.

Therefore, the researchers would like to develop the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” (how to repair fans, irons and rice cookers) so that there is a learning centre for the community. In this project, the researchers developed video clips entitled “How to Repair Electrical Appliances”, posters to provide them with knowledge about how to maintain electrical appliances, exhibition about the history of the local community and leaflets with information about the community so that the community members could use this knowledge to support their career by repairing electrical appliances to increase the value of old junk, resulting in more income for the community members. Moreover, the exhibition about the history of the local community and the leaflets would become the centre of attention for those interested and visitors to know more about the community.

II. RESEARCH METHODOLOGY

A. Research Objectives

1) To develop a learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances”

2) To assess the quality of the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances”

3) To examine the users’ satisfaction towards the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances”

B. Research Hypotheses

1) The quality of the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” would be at a high level.

2) The users would express their satisfaction towards the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” at a high level.

C. Expected Outcomes

1) There would be a learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” with a high level of quality.

2) The learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” could help the community members to repair the electrical appliances.

3) The research could be used as a model to develop a learning resource for other communities.

III. RESEARCH SCOPE

A. Contents in the Learning Resource

The contents in the learning resource for the Saleng community under the bridge of zone 1 entitled “How to Repair Electrical Appliances” would include an introduction to the tools for repair, how to repair fans, irons, rice cookers according to the conditions, and how to maintain electrical appliances. There would also be an exhibition about the history of the community and the leaflets with information of the community.

B. Population and Sampling Group

The population in this study consisted of 500 persons who live in the Saleng community under the bridge of zone 1 [3].

The sampling group in this study consisted of 30 persons who worked as itinerant junk buyers or Saleng pedal cart riders. They were chosen using specified sampling method out of those who previously participated in the research entitled “Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers”.

C. Expert Panel

The expert panel consisted of those who have skills and knowledge about how to assess the development of the learning resource for the Saleng community under the bridge of zone 1. They were chosen using specified sampling method. They had qualifications and they were voluntary to be in the expert panel of this research.

1) Experts in media: Experts in media ere those with at least a Masters degree or at least five years of working experiences. They have knowledge, experiences and skills in the development of learning resources, media production, publishing industry, print design, colour scheme and proofreading in order to assess the quality of the learning resource for the Saleng community under the bridge of zone 1 in terms of media. There were three experts in media.

2) Experts in contents: Experts in contents were those with at least a Masters degree or at least five years of working experiences. They have knowledge and understanding about steps and procedure to repair electrical appliances, which in this case are fans, irons, and rice cookers, and about the history of the community. There were three experts in contents.

IV. TOOLS IN THIS RESEARCH

The tools used in this research for the development of the learning resource for the Saleng community under the bridge of zone 1 were as follows:

A. The Learning Resource

- Video clips entitled “How to Repair Electrical Appliances”
- Posters about how to maintain electrical appliances
- Exhibition about the history of the community
- Leaflets about the information of the community

B. The Quality Assessment Form

- The quality in terms of contents with Likert’s 5-rating scale
- The quality in terms of media with Likert’s 5-rating scale

C. The Satisfaction Evaluation Form

- The satisfaction evaluation form for the sampling group with 5-rating scale

V. Procedure for Tool Development

A. The Development of the Learning Resource

The development of the learning resource for the Saleng community under the bridge of zone 1 began with (1) the identification of media format and genre and (2) the specification of the learning resource based on the demands of the community. The outcome was that the community would like to see the learning resource as a way to learn more entitled "How to Repair Electrical Appliances".

B. Research Plan and Development

1) The research objectives of the learning resource for the Saleng community under the bridge of zone 1 were identified.

2) The research scope and the preliminaries were determined.

3) The research duration was set.

C. Development according to ADDIE Model

1) Stage 1 Analysis: The learning resource for the Saleng community under the bridge of zone 1 was aimed to address the need of the community. Therefore, the demands of the community needed to be analysed and surveyed in terms of the topics, the information of the community and the location for the learning resource. According to the resolve from the analysis and survey, the community members would like to have the topics to cover how to fix fans, irons and rice cookers. As for the information about the community, according to the interview with the community committee and Toong-kru Office, it was agreed that Dutdao-Learning Centre Building would be used as the location for this research.

2) Stage 2 Design: The data were gathered and summarised in the form of dummy table which included the topics in the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances". Afterwards, it would be assessed by the expert panel and revised accordingly.

3) Stage 3 Development: After a conclusion was reached, many drafts would be drawn about the contents. The data and the articles as well as the information of the community would be gathered in order to create video clips, posters, exhibition material and leaflets. Another dummy table would be drawn and assessed by the expert panel before the implementation in the next stage.

4) Stage 4 Implementation: The software was chosen for the development of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances". Many applications were needed and these included software for magazine publication, image editing, graphic design, video editing and digital magazine before we could develop the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances".

5) Stage 5 Evaluation: The expert panel was invited to the learning centre of the Saleng community under the bridge of zone 1 in order to assess the learning resource entitled "How to Repair Electrical Appliances" in terms of media production, contents and quality. They were six experts in total. The learning resource would be revised according to the suggestions and comments by the expert panel in order to improve the quality, contents and media production.

D. Test or Experiment

The learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" was used by 30 itinerant junk buyers who were chosen using specified sampling method. Afterwards, they filled out the questionnaire and the data would be analysed.

E. Analysis

The data would be analysed to improve from the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances".

VI. Research Results

The learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" was developed and contained the following aspects: video clips, posters, leaflets and exhibition. These were shown in the figures below.

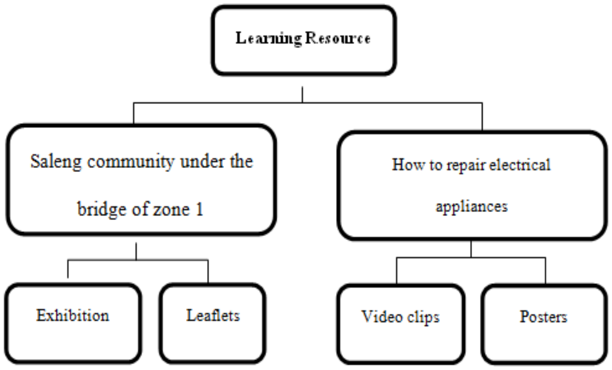

Learning Resource
Saleng community under the bridge of zone 1
How to repair electrical appliances
Exhibition
Leaflets
Video clips
Posters

Fig. 1. The learning resource

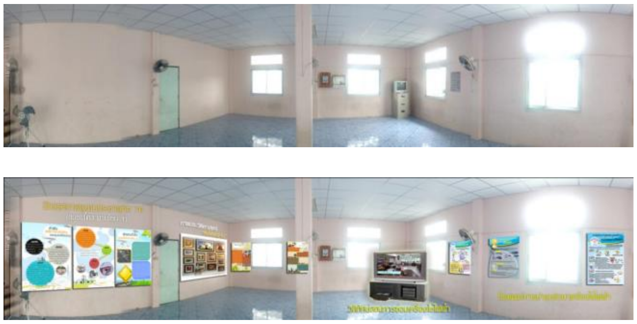


Fig. 2. The learning resource center before and after the development

A. Quality Assessment by the Expert Panel

The quality of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" could be summarized as below:

TABLE I. THE QUALITY ASSESSMENT BY THE PANEL OF EXPERTS IN CONTENTS AND EXPERTS IN LEARNING RESOURCE DEVELOPMENT

Item	Quality Assessment		
	Mean	***S.D.***	***Meaning***
1. Contents	4.83	0.29	Very Good
2. Learning Resource Development	4.47	0.63	Good

According to Table I, the quality assessment in terms of contents was 4.83 with S.D. of 0.29, or at a very good level. The quality assessment in terms of learning resource development was 4.47 with S.D. of 0.63, or at a good level.

B. *Sampling Group's Satisfaction*

The sampling group in this research consisted of 30 itinerant junk buyers who were chosen using specified sampling method who previously participated in the research project entitled "Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers".

TABLE II. THE SAMPLING GROUP'S SATISFACTION TOWARDS THE LEARNING RESOURCE

Item	Quality Assessment		
	Mean	***S.D.***	***Meaning***
1. Video clips	4.68	0.72	The Highest
2. Posters	4.71	0.63	The Highest
3. Leaflets	4.63	0.70	The Highest
4. Exhibition	4.79	0.50	The Highest
Average	**4.70**	**0.64**	**The Highest**

According to Table II, , the sampling group's satisfaction towards the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" was 4.70 on average with S.D. of 0.64, or at the highest level.

VII. RESEARCH DISCUSSIONS

According to the research results, the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" contained video clips, posters, leaflets and exhibition in order to increase the understanding entitled "How to Repair Electrical Appliances" which included fans, irons and rice cookers. There was also information about the community. These results could be discussed as follows.

A. *The Assessment by the Experts in Contents*

The quality of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" in terms of contents was 4.83 on average with S.D. of 0.29, or at a very good level. This was because the contents in the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" were accurate and suitable for the community. This comment is in accordance with the research study entitled "Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok" [1]. The contents of this learning resource also complied with the research study entitled "Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers" [2] which was approved at a national level. Therefore, the contents were accurate and complete.

B. *The Assessment by the Experts in Learning Resource Development*

The quality of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" in terms of learning resource development was 4.47 on average with S.D. of 0.63, or at a good level. To accomplish this result, the researchers applied theories about how to create video clips and how to choose the right angle for each context. There were also principles in graphic design and digital publishing and theories about how to run an exhibition in terms of layout and material presentation. The aspect of video clips in this research was in accordance with the research study by Petchthai Kerdchote and Sopa Meeyai [4] who developed the video clips to increase the public relations of the Department of Educational Communications and Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi and found that their video clips were of good quality and could be used for intended purpose. As for the printing aspect, the quality of the learning resource in this research was at a good level, similar to the research by Weeranan Homdok [5] who researched the publishing set for the public relations of the Masters degree in Educational Technology and Mass Media, because it followed the right approach to publication. Moreover, the quality of the exhibition in this research was at a good level and it was in accordance with the principle of Payungsak Prachusilpa [6] about the exhibition design so that the exhibition could be designed to meet the objectives of the sampling group.

C. *The satisfaction evaluation*

The sampling group of 30 persons showed the highest satisfaction towards the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" with mean score of 4.70 and S.D. of 0.64. This was because the contents of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" were accurate and suitable for the community. The contents would involve the information of the community and how to repair electrical appliances in a step-by-step approach, which is similar to the principle by Narumon Kitpaisarnrattana [7]. To analyse the satisfaction, the data were analysed using SPSS for Windows (Statistical Package for Social Sciences for Windows) and this is similar to the research by Jittima Puttacharoen [8] who examined the learning achievement and the satisfaction towards the webpage contents and found that the sampling group showed the highest level of satisfaction.

VIII. SUGGESTIONS

A. *Suggestions for application*

1) The learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical

Appliances" could be adopted for the self-study approach aimed at adults outside the formal education system or those without an opportunity to study the specialized discipline.

2) The learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" could be distributed on an online social network to gain broader usage.

B. Suggestions for further research

1) There should be research into the development of the learning resource for the Saleng community under the bridge of zone 1 about how to repair cars and household accounting.

2) There should be a study into the factors which affect the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" in an online social network environment.

3) The posters could be developed to teach skills in a step-by-step manner.

ACKNOWLEDGMENT

The development of the learning resource for the Saleng community under the bridge of zone 1 entitled "How to Repair Electrical Appliances" was granted research funding from Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi to support the extension from the previous research with the Saleng community under the bridge of zone 1.

REFERENCES

[1] The Research Committee for the Community Research Project, Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok, Photocopied Document, 2011, p. 4.

[2] The Working Team from Faculty of Industrial Education and Technology, Research Project Entitled "Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers", Bangkok, 2013, pp. 4-5.

[3] Head of the Community under the Bridge of Zone 1, Population in the Community under the Bridge of Zone 1, Bangkok, Interview, 12 February, 2013.

[4] P. Kerdchote and S. Meeyai, Research Project Entitled "The Development of Video Clips for Public Relations of the Department of Educational Communications and Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, 2007, p. B.

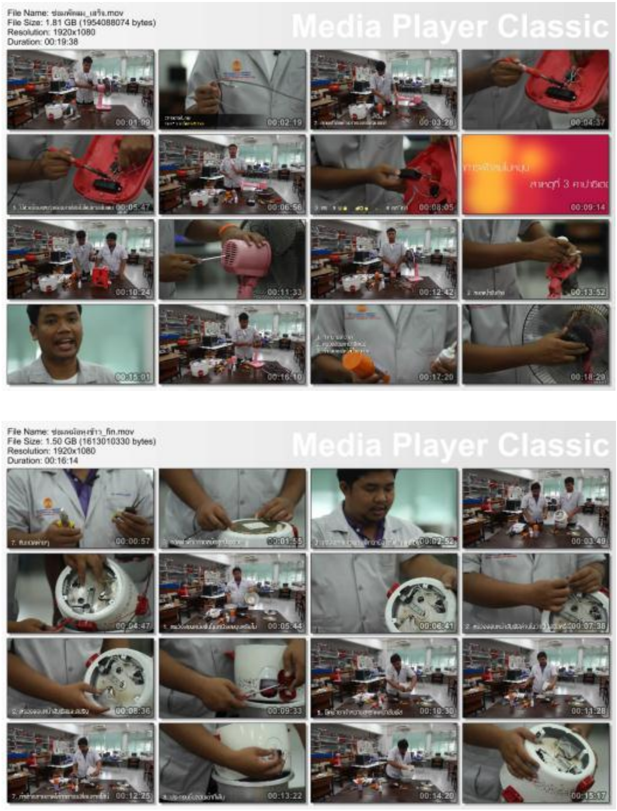
[5] W. Homdok, Publishing Set for the Public Relations of the Masters Degree in Educational Technology and Mass Media, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, 2010, p. B.

[6] P. Prachusilpa, Design for Exhibition, Compact Design, Bangkok, 1992, p. 1.

[7] N. Kitpaisarnrattana, Users' Satisfaction towards 7 Types of Service in Chulalongkorn University Library, a Master's thesis in Communication Arts, Chulalongkorn University, 2002, p. B.

[8] J. Puttacharoen, Learning Achievement and Satisfaction towards Web Pages with Different Contents and Presentation Styles, a Master's thesis in Education, Kasetsart University, 2000, p. B

APPENDIX



File Name: ซ่อมพัดลม_เสร็จ.mov
File Size: 1.81 GB (1954088074 bytes)
Resolution: 1920x1080
Duration: 00:19:38

File Name: ซ่อมหม้อหุงข้าว_fin.mov
File Size: 1.50 GB (1613010330 bytes)
Resolution: 1920x1080
Duration: 00:16:14

Fig. 3. Video clips entitled "How to Repair Electrical Appliances" and "How to Fix Fans"

Image fusion approach with noise reduction using Genetic Algorithm

Gehad Mohamed Taher
Faculty of computers and informatics
Suez Canal University
Ismailia, Egypt

Mohamed Elsayed Wahed
Faculty of computers and informatics
Suez Canal University
Ismailia, Egypt

Ghada El Taweal
Faculty of computer and informatics
Suez Canal University
Ismailia, Egypt

Ahmed Fouad
Faculty of computer and informatics
Suez Canal University
Ismailia, Egypt

Abstract— **Image fusion is becoming a challenging field as for its importance to different applications, Multi focus image fusion is a type of image fusion that is used in medical fields, surveillances, and military issues to get the image all in focus from multi images everyone is in focus in a different part, and for making the input images more accurate before making the fusing process we use Genetic Algorithm (GA) for image de-noising as a preprocessing process. In our research paper we introduce a new approach that begin with image de-noising using GA and then apply the curvelet transform for image decomposition to get a multi focus image fusion image that is focused in all of its parts. The results show that Curvelet transform had been proven to be effective at detecting image activity along curves, and increasing the quality of the obtained fused images. And applying the mean fusion rule for fusing multi-focus images gives accurate results than PCA, contrast and mode fusion rule, Also, GA shows more accurate results in image de-noising after comparing it to contourlet transform**.

Keywords—Multi-focus image fusion; Curvelet transform; genetic algorithm Introduction

I. INTRODUCTION

The driving forces in today‘s manufacturing environments especially with recent rapid developments in the field of sensing technologies are quality improvement and cost reduction. The quality of many raw materials, parts, and products can be measured by visual inspection. However, the Inspection by eye is costly, subjective, qualitative, inaccurate, eye-straining, and time consuming. For high speed and real time applications, manual inspection is not possible. The result of the use of these applications is a great increase in the amount of data. As the volume of data grows, the need to combine data gathered from different sources to extract the most useful information also increases. The technique which performs this is Image fusion that is widely recognized as an important tool for improving performance in image based applications such as remote sensing, machine vision, medical imaging, and optical microscopy and so on.

Image fusion is a process of combining set of images to integrate complementary and redundant information to provide a composite image which could be used to better understanding of the entire scene and will be more informative and complete than any of the input images. When a lens focuses on a subject at a certain distance, all subjects at that distance are sharply focused. Subjects not at the same distance are out of focus and theoretically are not sharp. It is often not possible to get an image that contains all relevant objects in focus. One way to overcome this problem is multi-focus image fusion, in which one can acquire a series of pictures with different focus settings and fuse them to produce an image with extended depth of field.

In the past years Genetic algorithm was used with image fusion for solving optimization problems such as used for estimating the weights of the weighted average Pixel level weighted average [4], then GA used for solving optimization problems for image fusion in another manner in which it is used for determining the best size of the block in [5], Michifumi Yoshioka presented an approach based on genetic algorithm for minimizing noise from original image. Most of algorithms proposed in literature are either noise dependent or threshold governed. In real time environment the type of noise in the image signal is unknown. So applying an algorithm specific to noise, will never be successful under these conditions. These disadvantages can be reduced by using a hybrid filter that consists of de-noising filters [1], In recent years image fusion had been used for many applications and many techniques had been used for achieving this one of the most popular techniques was Pyramid [7] and wavelet [6] are the most widely studied and used multi-resolution image fusion schemes. There are many types of pyramid and wavelet decomposition algorithms in recent years; however, not much research has been conducted on fusion rules. wavelets transform can only reflect "through" edge characteristics, but cannot express, To overcome the limitation of the wavelet transform, Donoho et al. has proposed the concept of Curvelet transform, which uses edges as basic elements, possesses maturity, and can adapt well to the image characteristics. “Along edge characteristics” [8]. Moreover, curvelet Transform has anisotropy and has better direction than wavelet can provide more information to image processing [9] [10].

We introduce in this paper a multi focus image fusion using curvelet transform and image de-noising using Genetic algorithm (GA). The basic idea is to apply the GA to filter sequence to get the best filter sequence for de-noising the input images. And then using the curvelet transform as a tool for fusing the two images to get a more focus fused image.

The paper is organized as follows. Section 2 explains the proposed fusion approach and an introduction of using GA for image de-noising. Section 3 the experimental results of the proposed approach, performance analysis showing the results of applying the quality measures and its graph and comparison with other schemes then it is followed by conclusion in section 4.

II. Proposed Image Fusion Approach

In our proposed approach we first used Genetic Algorithm (GA) as a de-noising tool, Then this de-noised image which comes from applying GA to the Hybrid filter entered to the next stage which is applying the image fusion process using curvelet transform to the two input images to obtain one fused image which is better in it's focusing from the other two input images.

Block diagram for our proposed approach is illustrated in (Fig. 1)

A. Image de-noising using GA

A variety of algorithms have been evolved from nature. GA is one of the simplest and most popular evolutionary algorithms. Genetic Algorithms called as (GA) are based on natural selection discovered by Charles Darwin. GA makes use of the simplest representation, reproduction and diversity mechanism. Optimization with GA is performed through natural exchange of genetic material between parents. Offspring's are formed from parent genes. Fitness of offspring's is evaluated. The fittest individuals are allowed to breed only. GA are being used in different applications such as function Optimization, System Identification and Control, Image Processing, Parameter Optimization of Controllers, Multi-Objective Optimization, etc.

Hybrid filter is a sequential filter where different filters are arranged in a sequence to obtain a noise free image. Peak Signal to Noise Ratio (PSNR) is one of the performance indices which determine the quality of the image.

Here we used PSNR as our fitness function for the GA which is directly proportional to the value of PSNR. Better the value of PSNR better is the quality of image.

Initial population:

De-noising using GA begins with initial population P0 with size µ and number of genes (filters) in a chromosome with size gnum, so we have the initial population is a matrix µ * gnum.

Selection: After applying the fitness function to the initial population we apply the selection function to select the highly fitness function chromosomes to be used for the next new generation and here we use the roulette wheel selection type for reproducing the new generation.

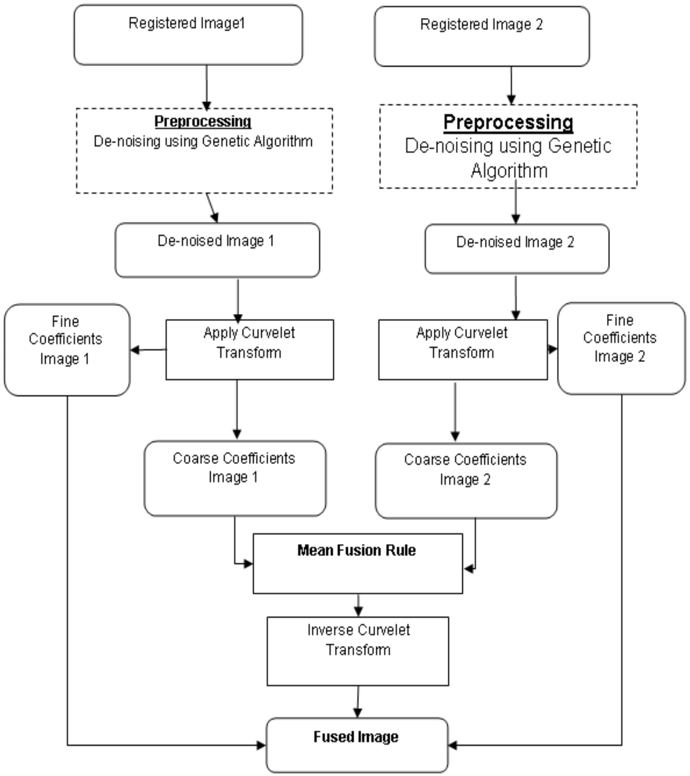

Registered Image1
Registered Image 2
Preprocessing
De-noising using Genetic Algorithm
Preprocessing
De-noising using Genetic Algorithm
De-noised Image 1
De-noised Image 2
Fine Coefficients Image 1
Apply Curvelet Transform
Apply Curvelet Transform
Fine Coefficients Image 2
Coarse Coefficients Image 1
Coarse Coefficients Image 2
Mean Fusion Rule
Inverse Curvelet Transform
Fused Image

Fig. 1. Block diagram for image fusion using Curvelet transform and de-noising using GA

Fitness Function:

The problem objective function can be defined as follows:

$$\text{Objective} = \text{Max}\ (f) \qquad (1)$$

Here PSNR can be defined as:

$$\text{PSNR} = 10\ \text{X}\ \log_{10}\left(\frac{(2552 \times M \times N)}{\sum\sum (x(i,\ j) - y(i,j))^2}\right) \qquad (2)$$

Where *y (i, j)* in PSNR can be defined as:

$$y\ (i,\ j) = \underbrace{I_5(W_5 * [\ldots \overbrace{[I_2(W_2 * \overbrace{[I_1(W_1 * y_1)]}^{(1)})]}^{(2)}]])}_{(5)} \qquad (3)$$

The sequence of the equation starts from 1 and ends by 5

y_1 is the initial corrupted image and * represents convolution.

3- W_k is the filters applied and I_k is the Boolean operators. Where k varies as $1 \leq k \leq 5$.

$$I_k(W_k * y_k) = \left\{\begin{array}{ll} y_{k+1}, & \text{if } I_k = 1; \\ y_k, & \text{if } I_k = 0. \end{array}\right\} \qquad (4)$$

In above equation, I_k=0 will imply that no convolution will take place and I_k=1 will imply that image y_k will be convoluted with the filter W_k to give a new image y_{k+1}.

Constraints:

$I_k \geq 0$ and $W_k \geq 1$ where $1 \leq k \leq 5$

$I_k \in \{0, 1\}$, $W_k \in [1, 5]$

W_k can be Mean, Contourlet, Average, Pyramid and Gaussian filters depending on the value W_k from 1 to 5.

Crossover: After individuals are selected, reproduction involves crossing the individual's chromosomes to produce their offspring's chromosome. Crossover is a random process, we use a single point crossover by choosing Pc where

$0 \leq Pc \leq 1$.

Mutation: By mutation individuals (chromosomes) are randomly changed. These variations (mutation steps) are mostly small. They will be applied to the variables of the individuals with a low probability Pm where $0 \leq Pm \leq 1$.

(Fig.2.) Shows the flow diagram of how GA used to de-noise a corrupted image Firstly, corrupted image and the smoothing filters are passed as an input to the GA function. GA analyses the system quality by comparing the values of the fitness function obtained by various sequences. GA uses SNR or PSNR as the fitness function for evaluating the best sequence of smoothing filters. After the completion of the first iteration, new set of sequences are created by the process of crossover and mutation. Mutation operator is used to avoid the local minima trapping of the algorithm. The probability of selection of a sequence from the set is directly proportional to the value of its fitness function. The new set of sequences then replaces the previous set. The process continues until the stopping criterion is achieved. The sequence, that gives the maximum value of SNR or PSNR, is said to be the best sequence. This sequence is passed as input to the Sequence Application Function. Sequence Application Function applies the filters on the corrupted image in that sequence. The resultant image is the noise removed image.

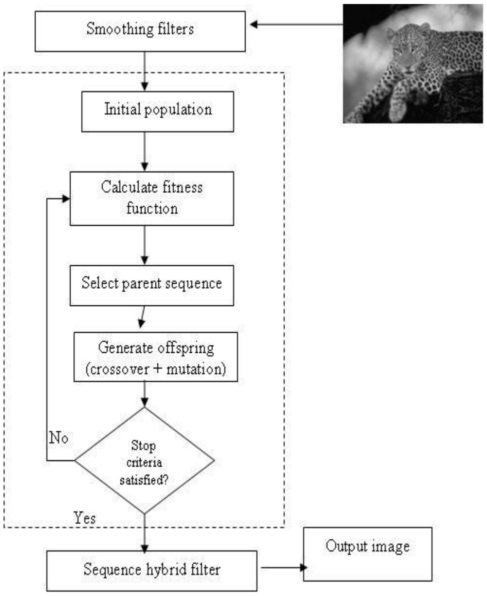


Fig. 2. Image De-noising using Genetic Algorithm (GA) applied to Sequence Hybrid Filter

B. Image Fusion by Curvelet Transform:

1) Curvelet Transform

Curvelet transform is a tool for representation of curved shapes in images. The concept of curvelet transform is based on the segmentation of the whole image into small overlapping tiles and then applying ridgelet transform on each tile.

Here we are using wrapping algorithm based curvelet decomposition.

The wrapping discrete curvelet transform is implemented using the following steps:

Step 1: FFT of the image is taken and the resulting Fourier samples is divided into collection of digital corona tiles as shown in "Fig. 3".
Step 2: For each corona tile, the tile is translated to the origin.
Step 3: The parallelogram shaped support of the tile is wrapped around a rectangle centred at the origin.
Step 4: The Inverse FFT of the wrapped support is determined and finally the resulting curvelet array is added to the collection of curvelet coefficients.

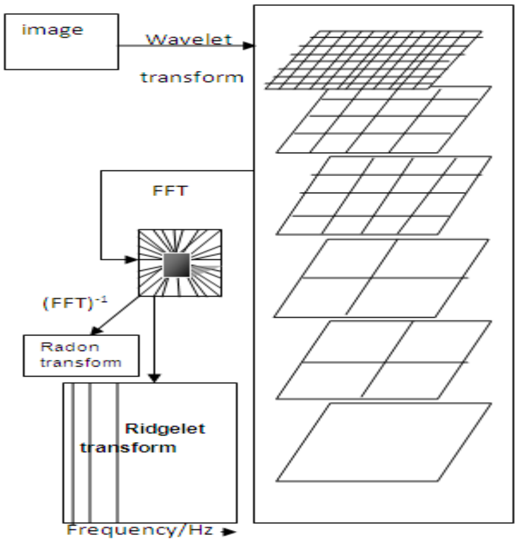


Fig. 3. Curvelet Transform

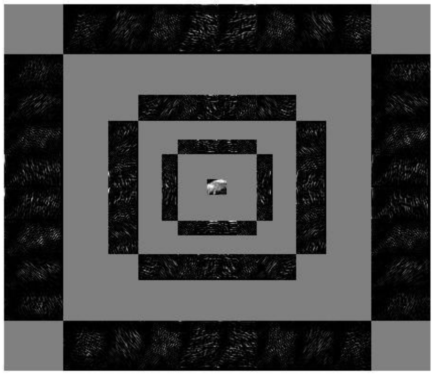


Fig. 4. Curvelet Coefficients of Tiger image

2) *Fusion Rules*

There are a variety of techniques that have been reported as valid image fusion processes. Some of these are Statistics based and Wavelet based.

Some of the popular fusion techniques based on statistical analysis of the images are max or min and mean, Principle Component Analysis (PCA) and contrast.

Assuming that images are collected simultaneously with accurate registration, images can be fused element wise, taking the maximum, the minimum, and the mean values.

The figures illustrated below show the registered and fused images using different fusion techniques like max or min, mean, Principle Component Analysis (PCA), contrast and wavelet based.

In our approach we use the popular mean fusion rule, as by applying the quality measure PSNR (Peak Signal to Noise Ratio) it gives the highest PSNR (Peak signal to noise Ratio).

III. Expermintal Results

Techniques for performing image de-noising and image fusion vary widely depending on the specific application, imaging modality, and other factors there is currently no single de-noising filter that can de-noise all types of noises and there is no single fusion method that yields acceptable results for all types of applications. The present research work proposes an approach that is more general and can be applied to a variety of image data. The performance of the proposed research work was analyzed using various experiments.

This section presents the experimental results obtained during performance analysis.

A. *Data Set*

The proposed approach was tested with six pairs of images (Fig. 5.). Each image is used as a representation of different scenes. All set of images represent the situation where, due to the limited depth-of-focus of optical lenses in cameras, it is not possible to get an image which is in focus everywhere.

Objective image quality measures play an important role in various image processing applications. There are different types of object quality or distortion assessment approaches. The fused images are evaluated, taking the following parameters into consideration.

Seven quality measures were used during experimentation to evaluate the efficiency of the proposed approach of image fusion using curvelet. They are Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR), Normalized Absolute Error, Normalized Cross Correlation, Maximum Difference, Average Difference and Structural Content.

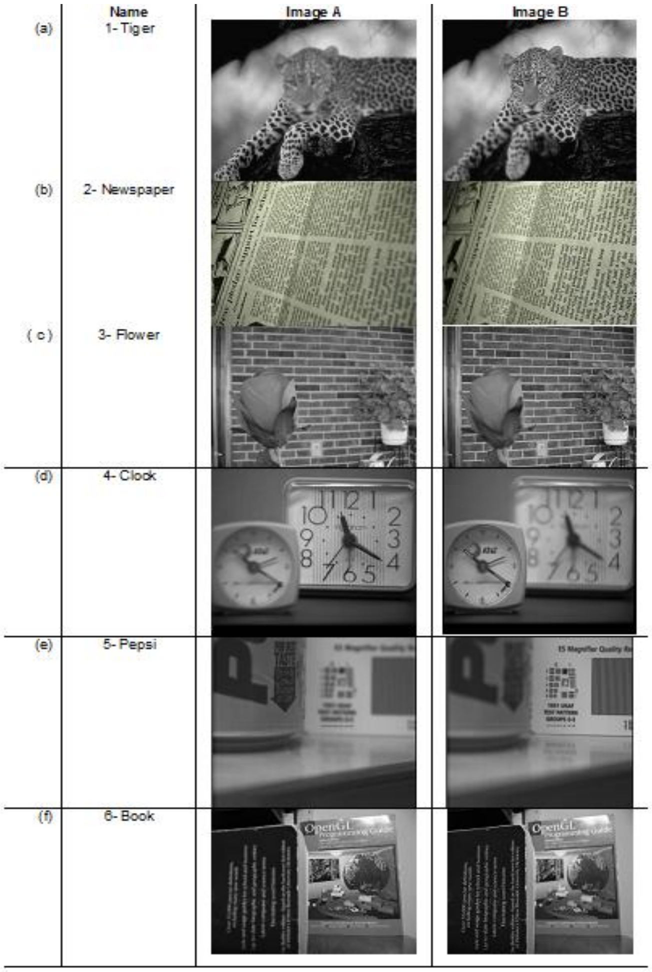


Fig. 5. Test Images

The following table is for the parameters used when using the GA:

TABLE I. Parameter setting for De-noising using GA

Parameters	Definition	Values
µ	Population size	25
Pc	Crossover probability	0.4
Pm	Mutation probability	0.01
Itrnum	Number of iterations	20
Gnum	Number of genes (filters)	5

A. *Performance Analysis*

The visual results of applying our approach are illustrated in the following figure:

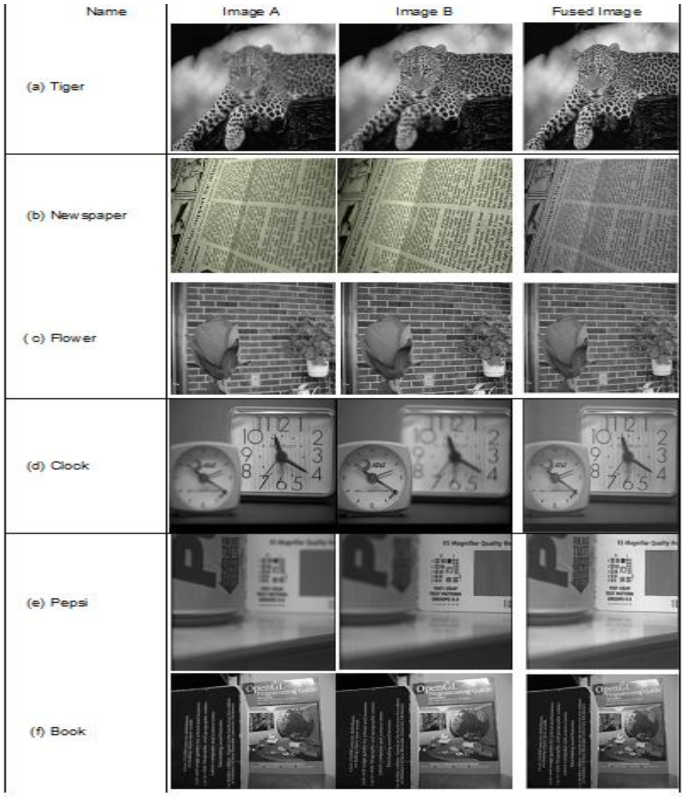


Fig. 6. Visual results of the approach

By applying the quality measure PSNR to set of images these are the result of the image de-noising using the GA which show that GA gives the best PSNR value as it is an optimization function. The next table [Table 2.] gives the value of the PSNR after each iteration by applying 20 iterations to the clock image, we found the best PSNR was 38.7426 and it was stable after iteration number 9 till the 20 iteration.

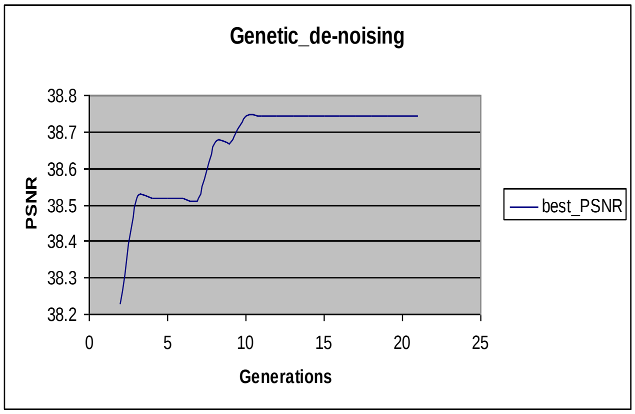


Fig. 7. Genetic performance using PSNR

TABLE II. THE BEST PSNR ALONG 20 ITERATIONS USING GA

Num_Iterations	Best_PSNR	
	1	38.2302
	2	38.5201
	3	38.5201
	4	38.5201
	5	38.5201
	6	38.5201
	7	38.6701
	8	38.6701
	9	38.7426
	10	38.7426
	11	38.7426
	12	38.7426
	13	38.7426
	14	38.7426
	15	38.7426
	16	38.7426
	17	38.7426
	18	38.7426
	19	38.7426
	20	38.7426

By applying quality measures to the set of dataset images these are the results of the image fusion approach using curvelet and mean fusion rule with RMSE, PSNR and Maximum Differenece.

TABLE III. THE RESULT OF APPLYING RMSE, PSNR, MAXIMUM DIFFERENCE TO DATASET

Image	RMSE	PSNR	Maximum Difference
1-Tiger	119.5014	27.3571	80.8271
2-Newspaper	1.4926e+003	16.3913	199.3773
3-Flower	86.6267	30.8525	53.4359
4-Clock	94.4028	28.3810	72.6575
5-Pepsi	20.1524	35.0875	34.6182
6-Book	38.6760	32.2564	83.0727

By applying Quality measures to different fusion rules for the six set of images these are the result tables and its graph:

1) The first table is for the Pepsi Image by applying the Fuse mode, PCA, contrast and fuse mean using curvelet transform, and its graph of these results.

Pepsi Image

TABLE IV. THE RESULT OF APPLYING DIFFERENT FUSION RULES TO PEPSI IMAGE USING DIFFERENT EVALUATION FUNCTIONS

	MSE	PSNR	MD
Fuse mod	323.7387	23.0289	83
PCA	242.8457	24.2775	98
Contrast	65.0607	29.9976	91.8697
Fuse mean	53.4359	30.8525	86.6267

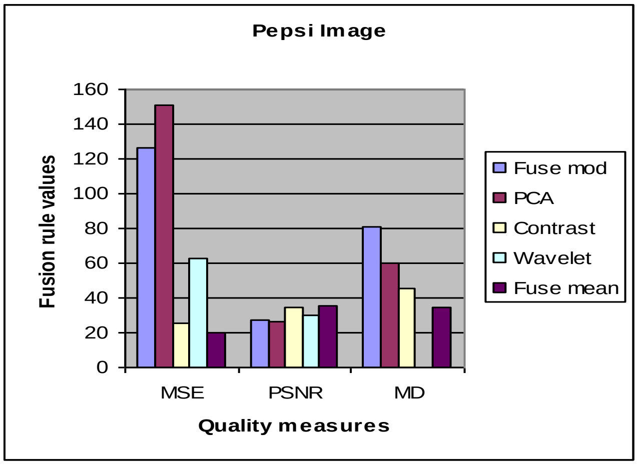


Fig. 8. Graph of table 4

2) The first table is for the Tiger Image by applying the Fuse mode, PCA, contrast and fuse mean using curvelet transform, and its graph of these results.

Tiger Image

TABLE V. THE RESULT OF APPLYING DIFFERENT FUSION RULES TO TIGER IMAGE USING DIFFERENT EVALUATION FUNCTIONS

	MSE	**PSNR**	**MD**
Fuse mod	174.9141	25.7026	87
PCA	179.1189	25.5994	90
Contrast	122.7919	27.2351	80.633
Fuse mean	119.5014	27.3571	80.8271

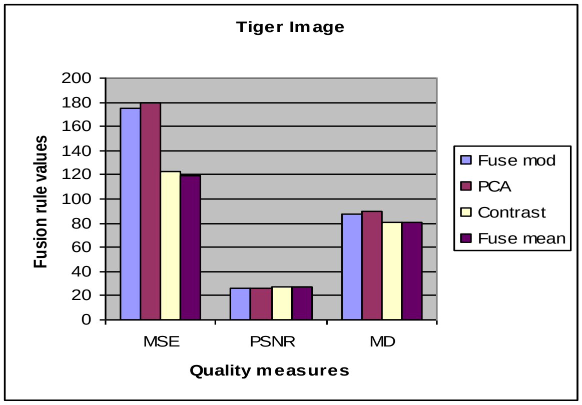


Fig. 9. Graph of table 5

3) The first table is for the Pepsi Image by applying the Fuse mode, PCA, contrast and fuse mean using curvelet transform, and its graph of these results.

Flower Image

TABLE VI. THE RESULT OF APPLYING DIFFERENT FUSION RULES TO FLOWER IMAGE USING DIFFERENT EVALUATION FUNCTIONS

	MSE	PSNR	MD
Fuse mod	126.7544	27.1012	81
PCA	150.8274	26.346	60
Contrast	25.1908	34.1184	45
Fuse mean	20.1524	35.0875	34.6182

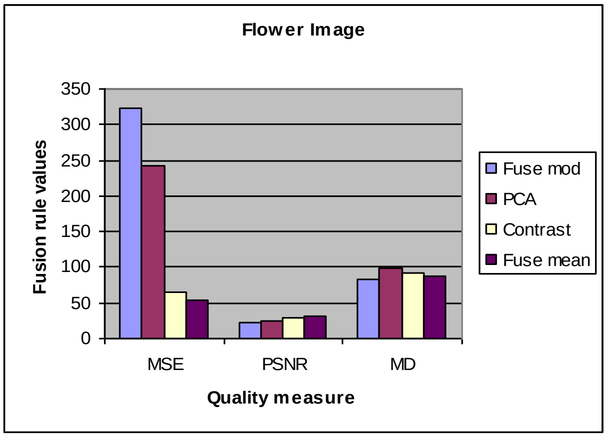


Fig. 10. Graph of table 6

B. Comparison with Other Schemes

From the tables of the previous section, The fusing of the images using curvelet and mean average fusion rule, we see that by applying the MSE for all the fusion rules the best MSE that give the low value which is the fusion mean, and for the PSNR the best value is the one that give the greater PSNR which is also fusion mean, but for the last one which is MD the best one is the one with the smaller value and it is the second one on this.

The below figures show the difference of the result of applying mean average and the wavelet for fusing the two images of the clock image.

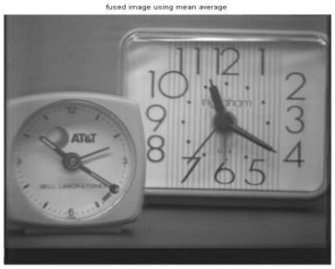


Fig. 11. Fusing using curvelet with mean average

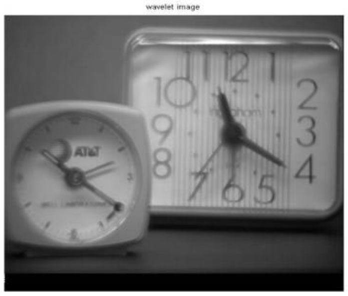


Fig. 12. Result of fusing using wavelet

Contourlet Transform that is seen as a discrete form of a particular curvelet transform had been used as a tool for image de-noising and it had been shown that it a good tool for this.

Now comparing the GA and the contourlet for image de-noising and using the PSNR as a quality measure to get the one that is better we had been given that the best PSNR is for GA which gave 38.7426 and contourlet gave 33.5483 by these results as an example for applying it to the Clock picture Image A and gave 35.0965 by applying contourlet for clock picture Image B.

IV. Conclusion

In this paper we present a new approach by applying GA as a de-noising process, and showed that it is a much more benefit as a de-noising techniques than the other techniques that we used for comparison, and then used image fusion using curvelet transform using mean fusion rule as a much more good method for fusing than other fusion methods and rules and applied to two grey scale images.

V. Future Work

The future scope for this approach is using it with one grey scale image and one colored image and get the same result that is quite better than any others. And further future work we can use other datasets and analysis.

References

[1] Siddharth Gupta, Rajesh Kumar, S. K. Panda, "A Genetic Algorithm Based Sequential Hybrid Filter for Image Smoothing," International Journal of Signal and Image Processing, Vol.1 2010/Iss.4 pp. 242-248.

[2] Y. Kiran Kumar, "Comparison Of Fusion Techniques Applied To Preclinical Images: Fast Discrete Curvelet Transform Using Wrapping Technique & Wavelet Transform," Journal of Theoretical and Applied Information Technology © 2005 – 2009 JATIT.

[3] J.L. Startck, E.J. Candes, D.L. Donoho, "The curvelet transform for image denoising," IEEE Transactions on Image Processing, Vol.11 (6), 2002, pp.670-684.

[4] Vjyothi,B.Rajesh Kumar,P.Krishna Rao,D.V.Rama Koti Reddy, " IMAGE FUSION USING EVOLUTIONARY ALGORITHM (GA)," Int. J. Comp. Tech. Appl., Vol 2 (2), pp.322-326, 2011.

[5] A. Angeline Nishidha, "Multi-Focus Image Fusion using Genetic Algorithm and Discrete Wavelet Transform," International Conference on Computing and Control Engineering (ICCCE 2012), pp.12 & 13 April, 2012.

[6] L. A. Ray and R. R. Adhami, "Dual tree discrete wavelet transform with application to image fusion," Southeastern Symposium on System Theory, pp. 430-433, 2006.

[7] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Transactions on Communications, vol. 31, no. 4, pp. 532-540, 1983.

[8] Sandeep, Yash Kumar Sharma, and Mahua Bhattacharya, "Curvelet Based Multi-Focus Medical Image Fusion Technique: Comparative Study With Wavelet Based Approach," International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), Worldcomp, July, Las Vagas USA , 2011.

[9] G.Pajares, J.M.Cruz, "A wavelet-based image fusion tutorial," Pattern Recognition, vol.37 no.9 pp.1855-1872, 2004.

[10] D.L.Donoho M.R.Duncan, "Digital Curvelet transform Strategy, implementation and experiments," SPIE vol. 4056 pp.12-29, 2004

Security of Mobile Phones and their Usage in Business

Abdullah Saleh Alqahtani
School of Computer Science, Engineering and Mathematics,
Faculty of Science and Engineering, Flinders University,
Adelaide SA 5001, Australia

***Abstract*—The purpose of this document is to provide an overview on the growth on mobile phone and PDA devices and use in business-oriented modern day life style. The explosion of smartphones in enterprise and personal computing heighten the concerns of security and privacy of users. Now a day's use of mobile is in every walk of life like shopping, trading, paying bill and even using internet banking. But with these facilities some draw backs are also there. Recent studies have shown that applications can host new types of malware. This discussion explores smartphone security through several research works and how user himself can avoid from data hacking and other insecurities.**

Keywords—Smartphone; PDA; security; business-oriented; applications

I. Introduction

The main intention of this manuscript is to give an impression on the growth of mobile phone and PDA devices, which are now indispensable for our business-oriented modern day lifestyle. This document also put emphasis on those technology risks that are associated with using these devices as well as available safeguards to diminish any problems. This information will help organizations to improve their levels of security and to decrease such occurrences concerning the use of these handheld wireless devices. It can be envisioned that the more mobile devices continue to accelerate, the more sophisticated applications can be predicted. [9] Internet connectivity has helped accelerate the tendency of mobile phones from 'voice-centric to data-centric' networking systems. These two worlds are gradually converging to support each other. [7]

Since security concerns have supreme importance in fiscal transactions or a mobile payment, that's why it is important that the attacker present in this application are examined by taking a holistic view of the vulnerabilities. [6] The increasing demand and functionalities of the hi-definition Smartphone category of mobile devices will make an attractive target for malware writers and infiltration of malware on mobile devices can raise serious business and safety concerns. [13] Moreover, Not sufficient security in wireless connection may tempt to numerous unlawful attackers including but not limited to hacking, fraud of system integrity, surveillance and loss or stealing of the device itself [14] In contrast, the security of handheld devices cannot be maintained without users' involvement. Users must be instructed about what measures to follow and what precautions to take while they use organization-issued equipment or a personally owned one. When taking full advantage of all the facilities afforded by cell phones or PDAs, it is important that the user should have a good knowledge of the security safeguards. [9]

On a related note the current economic environment and the advent of new technologies has interested organizations in availing the Cloud storage services provider model. Cloud storage providers can suggest cost-cutting measures by using equal storage capacity to meet organizations' needs to initiate transitional cost-savings measures for their customer base. [21] From 2010 on a majority of the cellular networks started undertaking switch-over from the existing third generation communications systems (3G) to the fourth generation systems (4G). This (4G) system mostly incorporates broadband IP-based heterogeneous multimedia services that let users use diverse networks on an anytime and anywhere basis. [22] The 4G system is a modified version of third generation mobiles and certainly is a 'must have' gadget for the organization. [18]

II. The Commonly Used Mobile Features

In today's high-tech business-oriented lifestyle cell (or mobile) phones and Personal Digital Assistants (PDAs) have turn into indispensable. For the most part these small and economical devices are being used for making calls, sending short text messages and supporting Enhanced Messaging Services (EMS). They can also be used for Personal Information Management (PIM), i.e. phonebook, calendar, and notepad, etc. These tiny devices can perform a whole array of functions that previously could only be done on desktop computers. Computers generally perform many functions such as sending and retrieving e-mails, web browsing, retrieving and modifying documents, delivering and making presentations, and can access available data from remote servers. Currently mobile device are now being equipped with built in devices like camera, GPS receiver, removable media card incision, and also are host to a vast assortment of wireless interfaces, including but not limited to, infrared, Wireless Fidelity (Wi-Fi), Bluetooth connectivity, and more than one type of cellular interfaces, etc. [9]

Research and advances in Information and Communication Technology (ICT) have resulted in today's high-end smart phones and Personal Digital Assistants (PDAs), which now possess fairly equal processing power and memory storage capacity that was previously the hallmark of Personal Computers (PCs). Today's smart phones converge with full-featured mobile phones in that they function much like computers. Users can now make phone calls and run applications, besides accessing and stockpiling useable data communications from shared networks and the Internet. In the meantime the memory cards of cell phones are now approaching the maximum of 8 GB capacity, which aims to provide adequate space to stockpile business information. Since highly developed mobile phones have the capacity to connect to the Internet and access websites and also have e-mail and multimedia set-ups, mobile devices are increasingly more "data-centric" compared to when they were traditionally "voice-centric" networks. [13]

Mobile manufacturing companies are taking the initiative to install those potential mobile enabling applications on mobile podiums, for instance Symbian™, OS, Microsoft™ and Windows™. Owing to the improved capacity, induction of new applications, available service support on cellular as well as local-area networks along with Bluetooth connectivity, the business community has an ever-increasing demand for mobile devices. These full-featured devices are improving workers' output by giving them ready access to the information they require. Even though these devices can boost competence and productivity, they also possess some inherent threats to organizations. What if the classified corporate and private data goes astray or a device is stolen? The types of threats that appear in the shape of spam, malware infections, and hacking will be discussed in the proceeding chapters. [13]

A. An overview of device transformation

The evolution of mobile telephony can be traced back to the continuous socio-historical development of the landline phone industry. One major reason for the acceptance of telephones at the beginning of the 20th century was "security". However, with the emerging American economy "business concerns" became yet another good reason for acquiring a landline phone. From1910 onwards the social use of the telephone became another factor for acquiring a landline phone. It is pertinent to mention here that the main disparity between the social setting of the early days of both landline and mobile telephones was the fear of breach of confidentiality, resulting from neighbors trying to 'snoop in' on their next door neighbors' conversations or business rivals abusing switch boards, etc.

Gradually, handheld devices began to appear in different shapes and sizes. The first cell or mobile phones appeared in the United State of America in 1978, after AT&T Company carried out its test communications under the auspices of the Federal Communications Commission in Chicago, Newark and New Jersey simultaneously. The device then had the weight and size of a brick and was restricted to voice communications only. From then on phenomenal improvements have been made in the appearance and performance of handsets, and the infrastructural capacity of their networking. By and large the capacities of these devices may vary but at the heart of technology there are certain similarities as well. In 1993 Apple introduced its very first PDA, "the Newton". Battery powered and compact in size PDAs are intrinsically designed for mobility as they stockpile a user's data in the form of solid-state memory rather than putting it on a hard disk. Nevertheless, certain vulnerabilities are accredited to the use of PDAs as preserved data in unpredictable memory may be lost and/or erased if the device is re-organized for some reason. In many ways, PDAs are akin to handheld (PCs) and are not used for telephony. Mobile phones on the other hand, are much like PDAs apart from an important disparity in that they support more than one radio interface to cellular telecommunications networks. Moreover, mobile phones have a different legacy as well in that mobile services have *enabled "follow me anywhere/always on" telephony"*.

What makes cell phones different from the rest of the handheld devices is their ability to communicate through cellular networks. In cellular networks - as their name suggests - cells play a pivotal role in reprocessing radio frequencies in a restricted radio band by allowing more and more calls to happen. Taking an example from the U.S.A. where a diversity of digital cellular networks is thriving which pursues dissimilar and incompatible sets of standards, it is evident that virtually everyone in that country is now covered by a digital network of some kind. [18] Two leading digital cellular network operating in the U.S.A are Code Division Multiple Access (CDMA) and Global System for Mobile Communications (GSm). Common cellular network are Time Division Multiple Access (TDMA) and integrated Digital Enhanced Network (iDEN). iDEN use proprietary protocol while other follow standardized open protocols. Last but not least a digital version of the ingenious parallel standards as designed for cellular telephone service also subsists in the field and to date has been referred to as recognized as Digital Advanced Mobile Phone Service (D-AMPS). [9]

B. Some common trends in mobile devices

The omnipresent exploitation of mobile systems, extensive use of the Internet as well as the speedy development in wireless technologies [18] in recent years is the contributing factors that have improved the functions and characteristics of mobile handheld devices. In particular, cell phones have witnessed those features which were once only accessible in high-end smart phones and gradually introduced into more basic phones. For instance, LCD screens have progressed from the colorless display to grayscale and finally to high resolution colored display technologies along with built-in cameras which, some time ago, were considered a scarcity. Today, however, they are just as ordinary feature. In the same way text messages have converted to chat messages and then multimedia messages and more enhanced messages, i.e. EMS and now ending with e-mail. At present mobile devices are expected to develop into more complex, powerful, communicative functions at high-speed and having better processing capabilities and greater bandwidth, apart from having the facility of "always on" connection such as those existing on desktop computers. As a consequence of the said developments, cell phones are becoming more of a depository for a wide range of private and organizational data and the core is changing to packet data as opposed to voice data. [9].

With the advent of broadband interactive multimedia applications in wireless devices the plethora of new trends and functionalities has overwhelmed certain device features. A few have been discussed in the preceding paragraphs and this does not include simple icon-menu, touch-pad, and artificial-intelligence-based natural languages. The said devices and communication systems have also developed especially designed software to curb any instances of encroachment, namely:

- Invasion detective system - detects intruders and takes defensive measures upon the information it has,
- Anti-virus software - restrains malicious codes and offers support for reliable servers and applications.
- Almost every individual device will perform as:
- personal organizer - alarms, clocks, calculators, time zones, flash lights, calendars, dictionaries, compasses, video and music player, pocket PCs with office-type applications (for example, Blackberry).
- Gauge meters - to measure temperature, air pressure, humidity, and heartbeat, etc.
- virtual keys - to secure ID cards, digital cash, tag-readers, remote control devices, pagers, locating sensitive devices, etc. [18]

In such devices there is still room for new applications by new invigorating platforms such as Android. Android is a relatively new and dynamic platform that possesses some of the most sophisticated phone applications and is theoretically in conflict with iPhone. Android is striving hard to launch as many operating systems as possible whilst iPhone intends to get the best users experience by imposing restrictions to its hardware and software standards. [19]

From 2010 a majority of the cellular networks started switching over from the existing third generation communications systems (3G) to the fourth generation systems (4G). This (4G) system mostly incorporates broadband IP-based multimedia services in diverse networks. 4G systems represent a modified version of third generation mobiles and certainly have an edge over them. Their basic objectives entail higher transmission rates, larger storage capacity, higher frequency and greater bandwidth, better coverage and conduit features, cheaper access costs, higher quality of service with lower system costs, and a single yet omnipresent, multi-functional and multi-band intelligent device that can handle a variety of contents. [18]

C. An insight into futuristic devices what technologies may prevail

As mobile devices accelerate in terms of services offered, the more sophisticated applications can be predicted. For instance, like “Google Earth”, today the Global Satellite Positioning (GPS) system has enabled wireless devices to run applications for designing software that can provide a full socio-historical account of an image or video of any building besides indicating its geographical location. Software can even detect whether an apartment house on a boulevard is for rent or sale, or whether it is mortgaged and even who owns it. If a picture has some text the enabling technologies of these devices can translate it into English. These devices have supermarket applications and can scan products’ barcodes; iPhone has the same applications. [9]

In Europe drink and potato chips vending machine services initially started using mobile applications for immediate purchases. In recent years, Telstra, one of Australia's leading telecommunications provider companies, has taken the initiative by carrying out an exclusive "dial a coke" on trial basis where everyone can purchase a coke by making a phone call from a “Telstra mobile” to a number on a coke vending machine. By utilizing location sensitive mobile applications a person may find the nearest petrol and/or gas station or comparing prices when they go shopping, etc.[17]

The futuristic wireless devices can be envisioned to perform endless business functions such as:

- Electronic wallets - mobile phones may be used to hold credit cards and other monetary information for making electronic transactions. For this purpose mobile handsets are basically used as a security coupon registered with the user’s identity. When the transaction is done its verification directly corresponds to the user either by a phone call or SMS. In some parts of the world they can practically be used to buy small-value things such as tickets, public transport, parking fees and/or retailing machinery, etc. [9]
- Speech recognition and converter - while the data available on mobiles continues to accelerate and mostly mobile devices have become “data centric” instead of “voice centric” , voice leads the field and will remain a dominant form of communications. New built-in technologies in the wireless devices can now translate verbal communications into transcript form to shun the use of alpha-numeric key pads, etc. [18]
- Speech fusion - this innovative technology can translate e-mails into understandable verbal communications in order to listen to the received e-mails. This system can be dubbed the “automated interpreter”. [18]
- Optical/Visual character recognition - wireless handheld devices can also assist and enhance learning. As its name implies, visual features of these phones may convert any hand-written content to a typed-written format with a high degree of precision and with the help of its built-in learning apparatus. [18]
- Voice activation - the enabling technologies of wireless devices can bring voice-control to steer websites and to replace a long chain of chronological input with an automatic “voice-menu-driven” phone system. [18]
- The expanding functionalities of mobile phones have opened the door to refined applications. In this context Android is exceptional in that Google is aggressively developing its Linux-based mobile phone operating system. Google has formed an alliance of hardware, software, and telecommunication companies to consolidate Android development. As long as fragmentation issues can be avoided, Android phones are well ahead in the development stage since Android devices are tailored for specific hardware and user interface upgrades. [19]

D. Some examples of user-related success stories

Mobile phone-enabling technologies and similar applications have widely been accepted in Australia. Australia is a country where a substantial amount of revenue stems from subscribers using mobile phones. In recent years, the trend of sending text messages has increased considerably and this has implications for M-commerce users. An empirical study indicates that approximately 300 million text messages are sent by 11.5 million mobile phone users on a per month average over Australia's three leading mobile networks. Nevertheless, it is widely believed that the inception of a third generation (3G) mobile networking system is the driving force in expanding M-commerce applications.[17]

Application writers and operators are working on new business models which are able to generate a sufficient amount of revenue that can pay for new high-speed wireless networks. The billing systems, however, have been identified as a significant factor in retarding M-commerce services and applications to an extent. Rapid technological progress is the answer to these issues, and it is envisaged that M-commerce will be much more widely accepted in the future. [17]

Another important issue that needs the attention of mobile network operators is to decide and What the importance of emerging world of M-Commerce if it is restricted for business-to-business or business-to-consumer and/or consumer-to-consumer transactions or will expand their services to act as a reservoir for offering credit for airtime, or for goods and can provide loans or billing services to intermediary companies. To capture the small business market, operators can carry transactions between consumers and business for cash withdrawals, payments through mobile phone for soft drinks, car wash, train tickets as well as for larger dealings through debit card, VISA™ and MasterCard™, etc.[8]

1) Swissair

IBM has developed an application through which they can facilitate the Swissair's preferred passengers by sending updates for their flights. Passengers get all updates on mobile screen and subsequently printed on their boarding passes, i.e. departure time, gate and seat number etc In case of any changes in the flight schedule the passengers receive automatic updates on their phone display.[8]

2) Woolwich

Woolwich was the first British bank who introduces internet banking for customers using WAP. With the help of this they can manage their bank account personally. Customer services make sure to customer that their transaction safely transmitted from WPA phone to Woolwich server. [8]

III. E-Business Trends Through Mobile Devices

Mobile commerce or M-Commerce as its name implies can be defined as "the use of handheld wireless devices to communicate, interact, and/or conduct business transactions using high-speed connection to the Internet." [17] M-commerce is gradually becoming a leading force for doing business and in society generally. For more than two decades there has been a persistent "push" for moving forward technologies and a common "pull" of public demand for low-cost, high-speed and cost-effective communications and for an omnipresent access to information on a "follow me anytime/anywhere" basis that has transformed the telecommunications industry. Consequently, Internet access and high computing capacity of wireless devices has heralded the induction of new broadband interactive multimedia applications. Apart from the fact that the wireless web market is still in its infancy stage, M-commerce is likely to evolve dramatically in the coming years due to the emergence of 4G systems integrating many wireless networks, for example WBAN, WPAN, WLAN, and WMAN. [18]

The developments in M-commerce applications are relatively more complicated than those concerning e-commerce and therefore require specialized knowledge. Regarding the present state of technology all technological requirements such as high-speed access and low power devices plus business requirements cannot be achieved all at once, because there are interests in the value chain that are clearly in conflict with each other. Amongst those M-commerce applications which are considered highly personalized, context aware and location sensitive the most inspiring of them all include digital cash (for micro payments), human-to-machine communications (from still to moving objects for access, safety, asset and logistics using RFID).

A. Mobile banking services

The broad dispersal of personal mobile phones in general and dependability on mobile communication technologies in particular have made mobile solutions suitable for an array of financial services including mobile banking and other micropayment solutions. Mobile banking services have won the confidence of their users because there is an absence of time and place restrictions as well as the need to make a physical effort. [2] In Australia, lawmaking and enforcement regimes are working at both federal and state levels to control and regulate mobile commerce. Different regulatory bodies have also been set up especially for the banking, credit and telecommunications industries. [17]

Banking services make it possible for users to retrieve information on their account balances using SMS. However, the new wireless devices using GPRS applications can now support many banking services, for instance transfer of funds between accounts, stock trading, and can verify direct payments through a phone's micro browser. Characteristically, mobile banking services are the modified edition of Internet banking services offered by each respective bank which are designed and financed by a banking industry syndicate, for instance Mobey forum and ECBS, etc. Now WAP replace GPRS, people can pay their utility bills, to do shopping and connect to solo market where pay can be made using WAP services. For secure business transaction user can change their passwords and WTLS. [2]

Mobile networks are being upgraded with WAP, GPRS and UMTS applications and other enabling technologies to deliver next-generation multimedia services. Consequently, customers are now able to check their account statements, transfer of funds, and they are also notified of larger payments. Generally, they have immediate and full control over their online finances. The next generation of mobile banking services will improve their user-friendly image including motivated instructions, direct access, safety issues and immediate transaction processing with minimum costs. The banks will receive more customer confidence and increased dependability by providing them with a secure form of instant banking. Customers will have less low administrative costs, facing no branch restrictions, and modernized call centers with lower handling charges. [8]

B. Other financial services ranging from macro to micro payments

One M-commerce application that is likely to emerge concerns mobile payments and how they can be further classified into macro and micro payments. [2] Macro-payments is of $10 whereas; micro-payments is of $10 or less. One major distinction is that for macro-payments, confirmation is required through a trusted financial organization that has to be performed over unrestricted wireless and/or wired-line backed networks besides invoking all defensive and safety measures. In contrast, micro-payments are utilized through an operator's communications systems or entail a cash card in addition to (user's ID card that stores the classified information such as a user's covert confirmation key) for making instant payments over short distances using Bluetooth, Infra-Red, RFID, and UWB technologies, etc. [18] Mobile macro payments are done for bigger purchases either electronically (including e-commerce, mobile ticketing, gaming, etc) or on manned and unmanned POS (i.e. restaurant bills, retail shopping, etc.). On a related note macro payments are facing staunch opposition from conventional payment instruments. Nonetheless, solutions have been developed for user confirmation while making macro payments which provide opportunities for many different services including but not limited to passage control, digital signatures, etc. [2]

The success stories in Europe and Japan as to the sale of wireless services and related products indicate that consumer becoming acclimatized to making small values purchases of digital content. In this context Apple's decision to offer 99¢ MP3 downloads back in 2003 would have been signaled the beginning of a new epoch for micro payments. It can be envisaged that in the future micro payment providers will be confronted by mobile payment systems that have some intrinsic advantages that can persuade them to keep their transactional costs low. Whether any micro payment provider can achieve this critical mass or whether micro payments system is ready for takeoff is debatable. [2]

C. Security mobile phones in the usage in business

It has been resolved that the mobile handheld devices are productivity enhancing tools and bring many benefits for the enterprise but at the same time they are also devoid of many risks for organization's security as a considerable amount of confidential corporate and personal data can amass on a wireless handheld device having enough potential to seek the interest of an attacker. The more the capabilities and functionalities of these devices increase, the more the associated risks increase. [9] Another security risk that is very common with wireless devices is that it provides a favorable setting for unauthorized users since it is rather difficult to track the users having no fixed geographic position hence, they can go online and offline with a practiced ease. Because of their small sizes one more risk that is very typical with mobile devices is the risk of loss or theft. Though the accumulated data on a lost device is proprietary in nature and can not be recovered however, there is a persistent risk that any malicious finder of the lost device can hack into the proprietary corporate systems such as, email servers and file systems and the like. [1]

Today mobile phones have become indispensible for doing business and to watch consumer activity. They have become the most cherished communication devices in the modern world. An enterprise can directly contact peoples lives when it is equipped with WAP enabled phone applications. Up to now, SMS services have provided an easiest and simplest way to communicate one on one basis over a mobile network. However, the WAP experts predict that mobile handheld devices will soon to become the universal personal interface to information as well as services. It is envisaged that WAP technologies will not only enhance the users interest in adopting the existing internet applications i.e., electronic banking but will also let those innovative mobile technologies to control their additional dimensions. Nevertheless, it is important to be familiar with the initial implementations of mobile Internet access which will not be the same with which most of users are familiar. Mobile screen do not have colours it has just tiny graphics capability, but this is not an issue because mobile screen has coloured graphics built in video cameras. GPRS technology will be crucial for enabling the users to stay online via high speed data transferring Technology, without plunging into the formalities of dialing up. To cote Jeff Bezos, the chief executive and founder of Amazon.com supporting m-commerce, "If you look five to ten years out, almost all of e-commerce will be on wireless devices." [8]

D. Probable and persistent threats, theft and loss of data due to asynchronization.

Being tiny in size and portable "anywhere anytime", mobile handheld devices are regularly lost or misplaced or stolen. Unless the proper measures are taken, gaining access to the missing and/or gone data will be too difficult to do, making it difficult to save and access classified data. Today, software has been developed and installed in almost all modern handheld devices besides the above-mentioned data collection websites, to recover the removed data from flash memory synchronization. Generally, manual resetting of mobile devices is used to clean up data and restore its original settings before selling it. From a rational viewpoint it appears as if the cleaned-up data has completely vanished but it has actually been preserved somewhere in the device and marked as unused space. Alternatively, a way to evaluate the risks associated with handheld devices is to compare them to desktop computers. The risk profile of handheld devices is incomparable to that of desktop computers. Nevertheless, the supplementary threats follow from two main sources: firstly, size and portability; and secondly, accessible wireless interfaces and associated services. [9]

Organizations will have to put a ceiling on the access to information resident on a mobile device. This will mean frustrating unauthorized access to data by erasing or encrypting the same data on the device. Or it may be necessary to issue a command from a long distance. On the other hand encryption and data wipe solutions are the best safeguards against data being lost or stolen from the mobile. [14] On a related note phone flasher units have been designed to rewrite and restore the memory of different types of cell phones and can easily be purchased online, etc. [9]

One more step in preserving precious data placed on a handheld device is to back up the contents on a regular basis. For instance, data can be synched and/or linked to a desktop computer as a principal means for having backup data. Backing up is only effective if the memory card is kept away from the device. Both device and card can be lost or stolen simultaneously, severely compromising the benefits of such data protection. [9]

E. Unauthorized access to e-mail content by vicious hackers

Access to the device and its contents have not escaped the clutches of malicious hackers through forgery and speculating the verification of a user's identification such as PIN or password and/or by bypassing the whole verification system. It appears that a good number of cell/mobile phones and PDA users seldom utilize appropriate security mechanisms built-in to their devices, and if they do so they usually inadvertently apply those settings that can be easily bypassed and/or invaded. Having some inherent vulnerability, cell phones are prey to malicious hackers if they are not properly secured. [9] In wireless networks before an attacker attempts to track a target, targets can very simply come into an attacker's proximity. Wireless devices pass through many different and practically unreliable networks from which service is derived and data can be swapped over. Consequently the information can be stolen or corrupted without the user knowing how or when it was done. Quite often a service may be interrupted and subsequently disengaged. Similarly, communication can become sporadic and then restored on a regular basis without having regard to re-authenticating principles. However, a simple attempt at "revitalizing" the browser to reinstate connection may accidentally invite some risks. [1]

Malicious hackers can find the middle ground wireless connections. An example refers to airline passengers who randomly check their stock portfolios at departure lounges and do business from their mobile phones and a malicious hacker creeps into their favorite financial online site by using a DNS system that drags information to the malicious hacker's site. [1] On a related note "Blue jacking" as suggested by the name is a method of attacking Bluetooth-enabled mobile devices. "Blue jacking" begins when an attacker hijacks users' Bluetooth-enabled devices by sending spontaneous messages which are subsequently used to persuade the user to act in response in some manner and the new contact is added to the device's address book. These messages cause harm when a user responds to "blue jacking" that is sent with a harmful intent. [13]

"Blue bugging" on the other hand, utilizes a security error in the firmware of some older Bluetooth devices to obtain access to the device and its commands. This attack uses the instruction without informing the user, and permitted the attacker to access data. [13] In 2004, Nicholas Tombros pleaded guilty to obtaining unauthorized access to wireless computer networks in order to send spam emails advertising pornographic websites using his laptop connected to insecure wireless access points. This was the first case that was prosecuted under the US CAN-SPAM Act 2003. In light of the above, mobile and wireless security needs to be addressed by users and technical experts as well as law enforcement agencies in order to control or suppress criminal misuse. law enforcement agencies need to be well aware of the ways in which criminals have begun to take advantage of the vulnerabilities of these new forms of information and communication technologies, for example in the case of "Wikkileaks" and 'Julian Assange's "anti-spam and defamation" charges. [15]

F. The factors that drive mobile hacking

Mobile phone hacking is done for reasons of economic gains and viruses allow a burglar to access passwords and/or corporate data amassed on cell phones. Invaders can maneuver from a victim's phone for making calls or send messages and this offense is commonly dubbed "theft of service". As the users of mobile devices are now making macro and micro payments and conducting other financial transactions over their cell phones, these devices are becoming an easy prey to attackers.

Business and finance experts have predicted that such an activity will boom in the next few years. Presently, mobile phone users store their credit cards and other financial information for making electronic transactions by using electronic wallet software. Mobile devices are becoming likely targets due to their extensive use, given that there are millions of prospective targets. They possess several vulnerabilities such as not being well equipped with antivirus software. Another pitfall is that mobile devices compared to desktop computers are more exposed to the outer world and hence face the perils of hacking. Since mobile devices are primarily built to make communication as easy as possible on an "anywhere anytime" basis therefore, "phone users want to communicate, and viruses want to be communicated." [4]

The entire mobile banking system needs to be evaluated because threats like spoofing, tampering, denial of service, information disclosure, disclaimer and elevation of advantages and the like are occurring on mobile phone banking systems. In order to protect the sensitive data that resides on the phone device it has become vitally important that such sensitive data is encrypted during data communication and when the same is stored on the phone and/or kept in external memory cards. In the United States of America the instances of stealing credit card records by hacking through a wireless connection started appearing in 2003.

The three suspects of this conspiracy allegedly used a laptop that hacked into the "Michigan Lowes Store's" wireless network in the early morning from a car parked outside the building, gaining access to the company's central data centre in the North Carolina and seven other Lowes Stores across the country. They intended to install a data capturing program used to process credit card transactions, thus enabling them to steal credit card details. In 2004, one of the three suspects was sentenced to nine years imprisonment whilst, his collaborator was awarded 26 months of imprisonment in addition to two years of court administered release. [15]

G. Onslaughts on SMS, MMS and emails

A modified version of pagers' SMS services does the same job as mobile phones GSM- and CDMA-based technologies to send concise text messages to mobile phones having little data storage capacity. They are not considered a useful means for spreading mobile viruses but they do allow an influx of damage-causing viruses when a massive amount of SMS traffic flows between different wireless devices. MMS is a highly developed form of SMS for those cell phones that are properly equipped with GPRS-based technologies and can carry up to 50 Kbits of data which is enough for many viruses. Today, most cell phones can run e-mail applications. However, it would be somewhat difficult for a virus author not to write mobile malware application using e-mail attachments to pass on to wireless devices as happens in case of desktop computers. The damage in that way would not be as widespread in that unlike SMS and MMS, most people do not use their cell phones to read e-mails. [4] By and large, users deem it appropriate to install antivirus software in their computers yet these precautionary measures are still not prevalent in the cellular or mobile phone setting. Since most cell phone users are not aware of the prospective mobile malicious code they are not prepared to protect their phones from any attacks. Some mobile companies have started to install antivirus software in their phone sets such as Japan's NTT DoCoMo via the new Symbian-based FOMA 901i phones with McAfee's VirusScan technology. Nokia has launched two phones having a Symantec Client Security software preloaded onto the memory card and can be subsequently upgraded via its Symantec LiveUpdate system. [4]

Mobile wireless devices have a larger attack facade including but not limited to Bluetooth connectivity, Wi-Fi, and supplementary cellular communications interfaces and furthermore protocols for web transactions, electronic mail (e-mails), instantaneous messaging (chat messaging) and SMS, EMS, and MMS messaging. Conversely, the cellular channel encryption that ends at the radio interface is not sufficient to systematize the back-to-back privacy requirements of an organization which requires application-level encryption to be used over the network in future. [9]

The risks involved in using mobile wireless devices may include:

- Wireless devices are vulnerable to attack from a virus emerging from SMS trafficking, Bluetooth connectivity and\or PCs. For instance, a security vendor company "SimWorks International" recently identified that the first Symbian virus is dispersed through MMS messages.

- Frequent file conversion using Bluetooth connectivity between mobile devices and PCs has made malicious code occurrence more manageable, but made data theft or device damage more likely.

- Spam messages - whether they are done for the purpose of marketing and/or fraud - are considered to be the biggest carriers of a virus.

- There is an inherent risk in eavesdropped or accessed by unscrupulous users. Some users store personal data in their devices. Sometimes people reveal sensitive information on mobile communication. [16]

Some recent empirical studies have indicated that 'Trojans' and not worms or viruses are the main enemy. 'Trojans' technically speaking do not need any transmission vector and purely depend on the user's inquisitiveness to download them onto their wireless devices. 'Trojans' camouflage themselves as utility programs and/or popular games and consequently users install such programs without knowing what they are; a spyware capable of recording their incoming and outgoing SMS messages and also snooping on their dialed numbers and received calls. Another 'Trojan' using malware (like PbStealer) can filch sensitive and classified data like the user's PIN from a user's cell phone. Such an attack has to be taken seriously bearing in mind the fact that there are some J2ME based schemes that can store sensitive data, i.e. a user's private key that is persistently stored or retained, etc. 'Trojans' are a big risk regarding the security of the m-payments system in that these transactions need an authentication through SMS messages. The huge potential for SMS fraud for the purpose of financial gain is obvious. [6]

H. Electronic eavesdropping, voice mail recordings and voice messaging hacking.

Most cells phone users feel comfortable if their phones cannot be eavesdropped some unscrupulous listeners. Likewise, any such endeavors to access and overhear from the air is yet another probable risk that needs to be avoided. The problem of 'electronic eavesdropping' occurs by installing spy software onto a device for collecting discreet information via another phone and/or server. This sort of application exists in some specific phone models and is frequently advertised as a means to check on the activities of a spouse or children. The most important feature is their capacity to distantly switch on the microphone and listen to and/or record conversations. Cell phones having certain vulnerabilities can consent to the spy software being subjected to such active communications interfaces e. organizing a Notebook computer at a legitimate access point in a busy public spot can permit data to be stolen from unsuspecting customers. [9]

Since the communications between a handset device and cell tower are cautiously designed with safety and privacy issues they will be exploited by wily attackers. Scientists and researchers in Israel and the USA have explored effective ways to break the encoding/programming system for GSM-enabled cell phone networks to facilitate eavesdropping. With reference to the networking structure, a more focused yet targeted approach is that while having conversations with their subscribers, cell phones can be secretly personalized to allow eavesdropping by networking companies. **[9]** Some social aspects of eavesdropping also persist and this has led to the advent of 'Flexispy' spyware to assist people probe possibly cheating partners. This application can be used and applied by distantly activating the device's microphone to overhear something. As the capabilities of 'Flexispy' are mostly being employed to monitor the activities of spouses and unruly children yet the same also has the potential to be used in the corporate world to keep a check on the activities of the workforce. Sometimes 'Flexispy' spyware logs information from the device to a central server without the owner's prior knowledge. Once the software is installed, a hacker can read private messages, examine logs from any computer connected through the Internet and may also overhear confidential conversations and thus compromise intellectual property rights, etc**.** [14]

Specific utility programs exist to record the voice data of the calls made by the user. With little modification a hacker very cunningly convert it into a voice recording spyware. The same can be combined with another Trojan such as PbStealer to send the recorded data via Bluetooth. This malware can obstruct the working of some security schemes that use voice recognition apparatus for verification, as the hacker may replay the message recorded with the help of this spyware. The above-mentioned attacks have to be taken seriously as the recent case involving mobile phone hacking by the *News of the World* newspaper testifies. [6]

I. Unwanted spam and instance of malware

Of all the transmission channels, communications networks are the simplest way to transport viruses and other forms of malware to handheld mobile devices. There are many instances of infiltrating malware into wireless devices. For example they can be received while being synchronized with desktop computers and through infected storage media. Malware can also be spread in a number of ways, including but limited to:

- **Internet Downloads** – A user may be at a risk of downloading any corrupted file. As a camouflage tactic the file can either be a game, security patch, utility program or any other useful application posted from somewhere as a free download. Downloading of legitimate content can also create problems if they possess intrinsic vulnerabilities that malware can exploit.

- **Messaging Services** – Generally, malware tainted attachments may be affixed to electronic mail (e-mails) and MMS messages that are transported to a cell phone device. Instant Messaging (IM) services that engage many phones are yet another means to transfer malware. The users have to make a choice to open the attachment and subsequently install the same to invite malware to corrupt the device.

- **Bluetooth Communications** – Bluetooth connectivity is the most convenient method to hook up devices for sending messages or reshuffling files between them. However, the communications via a Bluetooth device may be positioned in various forms: it becomes discoverable whenever it lets a device be noticed by another Bluetooth-enabled device; and it becomes connectable whenever it permits the device to respond to messages retrieved from connected devices until finally switched off. [9]

Mobile phones are exposed to unwanted SMS text messages, e-mail and/or voice messages from advertisers. No matter what sort of inconvenience their removal may cause, however, charges may apply for any interconnected events.

For example, a per-message tax is levied on each SMS message received and/or a further charge is levied for those messages above the outer monthly limit of a service package. On the other hand, downloading of data may cost extra charges if the attachment has visual images, which means that the charges will remain high. Mobile spam has a tendency to be used fraudulently and the aim is to convince users to make a call or send text messages to taxable service numbers by adopting a *modus operandi* based on the concept of social engineering. Conversely, spam may also be used to convince users to disclose their private and confidential data such as passwords, financial details or other sensitive data via web pages, e-mail, or text messages, or to download malware attached to the message or through a web page. Thus spam and fraud complement each other. As stated above that Instant Messaging (IM) and multimedia messages (MMS) are the most convenient method for spreading malware through spamming. Denial of service is also considered to be yet another leeway using spam techniques. [9]

By utilizing the foregoing delivery methods the user generally has to give approval for the malware to be installed and properly executed. An array of malware behaviors and their following consequences are quite extensive. Thus, malware can possibly overhear user input or otherwise filch sensitive information, tear down stored information and subsequently halt a device from working properly. Some malware can also pull together wireless communications fees against a subscriber, i.e. by sending SMS messages or making calls to chargeable tax numbers. The proliferation onto other handheld devices or even with PCs can also be done by malware and virtually compromise the entire communications network. Below are some distinct yet importantly identified malware categories:

- **Spoofing** – Malware provides spurious information to the user to activate an action in the name of security.

- **Data Interruption** – Malware that resides on the device's applications is susceptible to interruption or access data residing on the phone's memory, respectively.

- **Data stealing** – Occupant malware on the device is able to collect and send data out of the device.

- **Backdoor** – Malware resident on the device is able to put forward deliberations to improve functionality that lets an attacker gain access.

- **Abuse of Service** – Occupant malware can execute those functions which can cause higher than expected service provider costs for the user and thus cause embarrassing financial losses.

- **Accessibility** – Malware resident on the device impacts on the availability or reliability of either the device or the data in it.

- **Network Access** – Malware resident on the device uses the device for more than one unlawful and unauthorized network activity such as port scanning or using the device as a substitute for network communications.

- **Wormable** – Occupant malware uses available technologies to publicize itself in a semi-autonomous manner. [9]

One virus (Commwarrior-B) has appeared on Symbian Series 60 phones and it spreads via MMS message attachments and/or Bluetooth. MMS recipients were asked whether or not they wanted to open the attachment, while Bluetooth recipients were asked if they wanted to accept the file and subsequently run it. The moment the virus is installed it starts finding other Bluetooth-enabled devices to infect. These viruses illustrate that the ways of imitation are many. A classic Trojan (Brador) sends the invaders an email message that contains the IP address of the device as a warning that the backdoor on the tainted device is now activated. The invaders in this way can connect to the device, view and download files or even upload new malicious codes. [9]

J. *Electronic tracking*

Some cellular carrier companies have acquired the expertise to track a device's location through the development of 'location tracking services' for registered cell phone users. These allow users and their friends and family to be connected with each other 24 hours a day. As for the organizations these services are also made public to keep an eye on their employees' whereabouts and hence improve productivity. Before the tracking service is activated some carrier companies issue a warning for the user that they are about to commence the monitoring, while giving the user an option that he/she may conclude the service if deemed appropriate. Nevertheless other service providers issue no such warnings of monitoring to their customers, if the process of registration is completed. On the technical side radio isolation bags containing metallic fibers create a 'Faraday cage' that frustrates radio frequencies and prevent tracking. However, they make normal use of cell or mobile phones impossible and the battery to deplete quickly. [9]

Of the many security-related technologies being used, Radio Frequency Identification (RFID) can spot objects and users; locations. It is very likely to become an important and core technology in today's global mobile communication systems. This technology has particularly proved useful for organizations that have many functions such as retail, supplies, accounts, design, etc. [11] Tracking services have their own vulnerabilities such as the possibility of clandestinely registering someone else's phone for monitoring purposes. For instance, if the system contemplates completing the registration process in that case a phone will require a sign of authentication. In other words, an SMS must reply with an authenticator code, but uses a code value that is not distinctive. One more approach that can also be utilized includes and is not limited to an online SMS access to engineer the response required to complete registration, etc. [9]

IV. Is Increased Use Of Mobile Phones In Workplaces Raising Productivity Issues?

A statistical survey on the use of mobile phones in the workplace indicated that by the end of 2010 the number of people using such phones was 850 to 1000 million and rising. An estimated speedy growth of smartphones and their bulk consignments on a yearly basis now that they are being used more than ever by employees are the key factors which are forcing businesses to ponder their impact on organizational security. The increasing demand of mobile phones in the workplace is creating productivity concerns because the increased functionalities of these phones are being shadowed by more risk factors. Other empirical surveys have shown that the smartphone market will surpass the laptop market within a couple of years; the global shipments of smartphones will double at the rate of 30% compound annual growth within the next two years. From employees' perspectives the use of smartphones and PDAs will enable major business transactions to be done on a "follow me anytime/anywhere" basis without the need for desktop computers. From the business point of view besides voice telephony, employees can use these wireless handheld devices for the following functions including but not limited to:

- Send and retrieve e-mails where a transaction needs a prompt response,
- Send and receive instant messages (IM) for a quick chat,
- Use vertical applications for administration and business strategies such as Enterprise Resource Planning (ERP), Customer Resource Management (CRM) and Sales Force Automation (SFA),
- Scan barcodes for prices of goods using high definition smartphones like iPhone, etc.,
- Browse web pages,
- Download and share files on the Internet and via Bluetooth connectivity,
- Use Personal Information Management (PIM) for keeping records of phone book contact information to prepare agenda items and convene meetings,
- Store confidential personal and corporate data, etc. [13]

It is important to state here that many companies have started to value access to the above business applications on mobile handheld devices, namely: Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and/or Sales Force Automation (SFA). These applications invariably contain some classified and sensitive data that will not only be useful for customer dealings but can improve the worth of their future business operations. Consequently, such mobile applications have transformed the purpose and meaning of PDAs and mobile devices which are no longer an optional gadget but a much needed business tool. [13] All of the above mentioned functionalities mobile handheld devices meet the criteria of being the most effective method to raise the workforce's efficiency as well as improve an enterprise's security and privacy risks. [13]

Mobile devices also have certain drawbacks such as any violation of safety setting on the device can be expensive for the organization. The increased number of mobile phone users and global Internet connectivity has diminished the prospects of the conservative "fixed" boundaries for organizations since a network protected by a central firewall is no longer enough for today's hi-tech lifestyle. [13] Users frequently move beyond these boundaries are they have become more susceptible to data theft and similar threats. Moreover mobile devices are also more vulnerable to carrying viruses, spam and other malware, which can be released through the network when the user is connected behind the network firewall. Portability of these compact tiny devices is yet another drawback attributed to these devices as they can be lost or stolen very easily, compromising the data accessed or stored on them. Major security risks due to mobile devices can hamper productivity and create hazards for a business; this generates the following problem scenarios:

- loss of a company's classified and sensitive data and intellectual property (IP) due to theft or loss of mobile devices,
- loss of employee productivity due to malware and malicious codes,
- loss of intellectual property due to spyware,
- fraud and lost productivity due to hacking, etc. [13]

A. *Theft or loss of mobiles may endanger a company's confidentiality policies*

Statistics show that on a yearly basis hundreds and thousands of cell phones and PDAs are stolen and/or misplaced. [9] However, the worth and price of the hardware and software of a lost device becomes unimportant and irrelevant compared to the worth and price of the data residing on the device. Lost data always remains susceptible to tainted reputations, cut-throat business strategies and possible legal action, etc. People - whether they are general customers, patients, investors, entrepreneurs and business people - put their trust in those companies managing their private and sensitive information. In this context some national governments have developed specific legislation, amendments and/or regulations requiring those companies to protect and administer data from any leaks. [13]

The Australian government has enacted various laws and regulations to cope with different types of content, including personal information. At the federal level the most important statute is the "Privacy Act 1988" which protects personal information and how it is handled by private sector organizations and government agencies. The Act also has relevance to how consumers who use m-commerce services may have their personal information and private details collected and used, particularly by advertisers and service providers including telecommunication operators which already handle a huge amount of private information regarding subscribers to mobile phone services. An individual's privacy rights have also been described in the form of Ten National Privacy Principles (NPPs) which define the parameters of such organizations while collecting, storing, using, disclosing, protecting and transferring customers' personal information, etc. In the case of contravening any of the provisions of the said law(s) for releasing any private and corporate information severe penalties can be imposed. [17]

Organizations should take precautionary measures to protect their valuable data by restricting any unauthorized access to the data stored on a device in case the device is misplaced and/or stolen. In this case encryption and data restoration policies or data wipe-off solutions may provide the best defense. Only data-driven policies cannot completely eradicate all types of risks pertaining to mobile theft or loss of data; it can only persuade employees and managers to remain vigilant in minimizing the risk of data leakage and compliance violations that may hamper a company's reputation. [13] A study conducted by the Readers Digest organization in 2007 suggested that in many of the world's largest cities an estimated 32% of lost phones are never recovered. A cell phone following its reactivation could be used arbitrarily to make international calls that the original subscriber must pay for. If the lost device is able to be restored to its original settings manually or and is reused easily, the contents of the user's data may be expunged. [9]

B. *Threats posed by malware may hamper employee productivity*

It has become abundantly clear that today's viruses and worms are a routine hazard for desktop computers, but the increasing demand and functionalities of hi-definition smartphone mobile devices makes them an attractive target for malware writers. There is every likelihood that the infiltration of malware on mobile devices can increase business and safety concerns. One major attack relating to mobile handheld devices appeared in 2000 and from then on viruses and malware have threatened the most popular mobile operating systems such as Symbian OS, and Windows Mobile3, etc. The built-in email and text messaging faculty of smartphones has made them an easy prey to viruses, as improved functionalities simply increases the risks. Malware can easily be disseminated via built-in Wi-Fi and Bluetooth connectivity through peer-to-peer communication for mobile devices. The viruses can influence a mobile phone's built-in messaging facility and PIM data so that it is sent to other mobile phones. [13]

Some viruses infect other devices that sustain MMS text messaging service if they act in response to the retrieved messages, like the infamous Mabir virus. Offenses like fraud and economic loss are also associated with mobile malware. Mobile malware can disturb the whole Symbian OS system by sending premium-rated messages of which the user has no knowledge. Futuristic mobile spyware may use such unlawful methods that were once considered the domain of desktop computers. For instance, SMS spam disseminates through junk text messages and can reveal users' confidential and private data through SMS-based phishing attacks which at times is known as "smishing." [13]

Wireless handheld devices have a tendency to create a security gap in an organization's security firewall. It has been observed that employees who carry their own devices (employee owned equipment) will definitely try to synchronize them with their office terminal (organization-issued equipment) and/or use their own devices to get connected to the terminal's Internet. [9] Many employees would like to improve their productivity by synchronizing their mobile handheld devices to their laptops. Organizations have to protect their mobile devices in the same way as they protect desktops and laptops. Since mobile handheld devices operate from networks outside the controlled boundaries of an enterprise they can obtain access to the organization's network which causes the whole IT-based system to become infected. Such devices need to install anti-malware software to curb the risk of infection, which are not limited to Viruses, Trojans, Spyware, etc. The persistent danger remains that once malware gets installed on a device the same can not only steal the private and sensitive data residing on the device but may reduce productivity levels and escalate expenses for organizations. [13]

C. *Economic issues: customer receipts, taxation and running costs*

For organizations it is important to know what kind and model of devices are being used that can maintain their security policy and improve productivity without putting an unnecessary burden on their economic policies. [13] An omission in checking mobile handheld devices can cause financial losses so care must be taken when managing credit cards, which remain under the holders' control. If an organization's mobile phone is lent to a non-related person it is at risk of misuse and activation of malware and unwanted services such as mobile tracking, etc. Organizations may have to pay huge expenses regarding toll calls and if confidential data is misplaced. Tax laws in a number of countries have provided with a limited number of the services offered by mobile companies free of charge up to which a tax or surcharge is levied on each and every transaction hence, an organizational-issued mobile if becomes a prey of the clutches of an unscrupulous user may incur heavy losses to the organizations' reserves, etc. [9]

It is always best not to keep any confidential information on a mobile handheld device such as financial accounts. Verification mechanism can be thwarted by hackers and therefore any sort of authenticating data such as PINs, passwords, user IDs, and financial details should not be placed on a device's memory. If so such sensitive data should be retained only in an encrypted form. Most of today's smartphone devices like Symbian and iPhon support built-in encryption capabilities by providing a "wallet" that stores personal information when needed. Nevertheless, where the device is an organization-issued device then the aforesaid should comply with the company's policies. [9] Large organizations are also endorsing new productivity enhancing policies which have proven to be commercially viable and cost-effective. For instance, in some countries international food chain companies and sky shoppers have provided their employees with printer applications attached to their handheld mobile phone devices, particularly those associated with supply and delivery departments that issue billing receipts to customers to keep financial records. This is financially beneficial to companies so that they can check any billing irregularities. However, this practice is not suitable for smaller organizations such as Third World enterprises which cannot afford such expensive devices supported by a whole integrated workstation system. [9]

It is abundantly clear how mobile handheld devices are becoming indispensable for today's organizations to improve productivity. However, their global acceptance is only happening gradually because they are not integral to all aspects of organizational infrastructure. One core issue which is being faced by the organizations is how to differentiate between the "employee-owned equipment" versus "organization-issued equipment". At the outset, it seems practically workable to let "employee-owned" cell phones and PDAs to be used for business purposes in a cost-effective way. Nevertheless, it is difficult to develop the capabilities to control and handle these devices. More importantly, security concerns for cell phone handheld devices range from those that are commonly linked to computer equipment because they operate from platforms outside the restricted boundaries of "fixed" devices. Furthermore, a number of safeguards are invariably available for desktop and networked workstations but are not commonly available for a wide range of handheld devices. On the other hand "organization-issued devices" can be administered as the basic functionalities of these devices are known, their configurations can be sporadically managed in accordance with company policy. It is therefore suggested that the said functionalities can let organizational applications particularly those developed for PCs be more easily extended to the mobile platform. [9]

D. *Fraud and lost productivity are likely hacking targets*

Malware, Spam, Trojan as well as unwanted content are considered harmful for mobile security systems and susceptible to being attacked through hacking and/or denial of service, etc. Viruses take advantage of the limitations and vulnerabilities of mobile phone operating systems before initiating any attacks. For instance, one known malware called "Skulls" attacks all links on a mobile handheld device by neutralizing its applications. Consequently, if a device becomes infected with this kind of malware, the user cannot send any e-mails or instant messages; in fact all symbols on the phone device are replaced with the skull image of the "Skulls" virus. These threats can be alleviated by adopting and implementing industry best practices. IT administrators in large organizations can recommend integrated practices for protecting mobile devices from any sort of security-related risks. Advances in technology have led to improvements in security which is a three-fold system in that it involves people, events and blue-chip technology. All these elements need to be considered to create the best security system possible. Some of the prevailing policies have been discussed very briefly in the above sections but in a different context. [13]

The focal point of these integrated best practice policies should be how to protect handheld device from unscrupulous users and for this purpose Password protection is considered to be the most effective barrier to protect data intrusion. All mobile devices must have a power-on-password enabled facility so that phone users can be identified with their respective device. Nevertheless, a vigorous mobile security plan would empower administrators to execute reliable and integrated policies for all devices from a single location. For instance administrators are required to be proficient in preventing all brute force log-on attempts, i.e. multiple attempts with different login/password combinations and the like. [13] On the other hand, encryption is still considered to be the first line of defense against any invasion of a phone to prevent loss of data or theft. It is also important to protect data in transit (travelling data e-mails, etc.) during “device to server” transmission. There are some security protocols which help ensure that data is properly and safely transmitted. Of these, the “SSL” protocol protects data in transit because it is very economical and simple to implement and does not need any new client software on a mobile device. In contrast, VPNs also secure data in transit but they are expensive, have a propensity to drain battery life and require a client software. Administrators should be properly skilled to configure all forms of data encryption and how to use algorithms. [13]

Anti-malware and anti-spam solutions should be updated on a regular basis so that new prototypes for notorious malware can be discovered and dealt with. Moreover, mobile phone data has to be scanned immediately including data residing on mobile devices and on external memory cards whenever they are inserted. If the administrator deems it necessary a manually scan the devices then this should be done. [9] By and large, malicious programs can easily be disseminated to cell phones via communication channels, i.e. MMS or Bluetooth connections. Whenever, a message or contact is received on a mobile phone from an unknown number it deserves to be treated cautiously. Regularly received MMS messages or e-mails even from a familiar number and/or address, containing an attachment to be installed can become susceptible to a malicious program. [9] Whenever possible, Bluetooth settings must be constructed with the utmost security by sending phone users prior intimations regarding the incoming link requests and obtain their verifications before they take place. Most of today’s smart phones offer this service to manage Bluetooth functionalities by allowing only selective profiles which are required to support activation with another mobile handheld device. It has been suggested that device pairing should not be done at public places, but instead in places that are radio isolated and/or in Radio Frequency Identification (RFID)-free environments. This will deter the chance of being monitored or recorded over the air thereby using them to restore protection keys that may be used while eavesdropping. [9]

V. SAFEGUARDS AND PREVENTIONS

Mobile handheld devices are productivity enhancing tools even though they do have serious security problems. Yet organizations are still reluctant to realize their significance as a vital component of a particular organizational infrastructure. Without delving into the advantages and disadvantages of employee-owned tools as well as organization-issued tools, the “control factor” of these wireless devices is rather difficult to ascertain as they are not controlled by approved platforms vis-à-vis “fixed devices”. [9]

Since the security concerns pertaining to cell or mobile phone devices intrinsically vary from those of the desktop computers, many safeguards which are invariably available for desktop and other networked computers in the workplace are not as accessible for all kinds of mobile handheld devices. The reason behind this is that on the whole organization-issued devices are much easier to manage because the traits of these “fixed” devices are already identified; their prototypes can be easily managed and joysticks can be installed when needed to enhance the level of security, in conformity with the company’s policies. However, a workable suggestion would be that the said attributes of the organizational applications for desktop computers can be extended to mobile platforms as well. Our discussion will include an appraisal of the range of safeguards available for mobile handheld devices, and how they eradicate the associated risks for the organizations. [9]

A. User-oriented measures for maintaining security

The security of handheld devices cannot be maintained without users’ involvement. Users must be instructed about what measures to follow and what precautions to take while they use organization-issued equipment. For instance, numerous built-in configuration settings and security prototypes of handheld devices are seldom used. Taking full advantage of all the facilities afforded by cell phones or PDAs, it is vitally important that the user know all the security safeguards. [9] By and large, user authentication methods are available on a majority of devices such as PINs and passwords though they are considered to be the first barrier to any unscrupulous access, yet have certain pitfalls. For users it is rather difficult to understand and analyze the plethora of documents involving all the features and options available on a handheld device for authentication, as it entails accurate and secure choices. [9] Users should prevent keeping any sort of confidential data on a handheld device since the authentication mechanism is not devoid of certain weaknesses and can easily be bypassed or wrecked and/or recycled from the deleted data. Even if confidential data is kept on detachable memory cards, it should be kept away from the device unless required. When it becomes imperative to keep the sensitive data on devices, it should be kept in encrypted form since most of today’s smart phones do support built-in encryption capabilities to meet this requirement. [9]

Yet another simple protective measure against various forms of malware that users can employ is just to simply turn off Bluetooth, Wi-Fi, infrared, and other wireless interfaces, unless absolutely needed. This is because Bluetooth devices are prone to escalating risk factors due to mobile malware, particularly in crowded surroundings such as airports, sports events and/or music concerts that proffer a target-enriched environment. Immobilizing a wireless interface also has the advantage of extending the battery life of the device. In addition to the above, automatic connections to cellular data services, that is, GPRS or EDGE systems, are also better to be turned off while not being used. Staying offline brings many fringe benefits as well since it averts the risks posed by malware infection and it can also thwart an infected device from sending contaminated data to other parties. If a phone device automatically connects to data services it can also be a direct warning that the phone has been infected by malware and is now attempting to spread itself through various applications. [9]

In case the device is misplaced or stolen the user can take precautionary measures even from a distance, such as disabling service, locking the device and/or completely wiping out its contents by immediately reporting the incident to the cellular carrier company. In this regard GSM carriers in many countries have taken a quantum leap as they can now register the identifier of the phone, for instance International Mobile

Equipment Identity or (IMEI) in a global database that prevents it being used elsewhere. It is, however, important to understand regarding the whole reporting system prior to the incident what kind of information is essential. Stolen devices may accrue substantial charges that the subscriber of the phone must pay until the device is reported as stolen. A copy of the filed police report may be required for phone charges to be dropped. [9]

B. *User authentication and physical control on the mobile device*

Of all the available safeguards concerning the security of mobile handheld devices, users' authentication techniques are commonly available in many devices, for example PINs and passwords. Although these knowledge-based authentication techniques are not infallible and have certain vulnerabilities, they are considered to be the first line of defense against unauthorized users' access. There are three main categories of users' authentication techniques commonly being used for authentication:

- proof by knowledge (passwords),
- proof by possession (tokens, i.e. smart cards),
- proof by property (fingerprints).

The aforementioned techniques can be used either alone or in tandem with others. However, using more than one type of authentication technique is also feasible and affords better protection. Passwords on one hand are believed to be the oldest and most popular form of proof-by-knowledge technique for handheld devices. Likewise, smart card authentication is best known for its proof-by-possession technique. Having entrenched the computer chip operating system, programs and data storage systems, these credit card-sized security tokens have become an integral part of the internal security infrastructure of some organizations, which are extending the already installed smart cards to handheld devices. Smart cards are able to transmit users' security identification and policy rules to a device that administers users' authorization and permissible behavior. Passwords fingerprints are also considered to be the oldest proof-by-property technique that involves the biometric system. This technology is relatively complex compared to the rest of the above two and therefore, a small number of mobile handheld devices have built-in fingerprint-based technology authentication. [9]

The proof by knowledge technique (passwords) can be divided into two categories, where the former enables users to choose a series of displayed images and the latter refers to sketch a series of lines over a network or follows the icon pattern. The former, however, is being applied in various commercial security products for handheld devices. The most notable development in proof by possession (smart cards) technique is that now wireless smart cards insert a radio frequency chip or in a more compatible mode; some manufacturers have introduced removable media as well. **[9]** Organizations must adopt concrete and meticulous policies regarding passwords and PINs for cell phones and PDAs in a back-to-back and composite manner. Nevertheless, there should be restraints on using the same password for a handheld device which is to be used for network access or access to other devices and applications. In case the password is erased from memory, different techniques can be utilized to recover the same from various handheld devices. This infers collaboration access to the network or other devices in turn. Various authentication techniques incorporate a time-out feature that can automatically lock the device the moment it reaches the verge of a stipulated condition, such as a screen saver, etc. At times these techniques can be rather irritating, but they are meant to help protect a misplaced or stolen device or until the owner recovers it. [9]

Keeping physical control of a mobile handheld device is also vitally important. Like all precious possessions these devices should not be left unattended. The contents and confidential data that resides on a device's memory also be jeopardized if an unauthorized and dishonest user gains access to it. It is dangerous to let else to use the device as it inadvertently invites malware and/or activation of unwanted services, such as mishandling while retrieving messages and/or taking unwanted calls. Sometimes, even a slight change in the security settings of the device can expose it to other types of threats that remain unnoticed because the user does not know what kind of changes have been made in the security settings. [9]

C. *Minimized functionality of devices and decrease data exposure*

The more the augmented functionalities and innovative technologies are taking place the more the proportion of risk is mounting. To cope with this problem the most viable solution could be to decrease the number of functionalities offered by them except those which are particularly needed. Consequently, one good paradigm for getting the desired results is to minimize the wireless interfaces unless urgently required and by rendering all those superfluous features inoperative through configuration settings. However, in certain circumstances some features may also be removed permanently to avoid their involuntary reactivation. Likewise, reducing the use of attached applications and/or plug-ins may also provide desired benefits. As these applications do have certain vulnerabilities, if they are installed they can get into the user's content and compromise the programming interfaces of the device. It is therefore vitally important that prior to the installation of any such applications their advantages and disadvantages have been evaluated. [9]

Cellular service agreements and subsequent service settings are yet another method to manage or simplify the functionality issue. If data service is eradicated for the activation of voice service only the same may avert full access to the Internet, which is be considered to be an appropriate solution. On the other hand, it can be possible to have the carrier restrict access to international destinations which are not being used or blocking other services. An example concerns various cellular carriers who offer to block subscribers' text messages initiated directly from the Internet because this is a major cause of disseminating wireless spam. It has been suggested that it is better not to keep sensitive and private data on mobile devices. This is because the whole authentication system can be easily avoided and deleted information may be restored from a phone's memory. Regardless of how convenient it is for subscribers to verify their online financial records and/or to other devices via PINs, passwords, user IDs, and account numbers on wireless handheld devices, it should be avoided. Classified and sensitive data can also be stored on detachable memory cards but the same should be kept separate from the device. Moreover, matters pertaining to labeling and tracking of sensitive data residing on these devices can also divert attention. [9]

D. Restoration of back-up data and installation of preventive and detection software

It would be a catastrophe if someone starts using handheld devices as the sole depository for keeping important information. The device may be misplaced or stolen or damaged accidentally. For protecting the precious data residing on a device it is safer to restore back-up of data on a regular basis. For this reason, data can either be synchronized with a desktop computer for keeping a back-up or for any possible dual purpose. Alternatively, this back-up data can also be kept on the memory card but the card can only be supportive if it is kept detached from the handheld device. The chances of restoration of back-up data will be further narrowed down if both the device and the card are lost or stolen simultaneously. [9]

The operating systems and built-in technologies of mobile handheld devices are far more complex than desktop computers, and hence warrant extra security controls for the prevention and detection of attacks against them. In this regard the installation of "Prevention and Detection" software for defending and protecting against any kind of malware onslaught has become indispensable. Consequently, a large range of such equipment is now available for a number of today's handheld devices, especially, for smart phones and PDAs that may be used to supplement the already existing built-in security mechanisms in them. It should be mentioned here that these "add-on" security software systems do have certain vulnerabilities and they should be evaluated very carefully. These types of equipment generally contain one or more of the below mentioned capabilities:

- User authentication alternatives, including biometrics (proof by property) and token-based (proof by possession) techniques,
- Content and memory card encryption,
- Firewall and Intrusion detection system,
- Antivirus and antispam (anti-malware),
- Content and memory card erasure (wipe off technology), and
- Virtual private networking. [9]

Organizations must consider how to protect mobile handheld devices. Although these wireless handheld devices are controlled on the networks which are outside the ambit of organizations' "fixed" boundaries, when they connect back to the network they may pollute organizations' whole IT systems. Extending equal protection to mobile handheld devices with those of the desktops and laptops, these devices must have anti-malware software installed on them to reduce or eradicate infection. Once malware gets into a mobile device it may not only drain off all private and classified data, but may also severely compromise productivity levels and augmented support expenses. Anti-malware software works by finding solutions to scan all the upcoming mobile threats by not having them installed on the device. Commercially available software such as Flexispy has the same application. To be more precise, the most effective anti-malware and anti-spam solutions should obtain updates periodically, about all new prototypes for already identified malware with the lowest amount of users' and/or administrative intervention. It is suggested that data residing on a device's memory has to be scanned instantaneously, including the data residing on external memory cards whenever they are inserted. [13]

Mobile handheld devices also require a complete firewall protection to curb unauthorized access. Firewalls are best known for their cautious scrutinizing capabilities and once set in motion, they can restrict mobile traffic. Moreover, private firewalls are also indispensable for obstructing port scans that the attackers usually use to discover vulnerabilities when a device is linked to a public network. Firewalls also supposed to be the first barrier against any abuse of un-punched security holes in a device operating system and/or client applications. Thus, businesses must install a comprehensive firewall and intrusion detection systems with pre-defined security standards which can further be modified by the administrators in a particular workplace environment. An intrusion detection system (IDS) while implemented on a mobile handheld device can negate service attacks by identifying the prototypes of network traffic. On a related note, a better solution is where administrators are authorized to establish an exemption list to supersede security level settings or by blocking certain types of network traffic. For instance, administrators should be empowered to prohibit certain types of protocols, ports, and IP addresses from inspection because in a particular organizational setting users may have different levels of requirements and usage. The following measures will enhance the level of mobile security:

- Inspection and access perimeters for devices,
- Firewalls to curtail the type and origin of network traffic,
- Easy-to-deploy firewall solutions with pre-defined security levels to be modified by the administrator, and
- Intrusion detection systems to obstruct denial of service attacks, etc. [13]

E. Solution Methods (I-clouds)

A major global IT problem refers to disk storage and particularly its operating costs, which have been estimated as representing nearly 30-50% of gross capital expenditures per annum in many enterprises. Against this background, enterprises must manage effectively the costs of storing data, especially unformed data. Therefore, to address this need, Cloud storage services have emerged and become fashionable. [21] Nevertheless, the current economic state of affairs as well as the advent of new technologies has ignited the interest of organizations in Cloud storage services and/or provider models. Cloud storage providers can suggest cost-cutting measures by using equal storage capacity to meet organizations' needs, such as transitional cost-savings measures for their customer base. [21]

Cloud storage services/providers are commonly known as "Cloud Backup" which makes back-ups of enterprise archives online automatic, while data is safely stored externally in data centers and is easily recovered. For example, if an unfortunate fire incident happens at any workplace or home both computers and disks (i.e. CD-ROM or tape, etc.) can be damaged. With Cloud Backup this threat is eradicated. A number of platforms are being supported by Cloud Backup such as:

- Windows___ Windows Vista, Windows XP, 2000 Professional, Server 2000/2003,
- Linux_____ and its most popular versions RedHat, SuSE, Debian, and its supportive system like Ubuntu, all with Java 1.5 or higher,
- Mac OS X, etc. [19]

Security has been the core issue in the developing phase of Cloud Backup. Data is first encrypted and condensed and then it is sent to and stored on a Backup Server in a fully secured data center. This involves a fully secured (SSL) network connection that makes the Cloud Backup service as safe as online banking. Nobody can enter the data stored at data centers except the subscriber. In other cases, the moment the data enters into an insecure setting like the Internet, it is vulnerable to hackers. An answer to this is Cloud Backup which encodes data (in encrypted form) before it leaves the terminal and subsequently decodes it while it is restored on a workstation. [19] This is why security and accessibility issues come first when companies choose to transfer their sensitive data to the Cloud Backup system, via the Internet. [21] The market for Cloud Backup services is rapidly growing owing to the huge amount of private and corporate data now being stored on desktop computers and laptops and more recently on smartphones. Smartphones have the equivalent storage capacity of their larger counterparts. [20]

VI. Conclusions

The omnipresent exploitation of mobile systems, extensive use of the Internet as well as the speedy growth in wireless technologies and the broadband interactive multimedia applications in recent years have helped improve the functions and characteristics of mobile handheld devices. [9] The said devices and communication systems have also developed especially built-in software to curb any instances of encroachment. [18] Though mobile devices can provide many productivity benefits for organizations and businesses, these devices are subject to security risks such as malicious codes and communication attacks, theft of data, spam, etc. [15] Security concerns become vitally important vis-à-vis financial transactions so that the assault vectors in this application need to be examined with due diligence. [6] Therefore, mobile security must be tackled by both the users and law enforcement agencies to reduce the risk of criminal misuse. [14] In this economic environment new technologies have made organizations very interested in "Cloud Backup" because it creates automatic back-ups of enterprise archives online, while data remains safely stored in data centers and can easily be recovered. [19] Enhanced wireless security and the broad exploitation of 4G systems in the coming years will enable mobile commerce to become the best way of doing business. [18]

References

[1] Ghosh, A.K. & Tara M. Swaminatha (2001). Examining the risks in wireless computing that will likely influence the emerging m-commerce market: Software Security and Privacy Risks in Mobile E-Commerce. *Communications of the ACM*, vol. 44, no. 2, pp. 51-57.

[2] Mallat, Niina, Matti Rossi & Virpi Kristiina Tuunainen (2004). Adopting new and innovative mobile financial applications and service provisioning methods. *Mobile Banking Service*, vol. 47, no. 5, pp. 42-46.

[3] Wu, Min, Simson Garfinkel & Rob Miller (2004).Secure Web Authentication with Mobile Phones. MIT Computer Science and Artificial Intelligence Laboratory. DIMACS Workshop on Usable Privacy and Security Software.

[4] Leavitt, Neal (2005). Mobile Phones: The Next Frontier for Hackers? *Computers*, vol. 38, no. 4, pp. 20-23.

[5] Milanovic, Nikola, Miroslaw Malek, Anthony Davidson & Veljko Milutinovic (2004). Routing and Security in Mobile Ad Hoc Networks. *Computer*, vol. 37, no. 2, pp. 61-65.

[6] Agarwal, Shivani, Mitesh Khapra, Bernard Menezes & Nirav Uchat (2008). Security Issues in Mobile Payment Systems. Department of Computer Science and Engineering, IIT Bombay, India.

[7] Lehr, William & Lee W. McKnight (2002). A research and education initiative at the MIT Sloan School of Management. Wireless Internet Access: 3G vs. WiFi? Center for Business@MIT, 21p.

[8] Interforum (2001). M-Commerce: E-Business without boundaries. *Interforum: Helping Britain to Trade Electronically*, no. 8.

[9] Jensen, Wayne & Karen Scarfone (2008). PDA Security: Guidelines on Cell Phone and PDA Security: Recommendations of the National Institute of Standards and Technology Special Publication 800-124, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology.

[10] Varshney, Upkar, Ronald J. Vetter & Ravi Kalakota (2000). *Computer*, vol. 33, no. 10, pp. 32-38.

[11] Lee, Hyanjiin & Jeeyeon Kim (2006). Privacy threats and issues in mobile RFID. *Proceedings of the First International Conference on Availability, Reliability and Security*, 20-22 April, 2006. Korea Information Security Agency, 5p.

[12] Scarfone, Karen & John Padgette (2008). *Guide to Bluetooth Security: Recommendations of the National Institute of Standards and Technology*, Special Publication 800-121. Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology.

[13] Enterprise Mobile Security: Protecting Mobile Data and Increasing Productivity A Trend Micro White Paper, November 2007, Trend Micro, Incorporated

[14] Urbas, Gregor & Tony Krone (2006). Mobile and wireless technologies: security and risk factors Trends & Issues in Crime and Criminal Justice, no. 329, Australian Institute of Criminology.

[15] Ying, Liu, Huang Dinglong, Zhu Haiyi, & Patrick Rau (2007). Users' Perception of Mobile Information Security. *Hacker Journals White Papers*. Computer Security Knowledge Base Portal.

[16] Palen, Leysia, Marilyn Salzman & Ed Youngs (2000). Going Wireless: Behavior & Practice of New Mobile Phone Users. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. ACM, pp. 201-210.

[17] Consumer Affairs Victoria (2002). M-Commerce - What is it, What Will it Mean for Consumers? [work in progress]. Department of Justice.

[18] Grami, Ali & Bernadette H. Schell (2004). Future Trends in Mobile Commerce: Service Offerings, Technological Advances and Security Challenges. In *Proceedings Second Annual Conference on Privacy, Security and Trust*, October 13-15, 2004, Wu Centre, University of New Brunswick, Fredericton, New Brunswick, Canada, 14p.

[19] Cloud Backup: Cloud Backup - FAQs, April 2010, Version 1.6, https://backup.eu.businessitondemand.com

[20] Fu, Yinjin, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, Hong Jiang, Nong Xiao, Lei Tian, & Fang Liu (2011). AA-Dedupe, AA:Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment: IEEE Cluster 2011 Technical Paper TP-2b. National University of Defense Technology, China.

[21] Ju, Jiehui, Jiyi Wu, Jianqing Fu, & Zhijie Lin (2011), A Survey on Cloud Storage. *Journal of Computers*, vol. 6, no. 8.

[22] Hui, Suk Yu & Kai Hau Yeung (2003). Topics in Wireless Communications: Challenges in the Migration to 4G Mobile Systems. *IEEE Communications Magazine*, December, pp. 54-59

Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool

Ahmad Ashari
Department of Computer Science
and Electronics
GadjahMada University
Yogyakarta, Indonesia

Iman Paryudi
Institute of Software Technology and
Interactive Systems
Vienna University of Technology
Vienna, Austria

A Min Tjoa
Institute of Software Technology and
Interactive Systems
Vienna University of Technology
Vienna, Austria

***Abstract*—Energy simulation tool is a tool to simulate energy use by a building prior to the erection of the building. Commonly it has a feature providing alternative designs that are better than the user's design. In this paper, we propose a novel method in searching alternative design that is by using classification method. The classifiers we use are Naïve Bayes, Decision Tree, and k-Nearest Neighbor.**

Our experiments hows that Decision Tree has the fastest classification time followed by Naïve Bayes and k-Nearest Neighbor. The differences between classification time of Decision Tree and Naïve Bayes also between Naïve Bayes and k-NN are about an order of magnitude. Based on Percision, Recall, F-measure, Accuracy, and AUC, the performance of Naïve Bayes is the best. It outperforms Decision Tree and k-Nearest Neighbor on all parameters but precision.

Keywords—energy simulation tool; classification method; naïve bayes; decision tree; k-nearest neighbor

I. Introduction

Energy simulation tool is a tool to simulate energy use by a building prior to the erection of the building. The output of such simulation is a value in kWh/m^2 called energy performance. The calculation of the building energy performance must be carried out by developers as part of requirements to get permit to build the building. The building can only be built if the energy performance is below the allowable standard.

In order to get building energy performance below the standard, architects must revise the design several times. And in order to ease the design work of the architects, an energy simulation tool must have a feature that suggests a better alternative design.

Since the alternative design search is actually a classification problem, hence in this paper we propose a novel method to search alternative design by using classification method. The classification methods used in here are Decision Tree, Naïve Bayes, and k-Nearest Neighbor. We will then compare the performance of these three methods in searching alternative design in an energy simulation tools.

The rest of the paper is structured as follows: Section 2 describes the classification methods we use in this study. Section 3 explains the data preparation followed by the experiment in Section 4. The result and its discussion are presented in section 5 and 6 respectively. Section 7 concludes the paper.

II. Classification Method

Classification is the separation or ordering of objects into classes [1]. There are two phases in classification algorithm: first, the algorithm tries to find a model for the class attribute as a function of other variables of the datasets. Next, it applies previously designed model on the new and unseen datasets for determining the related class of each record [2].

Classification has been applied in many fields such as medical, astronomy, commerce, biology, media, etc. There are many techniques in classification method like: Decision Tree, Naïve Bayes, k-Nearest Neighbor, Neural Networks, Support Vector Machine, and Genetic Algorithm. In this paper we will use Decision Tree, Naïve Bayes, and k-Nearest Neighbor.

A. Decision Tree

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions [3].

The popular Decision Tree algorithms are ID3, C4.5, CART. The ID3 algorithm is considered as a very simple decision tree algorithm. It uses information gain as splitting criteria. C4.5 is an evolution of ID3. It uses gain ratio as splitting criteria [4].

CART algorithm uses Gini coefficient as the test attribute selection criteria, and each time selects an attribute with the smallest Gini coefficient as the test attribute for a given set [5].

The advantage of using Decision Trees in classifying the data is that they are simple to understand and interpret [6]. However, decision trees have such disadvantages as [4]:

1) Most of the algorithms (like ID3 and C4.5) require that the target attribute will have only discrete values.

2) As decision trees use the "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present.

B. Naive Bayes

Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, hence is called "naive" [3]. This classifier is also called idiot Bayes, simple Bayes, or independent Bayes [7].

The advantages of Naive Bayes are [8]:

- It uses a very intuitive technique. Bayes classifiers, unlike neural networks, do not have several free parameters that must be set. This greatly simplifies the design process.
- Since the classifier returns probabilities, it is simpler to apply these results to a wide variety of tasks than if an arbitrary scale was used.
- It does not require large amounts of data before learning can begin.
- Naive Bayes classifiers are computationally fast when making decisions.

C. k-Nearest Neighbor

The *k*-nearest neighbor algorithm (k-NN) is a method to classify an object based on the majority class amongst its *k*-nearest neighbors. The k-NN is a type of lazy learning where the function is only approximated locally and all computation is deferred until classification [9].

k-NN algorithm usually use the Euclidean or the Manhattan distance. However, any other distance such as the Chebyshev norm or the Mahalanob is distance can also be used [10]. In this experiment, Euclidean distance is used. Suppose the query instance have coordinates (a, b) and the coordinate of training sample is (c, d) then square Euclidean distance is:

$$x^2 = (c - a)^2 + (d - b)^2 \tag{1}$$

III. DATA PREPARATION

In classification method, training set is needed to construct a model. This training set contains a set of attributes with one attribute being the attribute of the class. Then the constructed model is used to classify an instance.

For this experiment, there are more than 67 millions of raw data available. This data comes from combination of 13 building parameters with each parameter has 4 possible values (4^{13} data).

The parameters and the values used in each parameter are as follows:

1. Wall U-value: 0.1; 0.15; 0.2; 0.25 W/m^{2}K
2. Wall Height: 2.5; 3.0; 3.5; 4.0 m
3. Roof U-value: 0.1; 0.15; 0.2; 0.25 W/m^{2}K
4. Floor U-value: 0.1; 0.15; 0.2; 0.25 W/m^{2}K
5. Floor Area: 70; 105; 140; 175 m^2
6. Number of Floors: 1; 2; 3; 4
7. Window U-value: 0.1; 0.7; 1.3; 1.9 W/m^{2}K
8. South Window Area: 0; 4; 8; 12 m^2
9. North Window Area: 0; 4; 8; 12 m^2
10. East Window Area: 0; 4; 8; 12 m^2
11. West Window Area: 0; 4; 8; 12 m^2
12. Door U-value: 0.1; 0.7; 1.3; 1.9 W/m^{2}K
13. Door Area: 2; 4; 6; 8 m^2

Since the data is very big, representative training set must be selected. Besides that the training set must be as small as possible. With the above considerations in mind, 5 candidate training sets created. They are with different number of data. The candidate training sets are:

- Training set 1: 2827 data
- Training set 2: 4340 data
- Training set 3: 5405 data
- Training set 4: 6819 data
- Training set 5: 8630 data

To select the best training set, an experiment using the three classifiers is carried out. The experiment is done by means of Weka data mining software. For this experiment we use 10-fold cross validation. The results are depicted in Fig. 1, 2, and 3.

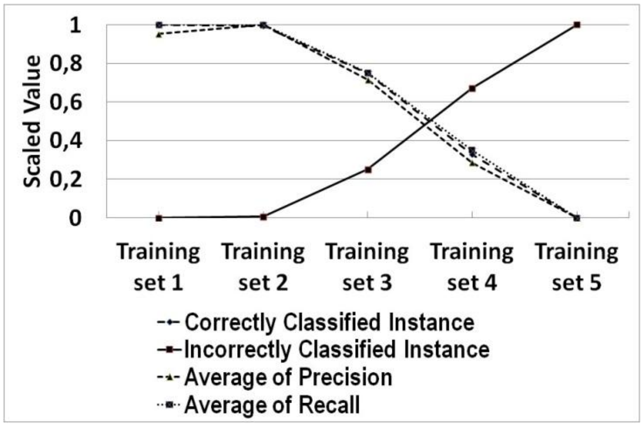


Fig. 1. k-NN performance on different training sets.

Fig. 1 shows performance of k-NN methods using the five training sets. The classifier shows the best performance when using training sets 1 and 2. However, k-NN performance has better precision when using training set 2 than training set 1. Fig. 2 shows performance of Naïve Bayes classifier using the same training sets. Naïve Bayes performs best when using training set 2. This is shown by the highest correctly classified instance and precision, and the lowest incorrectly classified instance.

Meanwhile Fig. 3 shows no performance difference on Decision Tree when using the training sets. From this result, training set 2 is chosen as the working training set.

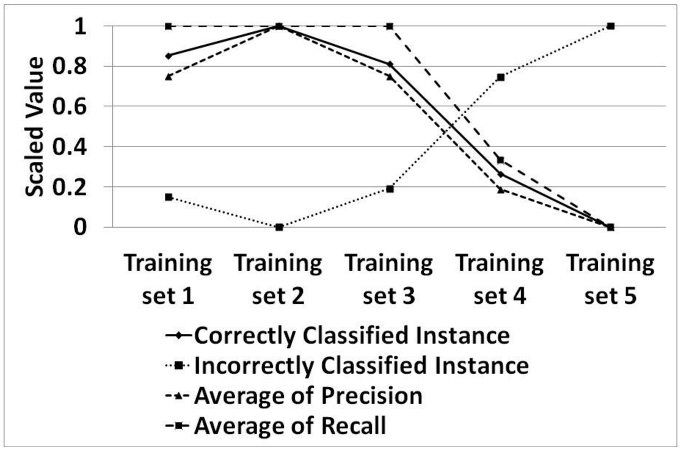


Fig. 2. Naïve Bayes performance on different training sets.

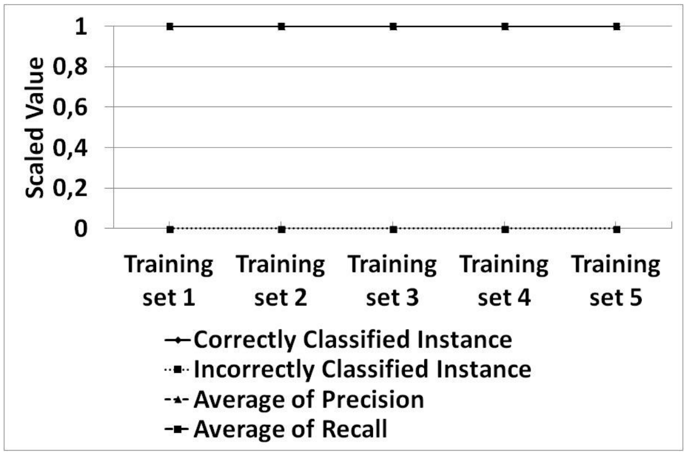


Fig. 3. Decision Tree performance on different training sets.

IV. Experiment

To carry out the experiment, a simple energy simulation tool using the three classifiers (Naïve Bayes, Decision Tree, and k-NN) is developed. For the Decision Tree we use C4.5 algorithm and for k-NNwe use k = 11.We did an experiment using 10 data and for each data, a classification time and performance values are recorded. We should mention here that the time we use is classification time only (without training time). The reason is that K-NN is lazy learner that does not need training. Hence to be fair, the time we use here is only classification time.

Except classification time, the output of the experiment is a confusion matrix. Using confusion matrix, performance parameters of a classifier can be calculated. The performance parameters include: precision, recall, accuracy, F-measure, and area under the curve (AUC).

We use AUC in this experiment because Provost et al., 1998 in [11] state that simply using accuracy results can be misleading. They recommended when evaluating binary decision problems to use Receiver Operator Characteristic (ROC) curves, which show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. This is supported byEntezari-Maleki, Rezaei, Minaei-Bidgoli [12]who state that ROC curve is a usual criterion for identifying the prediction power of different classification methods, and the area under this curve is one of the important evaluation metrics which can be applied for selecting the best classification method.

An ROC graph isactually two-dimensional graph in which True Positive Rate (TPR) is plotted on the Y axis and False Positive Rate (FPR) is plotted on the X axis [13]. It depicts relative trade-offs between benefits (true positives) and costs (false positives). One point in ROC space is better than another if its TPR is higher,FPR is lower, or both[14]. ROC performance of a classifier is usually represented by a value which is the area under the ROC curve (AUC). The value of AUC is between 0 and 1.

The experiment steps are as follows:

1) Enter user data. Values of all 13 parameters are entered. The application then calculates the energy performance. For instance the energy performance of the user data is X W/m^2. The energy performance is calculated using the following formulas:

Le = 1.0 * (wa – wina – da) * wuv + 1.0 * wina * winuv + 1.0 * da * duv (2)

Lu = 0.9 *ra * ruv (3)

Lg = 0.5 * fa * fuv (4)

tl = Le + Lu + Lg (5)

TL = 0.024 * tl * 3235 (6)

Lv = 0.33 * 0.6 *fa * wh * 0.8 (7)

VL = 0.024 * Lv * 3235 (8)

IG = 0.024 * 4 * fa * nof * 208 (9)

SG = 356 * (swa * 0.75) * 0.9 * 0.67 * 0.9 + 150 * (nwa * 0.75) * 0.9 * 0.67 * 0.9 + 210 * (ewa * 0.75) * 0.9 * 0.67 * 0.9 + 210 * (wwa * 0.75) * 0.9 * 0.67 * 0.9 (10)

EP = (TL + VL) – 1.0 * (IG + SG) (11)

where:

Le = exterior loss

wa = wall area

wina = window area

da = door area

wuv = wall u-value

winuv = window u-value

duv = door u-value

Lu = unheated space loss

ra = roof area

ruv = roof u-value

Lg = ground loss

fa = floor area

fuv = floor u-value

tl = thermal loss

TL = transmission loss

wh = wall height

VL = ventilation loss

IG = internal gain

nof = number of floors

SG = solar gain

swa = south window area

nwa = north window area

ewa = east window area

wwa = west window area

EP = energy performance

2) Setting classes of the training set. Every data in the training set having energy performance less than or equal to X W/m^2 is set to class Good, and those having energy performance greater than X W/m^2 is set to class Bad.Note that the attributes of training set are: Wall U-value, Wall Height, Roof U-value, Floor U-value, Floor Area, Number of Floors, Window U-value, South Window Area, North Window Area, East Window Area, West Window Area, Door U-value, Door Area, Energy Performance, Class.

3) Create working data.The working data is created by querying on the raw data. Since there are 13 parameters, there will be 13 queries. The condition on each query is taken from the value of the respective parameter on the user data.The queries are done one after another. It means that the data resulted from a query will be queried again by the next query.This is done 13 times.Note that the attributes of working data are:Wall U-value, Wall Height, Roof U-value, Floor U-value, Floor Area, Number of Floors, Window U-value, South Window Area, North Window Area, East Window Area, West Window Area, Door U-value.

4) Classification. Data from working data is taken one by one. This data is then classified against the training set using one of the three classifiers (Naïve Bayes, Decision Tree, k-Nearest Neighbor). The classification time is recorded starting from the beginning until the end of the classification. After the classification, the energy performance of this data is calculated. Note that the data resulted in this step has the following attributes: Wall U-value, Wall Height, Roof U-value, Floor U-value, Floor Area, Number of Floors, Window U-value, South Window Area, North Window Area, East Window Area, West Window Area, Door U-value, Door Area, Energy Performance, Class, Classification time.

5) Create confusion matrix. Count True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). A data is included in TP if it has energy performance less than or equal to X W/m^2 and class Good. A data is included in TN if it has energy performance greater than X W/m^2 and class Bad. A data is included in FP if it has energy performance greater than X W/m^2 but has class Good. Meanwhile a data is included in FN if it has energy performance less than or equal to X W/m^2 but has class Bad.

6) Select alternative design. Of all data included in TP, the one having the best energy performance will be selected as the alternative design.

V. Result

The classification times of the three classifiers that are used to classify 10 data are shown in Fig. 4.This figure shows that Decision Tree has the fastest classification time followed by Naïve Bayes and k-Nearest Neighbor. The differences between classification time of Decision Tree and Naïve Bayes also between Naïve Bayes and k-NN are about an order of magnitude.

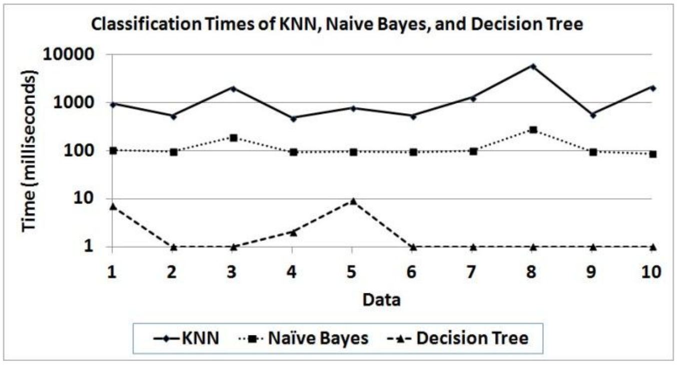


Fig. 4. Classification times of k-NN, Naïve Bayes, and Decision Tree.

The average precisions and recalls for k-NN, Naïve Bayes, and Decision Tree are: 0.819 and 0.543; 0.799 and 0.794; 0.779 and 0.663 respectively(Fig. 5 and 6). Since F-measure is the harmonic mean of precision and recall, hence to know which classifier is the best in terms of precision and recall, we can calculate the F-measure value (Fig. 7). The average F-measure value of Naïve Bayes is the biggest among the three, that is 0.780. Decision tree has average F-measure of 0.676 and k-NN of 0.543. Therefore we can say that Naïve Bayes is the best in terms of precision and recall followed by Decision Tree and k-NN.

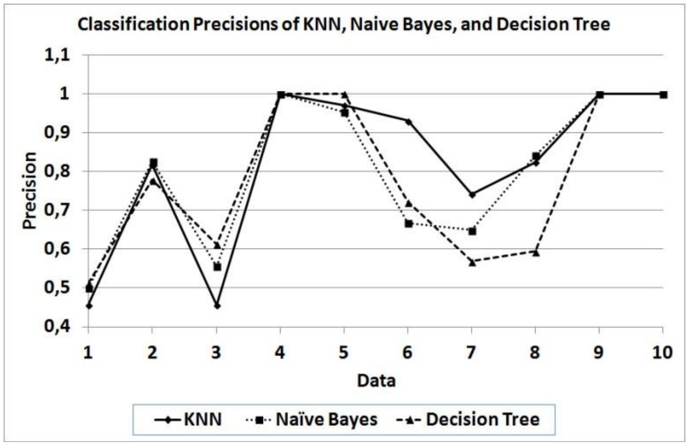


Fig. 5. Classification precision of k-NN, Naïve Bayes, and Decision Tree

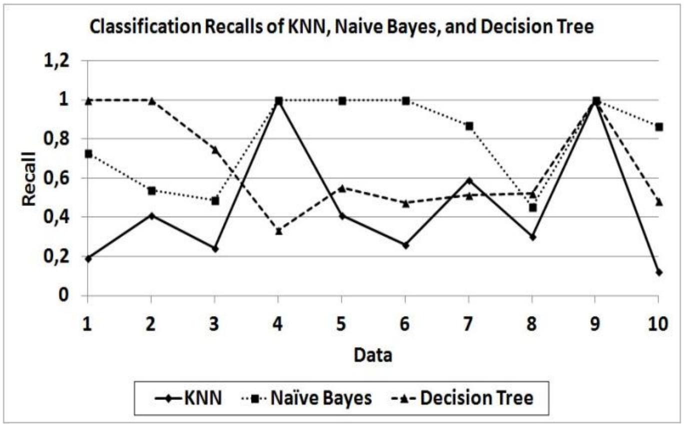


Fig. 6. Classification recall of k-NN, Naïve Bayes, and Decision Tree

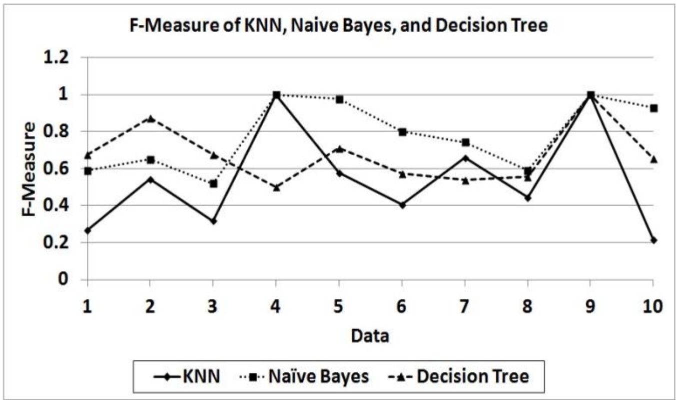


Fig. 7. F-measure of k-NN, Naïve Bayes, and Decision Tree

Naïve Bayes is again the best in accuracy (Fig. 8). Naïve Bayes is the most accurate classifier compared to Decision Tree and k-NN with the average accuracy of 0.737. Meanwhile the average accuracies of Decision Tree and k-NN are 0.589 and 0.567, respectively.

The last parameter for comparing classifier performance is area under the curve (AUC). In this parameter Naïve Bayes is also the biggest among the three classifiers (Fig. 9). The AUC of Naïve Bayes is 0.605, followed by Decision Tree 0.585 and k-NN 0.570.

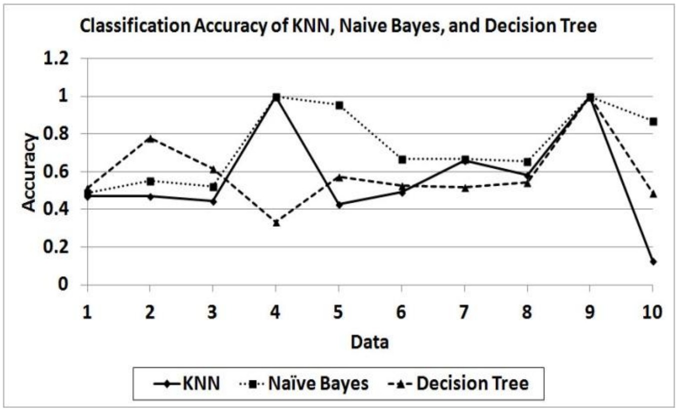


Fig. 8. Classification accuracy of k-NN, Naïve Bayes, and Decision Tree

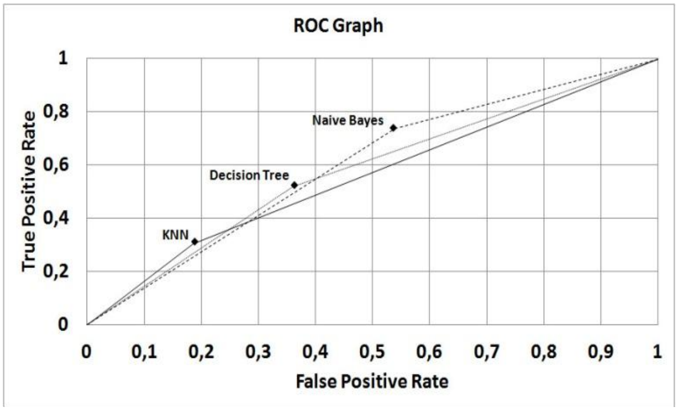


Fig. 9. Area under the curve (AUC) of k-NN, Naïve Bayes, and Decision Tree

VI. Discussion

As stated in the previous section, the experiment we carried out reveals that Naïve Bayes outperforms Decision Tree and k-NN. It is the best in all performance parameters but precision, they are: recall, F-measure, accuracy, and AUC. This result is similar to previous studies.

When comparing Naïve Bayes and Decision Tree in the classification of training web pages, Xhemali,Hinde, and Stone[15] find that the accuracy, F-measure, and AUC of Naïve Bayes are 95.2, 97.26, and 0.95 respectively. This is better than Decision Tree whose accuracy, F-measure, and AUC are: 94.85, 95.9, 0.91, respectively.

Li and Jain [16] investigate four different methods for document classification: the naive Bayes classifier, the nearest neighbour classifier, decision trees and a subspace method. Their experimental results indicate that the naive Bayes classifier and the subspace method outperform the other two classifiers on the data sets. Their experimental results show that all four classification algorithms perform reasonably well; the naïve Bayes approach performs the best on test data set1, but the subspace method outperforms all others on test data set2.

Other studies in references [17] - [20] also obtain the same results when comparing performance of Naïve Bayes and Decision Tree.

A Naive Bayes classifier is a simple classifier. However, although it is simple, Naive Bayes can outperform more sophisticated classification methods. Besides that it has also exhibited high accuracy and speed when applied to large database [3]. Moreover, it is very fast for both learning and predicting. Its learning time is linear in the number of examples and its prediction time is independent of the number of examples [21].Naïve Bayes classifier is also fast, consistent, easy to maintain and accurate in the classification of attribute data [15]. And from computation point of view, Naïve Bayes is more efficient both in the learning and in the classification task than Decision Tree [22].

The reason for good performance of Naïve Bayes is described by Dominggos and Pazzani [23]as follows:"Naïve Bayes is commonly thought to be optimal, in the sense of achieving the best possible accuracy, only when the independence assumption holds, and perhaps close to optimal when the attributes are only slightly dependent. However, this very restrictive condition seems to be inconsistent with the Naïve Bayes' surprisingly good performance in a wide variety of domains, including many where there are clear dependencies between the attributes." In a study on 28 datasets from the UCI repository, they find that Naïve Bayes was more accurate than C4.5 in 16 domains. They further statethat: "the Naïve Bayes is in fact optimal even when the independence assumption is grossly violated, and is thus applicable to a much broader range of domains than previously thought. This is essentially due to the fact that in many cases the probability estimates may be poor, but the correct class will still have the highest estimate, leading to correct classification". Finally they come to conclusion that "the Naïve Bayes achieves higher accuracy than more sophisticated approaches in many domains where there is substantial attribute dependence, and therefore the reason for its good comparative performance is not that there are no attribute dependences in the data".

Frank, Trigg, Holmes, and Witten[24] explain why naive Bayes perform well even when the independence assumption is seriously violated: "most likely it owes its good performance to the zero-one loss function used in classification. This function defines the error as the number of incorrect predictions. Unlike other loss functions, such as the squared error, it has the key property that it does not penalize inaccurate probability estimates as long as the greatest probability is assigned to the correct class. There is evidence that this is why naive Bayes' classification performance remains high, despite the fact that inter-attribute dependencies often cause it to produce incorrect probability estimates".

Meanwhile Zhang [25] explains the reason of good performance of Naïve Bayes as follows:"In a given dataset, two attributes may depend on each other, but the dependence may distribute evenly in each class. Clearly, in this case, the conditional independence assumption is violated, but naive Bayes is still the optimal classifier. Further, what eventually affects the classification is the combination of dependencies among all attributes. If we just look at two attributes, there may exist strong dependence between them that affects the classification. When the dependencies among all attributes work together, however, they may cancel each other out and no longer affect the classification". Therefore, he argues that "it is the distribution of dependencies among all attributes over classes that affect the classification of naive Bayes, not merely the dependencies themselves".

Similar to the result of our study, previous studies also show that k-Nearest Neighbor is worse than both Naïve Bayes and Decision Tree. In their study to classify arid rangeland using Decision Tree and k-Nearest Neighbor, Laliberte, Koppa, Fredrickson, and Rango[26] obtain that the overall accuracy of Decision Tree (80%) is better than that of k-Nearest Neighbor (78%). Pazzani,Muramatsu, and Billsus[27] find that in identifying interesting web sites, the naive Bayesian classifier has the highest average accuracy with 20 training examples: 77.1 (standard deviation 4.4). In contrast, backprop is 75.0 (3.9), k-Nearest Neighbor is 75.0 (5.5), and ID3 is 70.6 (3.6). The only study which shows that k-NN outperforms Decision Tree and Naïve Bayes is by Horton and Nakai[28]. However, they do not have a solid answer as to why k-NN performs better on this task.

The performance of k-NN in this and previous studies is the worst among the three classifiers. Since k-NN uses number of nearest neighbor k as one of the parameter in classifying an object, then this value might affect the performance of the classifier. In their study using k-NN to classify credit card applicants, Islam,Wu, Ahmadi, Sid-Ahmed[29] find that the best performance of k-NN is when k=5. Using this k value, k-NN outperforms Naïve Bayes. Using bigger and smaller k value, the k-NN performance is worst. Meanwhile, Batista and Silva [30] study three parameters affecting the performance of k-NN, namely number of nearest neighbors (k), distance function, and weighting function. They find that for all weighting function and distance function, the performance increases as k increases up to a maximum between k = 5 and k = 11. Then, for higher values of k, the performance decreases. Based on this study, we use k = 11 in this experiment. And the reason why we choose the upper boundary is because larger k values help reduce the effects of noisy points within the training data set [29].The choice is also based on our experiment onk-NN performance with different k values. The k values we use are: 11, 21, 31, 41, and 51. The experiment use 10-fold cross validation. The result is shown in Fig. 10. The figure shows thatk-NN reaches the best performance when we use k = 11. For k values greater than 11, the performance decreases. Since we have not tested the k values smaller than 11, hence it is worth trying to use those values in the future work.

Beside low performance, another weakness of k-NN is slow runtime performance and large memory requirements [31]. The k-NN classifier requires a large memory to store the entire training set [32]. Hence, the bigger the training set, the bigger memory requirement and the larger distance calculations must be performed. This causes the classification is extremely slow. This is the reason why the classification time of k-NN in our experiment is very big, the worst among the three classifiers.

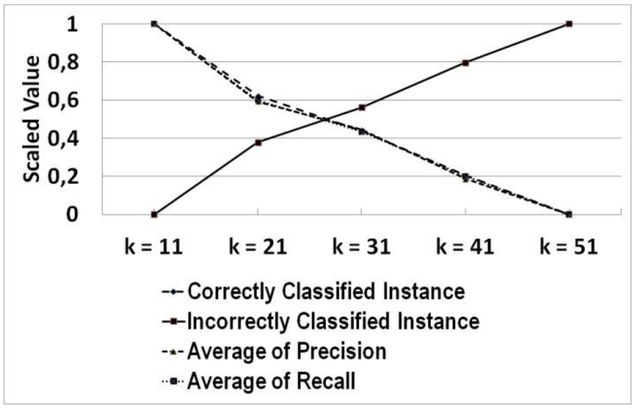


Fig. 10. k-NN performance on different k values.

The fast classification time by Decision Tree is due to the absence of calculation in its classification process. The tree model is created outside the application, using Weka data mining tool. And the model is converted into rules before being incorporated into the application. Classification by way of following the tree rules is faster than the ones that need calculation as in the case of Naïve Bayes and k-NN.

VII. Conclusion

A novel method to search alternative design in an energy simulation tool is proposed. A classification method is used in searching the alternative design. There are three classifiers used in this experiment namely Naïve Bayes, Decision Tree, and k-Nearest Neighbor. Our experiment shows that Decision Tree is the fastest and k-Nearest Neighbor is the slowest. The fast classification time of Decision Tree because there is no calculation in its classification. The tree model is created outside the application that is using Weka data mining tool. And the model is converted into rules before being incorporated into the application. Classification by way of following the tree rules is faster than the ones that need calculation as in the case of Naïve Bayes and k-NN. Meanwhile k-Nearest Neighbor is the slowest classifier because the classification time is directly related to the number of data. The bigger the data, the larger distance calculations must be performed. This causes the classification is extremely slow.

Although it is a simple method, Naïve Bayes can outperform more sophisticated classification methods. In this experiment, Naïve Bayes outperforms Decision Tree and k-Nearest Neighbor. Dominggos and Pazzani[23] state that the reason for Naïve Bayes' good performance is not because there are no attribute dependences in the data. In fact Frank,Trigg, Holmes, and Witten[24] explain that its good performance is caused by the zero-one loss function used in the classification. Meanwhile Zhang [25] argues that it is the distribution of dependencies among all attributes over classes that affect the classification of naive Bayes, not merely the dependencies themselves.

Acknowledgment

ImanParyudi would like to thank Directorate of Higher Education, Ministry of Education and Culture, Republic of Indonesia for the scholarship awarded to him.

References

[1] G. K. Gupta, Introduction to Data Mining with Case Studies. Prentice Hall of India, New Delhi, 2006.

[2] P-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. Addison Wesley Publishing, 2006.

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan-Kaufmann Publishers, San Francisco, 2001.

[4] O. Maimon and L. Rokach, Data Mining and Knowledge Discovery. Springer Science and Business Media, 2005.

[5] X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010.

[6] V. Mohan, "Decision Trees: A comparison of various algorithms for building Decision Trees," Available at: http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf

[7] T. Miquelez, E. Bengoetxea, P. Larranaga, "Evolutionary Computation based on Bayesian Classifier," Int. J. Appl. Math. Comput. Sci. vol. 14(3), pp. 335 – 349, 2004.

[8] M. K. Stern, J. E. Beck, and B. P. Woolf, "Naïve Bayes Classifiers for User Modeling," Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.118.979

[9] Wikipedia, "k-Nearest Neighbor Algorithm," Available at: http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

[10] V. Garcia, C. Debreuve, "Fast k Nearest Neighbor Search using GPU," IEEE, 2008.

[11] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006.

[12] R. Entezari-Maleki, A. Rezaei, and B. Minaei-Bidgoli, "Comparison of Classification Methods Based on the Type of Attributes and Sample Size," Available at: http://www4.ncsu.edu/~arezaei2/paper/JCIT4-184028_Camera%20Ready.pdf

[13] Wikipedia, "Receiver Operating Characteristics," Available at: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

[14] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," Kluwer Academic Publishers, Netherland, 2004.

[15] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," International Journal of Computer Science Issue, Vol. 4(1), 2009.

[16] Y. H. Li and A. K. Jain, "Classification of Text Document," The Computer Journal, Vol. 41(8), 1998.

[17] R. M. Rahman and F. Afroz, "Comparison of Various Classification Techniques," Journal of Software Engineering and Applications, Vol. 6, 2013, 85 – 97.

[18] Z. Nematzadeh Balagatabi, "Comparison of Decision Tree and Naïve Bayes Methods in Classification of Researcher's Cognitive Styles in Academic Environment," Journal of Advances in Computer Research. Vol. 3(2), 2012, 23 – 34.

[19] L. Dan, L. Lihua, Z. Zhaoxin, "Research of Text Categorization on WEKA," Third International Conference on Intelligent System Design and Engineering Applications, 2013.

[20] J. Huang, J. Lu, C. X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy," Third IEEE International Conference on Data Mining, 2003.

[21] M. Pazzani and D. Bilsus, "Learning and Revising User Profiles: The Identification of InterestingWeb Sites," Machine Learning, Vol. 27, 313 – 331, 1997.

[22] N. B. Amor, S. Benferhat, Z. Elouedi, "Naive Bayes vs Decision Trees in Intrusion Detection Systems," ACM, 2004.

[23] P. Domingos, M. Pazzani, "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier," Available at: http://www.ics.uci.edu/~pazzani/Publications/mlc96-pedro.pdf

[24] E. Frank, L. Trigg, G. Holmes, I. A. Witten, "Naïve Bayes for Regression," Machine Learning, Vol. 000, 1 – 20, 1999.

[25] H. Zhang, "The Optimality of Naïve Bayes," American Association for Artificial Intelligence, 2004.

[26] A. S. Laliberte, J. Koppa, E. L. Fredrickson, and A. Rango, "Comparison of nearest neighbor and rule-based decision tree classification in an object-oriented environment," Available at: http://naldc.nal.usda.gov/download/44074/PDF

[27] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying interesting web sites," Available at: http://www.ics.uci.edu/~pazzani/RTF/AAAI.html

[28] P. Horton and K. Nakai, "Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier," ISMB-97 Proceedings, AAAI, 1997.

[29] M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. Sid-Ahmed, "Investigating the Performance of Naïve- Bayes Classifiers and K- Nearest Neighbor Classifiers," Journal of Convergence Information Technology, Vol. 5(2), 2010.

[30] G. E.A.P.A. Batista, D. F. Silva, "How k-Nearest Neighbor Parameters Affect its Performance," Simposio Argentino de Inteligencia Artificial (ASAI 2009), 95 – 106, 2009.

[31] S. D. Bay, "Nearest Neighbor Classification from Multiple Feature Subsets," Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.9040&rep=rep1&type=pdf

[32] Y. Lee, "Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function, and Backpropagation Neural Networks," Neural Computation Vol. 3, 440 – 449, 1991

Modification of Contract Net Protocol(CNP) : A Rule-Updation Approach

Sandeep Kaur
Dept. of Comp. Sci. &Engg.,
BCET,Gurdaspur, PUNJAB,
INDIA

Harjot Kaur
Dept. of Comp. Sci. &Engg.,
GNDU Regional Campus,
Gurdaspur, PUNJAB, INDIA

Sumeet Kaur Sehra
Dept. of Comp. Sci. &Engg.,
GNDEC,Ludhiana, PUNJAB,
INDIA

***Abstract*—Coordination in multi-agent system is very essential, in order to perform complex tasks and lead MAS towards its goal. Also, the member agents of multi-agent system should be autonomous as well as collaborative to accomplish the complex task for which multi-agent system is designed specifically. Contract-Net Protocol (CNP) is one of the coordination mechanisms which is used by multi-agent systems which prefer coordination through interaction protocols. In order to overcome the limitations of conventional CNP, this paper proposes a modification in conventional CNP called updated-CNP. Updated-CNP is an effort towards updating of a CNP in terms of its limitations of modifiability and communication overhead. The limitation of the modification of tasks, if the task requirements change at any instance, corresponding to tasks which are allocated to contractor agents by manager agents is possible in our updated-CNP version, which was not possible in the case of conventional-CNP, as it has to be restarted in the case of task modification. This in turn will be reducing the communication overhead of CNP, which is time taken by various agents using CNP to pass messages to each other. For the illustration of the updated CNP, we have used a sound predator-prey case study.**

Keywords—Multi-Agent System; Coordination; Communication Language; Norm-based Contract Net Protocol; utility parameters

I. INTRODUCTION

A multi-agent system (MAS) can defined to be a system comprised of multiple interacting intelligent agents [14] inside surroundings (environment). It can be accustomed to solve troublesome issues and problems that are not possible for a private agent or a monolithic system to unravel. So, it is having tremendous applications in any problem-solving domain as explained below. Intelligence is basically ability to reason, learn, act and react. Multi-agent system carries with and within it, its surroundings and agents. Generally multi-agent systems research refers to software system agents actively functioning, in order to achieve the goals of a multi-agent system or their individual goals. However, in a multi-agent system the agents could equally well be robots or human-beings. A multi-agent system may contain combined human-agent teams.

A. Applications of MAS to real world

Multi-agent systems [13] are applied in the real world to graphical applications such as computer games. Agent systems have been used in films as well. They are also used for coordinated defence systems. Other applications include transportation, logistics, graphics, GIS , disaster management as well as in many other fields. It is widely being advocated for use in networking and mobile technologies, to achieve automatic and dynamic load balancing, high scalability, and self-healing networks.

In MAS, a single agent alone is not sufficient to solve any complex problem for which actually the MAS are designed, as it has not sufficient resources, information or competence. Therefore, in order to achieve the goals of the system and hence performing the tasks of the system, agent has to coordinate (cooperate) with rest of the agents of the system. And in order to ensure coordination (cooperation) [15], agents communicate with each in MAS by using various communicative acts of communication languages[1][2] like FIPA-ACL (Agent Communication Language), KQML (Knowledge Query and Manipulation Language), FLBC,UCL (Universal Communication Language), FACL (Form-based Agent Communication Language), and DBACL (Database Agent Communication Language). Also, this communication is governed by a set of protocols and coordination mechanisms. One of the best and oldest mechanisms used for coordination is CNP (i.e. Contract Net Protocol) [11][12].

There are numerous issues related to CNP, which have been modified and added later on to it. One is presented by Sun and WU in [6] , in which they have modified CNP by adding the concept of norms to conventional CNP and have termed it as Norm-based CNP by removing the limitation of conventional CNP of being inefficient to handle specialized interactions and hence, coordination.Another one is presented by Elmahalawyin [21], which is called Round Contract Net Protocol (RCont), which is modification of CNP using an acquaintance model. There are many versions of modifications available to CNP, and discussion of all them requires in itself a complete book. This paper is basically contributing to the modification of CNP by modifying one of its rules or phases, i.e.in the task processing phase, modification can be done at the level of manager agent for the task which is already being allocated to contractor agent corresponding to the changes in requirements of set of tasks which are to be performed by manager agents, which otherwise can only be performed by task repetition and restarting CNP from the beginning once again. This in turn will reduce the communicationoverhead or the time of processing required by CNP.

In addition to this, this modification is also studied comparatively in the agent communication languages, FIPA-ACL and KQML. To demonstrate the implementation of this modification; a predator prey case study is used. This paper is organized as follows. Section II is related to introduction of various coordination mechanisms used in MAS, Conventional Contract-Net Protocol (CNP), Norm-based Contract Protocol, as well as description of the limitations present in both of them. Section III is related to description of various agent communication languages, hence highlighting various importance differences between all of them. Section IV is related to the brief description of predator-prey case study and work done by us for modification of CNP by using proposed approach, and results derived out of this modification. Section V summarizes the conclusions.

II. Coordination In Mas

A. Coordination

Coordination [15] comprises of a set of mechanisms necessary for the effective operation of Artificial Agent Societies (AASs). It is also defined as other process of managing dependencies between activities. Amongst its fundamental components, are the allocation of scarce resources; communication between the agents about intermediate results, coordination goals, capabilities and plans, status of the different aspects of the environment as well as providing some meta-level information. Coordination is required and is normally available also in cases in which there is not full cooperation amongst the agents or groups of agents. In a human society, for example, competition is constrained by consumer protection, various government agencies and antitrust laws. People and organizations antagonistic to one another may interact via prescribed legal channels. Coordination theory can be defined as a set of axioms, mathematical and logical constructs, and analytical techniques used to create a model of dependency management in AASs.

B. Types of Coordination in MAS

According to Bergentti and Ricci [16], there are basically three main coordination approaches used in any MAS for managing coordination amongst a set of agents and they are based on the use of

- Tuple centres;
- Interaction Protocols; and
- Semantics of ACLs

In our case, for exercising coordination in MAS, agents are using Contract Net Protocol (CNP), which is one of the best approaches used for coordination, and is described in subsection below.

C. Contract Net Protocol (CNP)

In multi-agent system, in order to accomplish any task agents need autonomy and collaboration (in terms of coordination). Contract net protocol (CNP) is coordination mechanism often used in a multi-agent system so as to coordinate amongst a set/group of agents. The original Contract Net Protocol (CNP) , was originally developed by Smith and Davis [11].

CNP works like a business market where manager agent asks for bids from the contractor agents and then awards tasks to suitable contractor agent. In CNP, tasks are accomplished by breaking them down into sub-tasks by manager agent and then asking for bids for those sub-tasks from the contractor agents. Contractor agents replies with bids or refuse within a given deadline. Once deadline is reached, manager agent awards the task to the most suitable contractor agent having lowest bid.

Another minor modification of CNP is FIPA-Contract-Net-Protocol [5], in which there is addition of rejection and confirmation communicative acts. For the detailed study of CNP, the readers can refer[5].

Various limitations of Conventional CNP are in the form of attributes like responsiveness, load balancing, and fairness, utilization of resources, communication overhead, robustness, modifiability and scalability [17]. Although, all of them cannot be improved at the same time by enhancing CNP because this enhancement in itself will become a very complex problem. Here in our paper, we have tried to update CNP by modifying its rules used in communication to decrease the communication overhead, which will in turn increase the efficiency of CNP. This updated version of CNP is basically implemented using the predator-prey case study by modifying the rules of communication for predator agents.

III. Agent Communication In Mas

For achieving collaboration and hence coordination in MAS, agents interact with each other. And, for communication, they use various communicative acts of agent communication language (ACL), as; possibility for different agents to interact in an open environment heavily depends on the adoption of a common, standard Agent-Communication Language. The two most-widely used ACLs in practice are KQML and FIPA-ACL. But neither have yet been considered as standards as they are not capable of letting heterogeneous agents communicate as there are numerous other agent communication languages like Universal Communication Language(UCL), Database Agent Communication Language(DBACL), Formal Language For Business Communication(FLBC), Form –Based ACL(FACL)) available for agent communication. These languages are actually application-specific languages, as their use varies from application-to-application. Like any other communication language, an Agent-Communication Language (ACL) also includes the definition of the syntax and the definition of the semantics.

Definition of the syntax is the way in which single words are put together and the definition of the semantics is the meaning of the communicative acts. By means of an agent-communication language, an agent can coordinate, communicate and exchange knowledge with other agents despite differences in their hardware platforms, operating systems, architectures, programming languages and representation and reasoning systems. Language is assumed to be the fundamental component of every interaction or communication. In a multi-agent environment, agents “talk” to each other by using an agent communication language.

For implementation of modified CNP, we have used a predator-prey case-study[3]. And, we have studied it comparatively, being implemented in FIPA-ACL and KQML languages. Therefore, this implementation and comparison will be incomplete without a brief introduction to these languages, which is described in subsections.

A. FIPA-ACL

The FIPA (Foundation for Intelligent Physical Agents) - Agent Communication Language (ACL) is based on speech act theory [13]: messages are actions, or communicative acts, as they are intended to perform some action by virtue of being sent. The specification consists of a set of message types and the description of their pragmatics (linguistics), i.e. the effects on the mental attitudes of the sender and receiver agents. Every communicative act is described with both a narrative form and a formal semantics based on modal logic. The specifications embrace steering to users who are already familiar with KQML in order to facilitate migration to the FIPA- ACL.The specification also provides the normative description of a set of high-level interaction protocols, including requesting an action, contract net protocol and several kinds of auctions [4][5].

B. KQML

KQML (Knowledge Query and Manipulation Language) is complementary to work on representation languages for domain content, including the DARPA Knowledge Sharing Initiative's Knowledge Interchange Format (KIF)[9]. KQML has also been used to transmit object-oriented data, and a wide range of information can be accumulated using it. KQML is a language for programs which use to communicate attitudes about information, such as querying, stating, believing, requiring, achieving, subscribing, and offering. KQML is indifferent to the format of the information itself, thus KQML expressions will often contain sub-expressions in other so-called content languages [10].

All the communicative acts, which are used by agents in MAS for interacting with each other, are governed by a set of rules and regulations, which are termed as agent communication protocols, as described in the next subsection below:

C. Agent Communication Protocols

Communication protocols [4] are widely recognized as a major and efficient concept to support many forms of interaction among agents, such as information sharing, task sharing, resource sharing, and coordination of actions, conflict resolution or commitments. When a set of agents interact through a protocol, each one is assumed to know when it may or must perform a communication act and what will be the effect of this performance[7][8]. Thus a protocol is a behavioural structure defined by:

- a set of (types of) communication acts feasible by agents;
- a set of roles that are played by agents;
- a set of behavioural rules stating under which circumstances an agent playing a particular role may or must perform a particular communication act. When an agent engages in a protocol, it chooses a role and commits to obey the protocol's rules.

Here, in this paper, for illustration of communicative acts and hence communication in FIPA-ACL and KQML, we are using once Contract Net Protocol (CNP) and then later on, its updated version in predator-prey case-study which is described in next section. Also, after this the comparative results of CNP and its updated version in both FIPA-ACL and KQML are analyzed graphically using various parameters used in CNP and these communication languages.

IV. Work Done

A. A Predator-Prey Case Study

We have used Predator-Prey case study for implementing the Contract-Net interaction protocol (FIPA-CNP) in FIPA-ACL and KQML languages, using their available per formatives (communicative acts), once using CNP and then its updated version, which is presented by us. The reason for using Predator-Prey system as a case study for studying the interaction in multi-agent systems is because it is very difficult to implement a real-world multi-agent system and study agent interaction in it. So, a directed test-bed or toy-domain like predator prey system is selected as a case study.

The Predator- Prey system is a pure-pursuit domain which involves multiple goal-oriented predator agents, prey agents and environment. The goal of predator agents is to chase and capture prey agents before it reaches the goal. The prey agent's goals are simply to evade the predator for a period of time, or to find and enter a goal square before they are captured. In the course of achieving goal, predator agents need to communicate with each other for passing information in space as they can communicate with each other about prey's location and form good strategy of capturing it. For the coordination between predator agents, we have used Contract-Net Protocol implemented once with conventional rules and secondly, using updated rules for decreasing the communication overhead in FIPA-ACL and KQML both. Both predators and prey cooperate to solve their "goals". The game takes place on an arbitrarily sized grid of squares [3].

Each predator or prey agent initially was completely autonomous. We have achieved this be writing and defining separate JAVA classes for both predator and prey agents. Later on, also they are also given ability to communicate through communicative acts of FIPA-ACL and KQML. Both of these are briefly described in the previous section. For implementation of these predator and prey agents, we have used JAVA-based platforms JADE (Java Agent Development Environment)[18][19] and JATlite [20].

B. Description of Updated-CNP

The conventional Contract-Net Protocol [11], coordinating the communication of contractor and manager agents, comprises of five phases as described below, In this we have also added our additional step to make it updated and efficient in its last and fifth phase which is also described after the description of CNP:

1) Task Announcement:This phase is related to task announcement preparation by the manager agents for issuing them to every agent. This phase comprises of subtasks such as task abstraction, bid specification and expiration time specification. Task abstraction is the description of information related to tasks in abstract form which are to be *performed by contractor agents, for instance task name, task content description, bid specification is specifies the mandatory requirements which are to be fulfilled by contractor agents in order to be eligible to bid. Expiration time is the deadline for accepting the bid. This phase is also said to formulate in general, CFP(Call for Proposal).*

2) Task Announcement Processing : In this phase, according to the type of the task, the manager agent maintains a rank-ordered list of announcements that have been received and have not yet expired. Only those announcements which satisfy the bid specification criteria, will be one, to be allocated a rank and order in the list.

3) Bidding :This phase is related to contractor agents, in which they evaluate the tasks which are announced by manager agents and bid accordingly, if they meet all the necessary requirements specified in bid specification critieria.

4) Bid Processing :This phase is performed at the end of manager agent, who after receiving all the bids from the qualifying contractor agents, will evaluate them according to the task specification template, and then will process all the bids by ranking them. Then, from the ranked list or set, bids with the lowest cost are selected and tasks are allocated to them. This is further done, by informing to the contractor agents who had sent those bids, that their bids have been selected, and these particular set of agents will be now responsible for performing that particular set of tasks allocated to them. This all is performed by manager agents by making use of announced award message.

5) Contract Processing, Reporting Results and Termination : This is the last and the final phase of CNP, which actually marks the completion of CNP and its usage by contractor and manager agents for their task completion. In this phase, all the contractor agents who are allocated contracts, they are working for it, to complete the task allocated to them by means of the contract. During the processing of the contract, an information message is used for the general communication, whenever, it is occurring at any instance between the contractor and the manager agents. Also, in addition to that according to our updated-CNP, the task modification if any is required at the manager agent will be done be the same in the form of step 5 a) mentioned below, and for this again, manager agent will be informing contractor agent by making use of information message:

*a) Task Modification in CNP:*While the contract is processed by contractor agent in CNP, an interim report is sent by contractors to the manager agent, which will be summarizing work in progress or partially executed tasks which are being performed by the contractor agents. After, all the tasks are completed, a full report called final report is being submitted by contractor agents to manager agents, which will be containing a summary of all tasks which are allocated to contractors , there deadline, along with the time of completion.

But before, the final report is to be dispatched by contractor agent, the manager agents can send task change request to the contractors, if any change in task processing is required, i.e., any time before tasks are accomplished and results are sent back to the manager. Hence, this will be saving extra efforts incurred by task repetitions, if in case task change is done and again the commiuncation between the managers and the contractors is to be started from scratch. The changes in task can be anything from a contract setup between contractor and worker agents, or changes in bidding specification, as managers are only communicating with the eligible agents. Therefore, with this repetitions and hence wastage of efforts can be saved by changing requirement of tasks or terminating task execution at all. This will save communication overhead, which occurs in case of Conventional CNP if task repetition is done.

C. Implementation with Case-Study

We have implemented the predator-prey case-study, once with conventional CNP and then with updated CNP in both FIPA-ACL and KQML langauges. In case of old CNP it is not possible to change tasks once they are allocated to all predators to capture prey but in updated CNP we are asking some of the predator agents to change their goal and capture a specific prey.

In Predator prey system, an agreement of predators is intially done with environment where environment sends out prey's details while tasks of capturing prey are still in process. Environment can change the task of capturing prey to revoke chasing prey and work on something else like ignoring some of the preys (this is specifically in the case of stronger predator agents) and going after only a specific prey, which is more dangerous than others. In our implementation, it is the same set of predator agents, out of which few are working as manager predator agents and rest as contractor predator agents. All the feedback related to the prey agents in this case is provided by an environment. So, the initiation of CNP is between predator agents only. In case of old CNP it is not possible to change tasks once they are allocated to all predators to capture prey but in updated CNP, we are asking some of the predator agents to change their goal and capture a specific prey.

For the implementation of the above mentioned case-study and checking the performance of conventional and updated CNP in both FIPA-ACL and KQML languages, we have implemented predator and prey agents and hence system, once in JADE and then JATlite, these agents are communicating using communicative acts of once FIPA-ACL and then KQML. The coordination between predator agents for speeding up the process of prey-catching is done using CNP and updated-CNP. After the implementation, the performance is analyzed comparatively using graphs.

D. Results

The comparison of CNP and updated CNP by using three parameters, i.e., Updated Tasks, Tasks Repetitions and communication overhead in terms of time elapsed is illustrated in graphs, which are giving the summarized results of the execution of the predator-prey case study created with JADE and JATlite platforms for FIPA-ACL and KQML languages.The first set of graphs is related to performance of CNP and updated-CNP in case of predator-prey case study implemented in JADE, in which communication between predator and prey agents is done using FIPA-ACL performatives.

Graph1 in Figure 1 shows performance of CNP and updated-CNP for FIPA-ACL, while updation of tasks is done from the side of manager predator agents for contractor predator agents while CNP is being processed. Taken a set of 5 tasks, 2 were changed during execution.As, updated CNP accommodated almost all task changes immediately as compared to Conventional-CNP, so time was saved while processing of coordination between a set of predator agents, trying to chase and catch a prey.

Graph 2 in Figure 2 shows the performance of CNP and updated-CNP when task repetition was performed by contractor predator agents, the requirements corresponding to those tasks changed while CNP was in execution, so they were rescheduled by manager predator agents in conventional CNP, but in updated-CNP this change was absorbed within the protocol communication.

Graph 3 in Figure 3 shows the comparative performance of CNP and updated-CNP in terms of time taken for accomplishment of tasks after execution of tasks (5 tasks), including the changes made in tasks in between task execution.

The second set of graphs is related to performance of CNP and updated-CNP in case of predator-prey case study implemented in JATlite, in which communication between predator and prey agents is done using KQML performatives.

Graph 4 in Figure 4 shows performance of CNP and updated-CNP for KQML, while updation of tasks is done from the side of manager predator agents for contractor predator agents while CNP is being processed. Taken a set of 5 tasks, 2 were changed during execution. As, updated CNP accommodated almost all task changes immediately as compared to Conventional-CNP, so time was saved while processing of coordination between a set of predator agents, trying to chase and catch a prey.

Graph 5 in Figure 5 shows the performance of CNP and updated-CNP when task repetition was performed by contractor predator agents, the requirements corresponding to those tasks changed while CNP was in execution, so they were rescheduled by manager predator agents in conventional CNP, but in updated-CNP this change was absorbed within the protocol communication.

Graph 6 in Figure 6 shows the comparative performance of CNP and updated-CNP in terms of time taken for accomplishment of tasks after execution of tasks (5 tasks), including the changes made in tasks in between task execution.

V. Conclusions

Coordination and Communication are two vital parts of MAS for its proper functioning, i.e. for performing a set of complex tasks and in order to fulfill its goals. For exercising coordination in MAS, it uses a set interaction protocols, CNP is one of them. The motive of CNP is to enhance communication between a set of agents (in the form of contractor and manager) which are using it. In case, if the task modification is required at the end of contractor agent for the task which is allocated by manager agent, then, it is only possible in conventional CNP after the termination of the protocol. Because, there is no means present in conventional CNP for task modification, only task repetition can be performed.

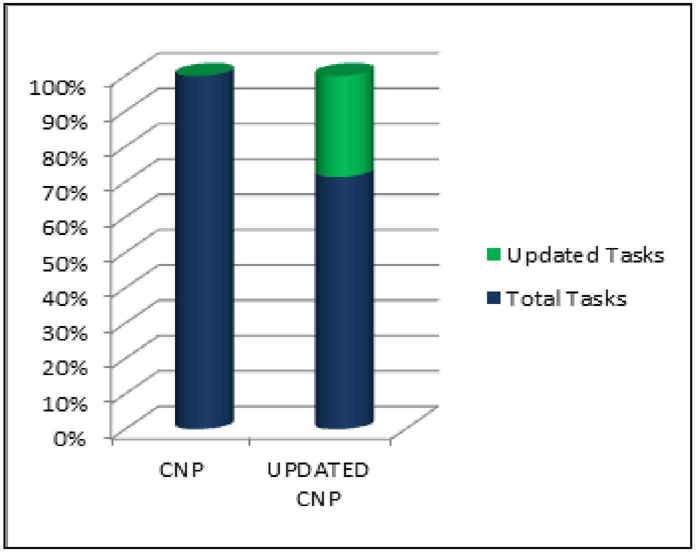


Fig. 1. Performance of CNP and updated CNP for task changes during CNP execution for FIPA-ACL.

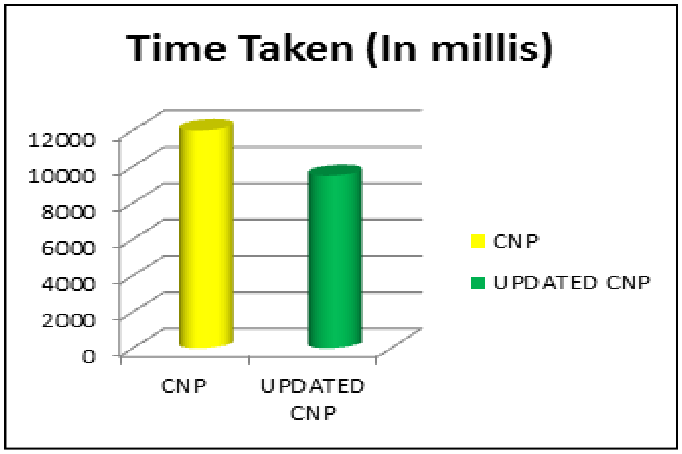


Fig. 2. Performance of CNP and updated CNP for task repetition during CNP execution for FIPA-ACL.

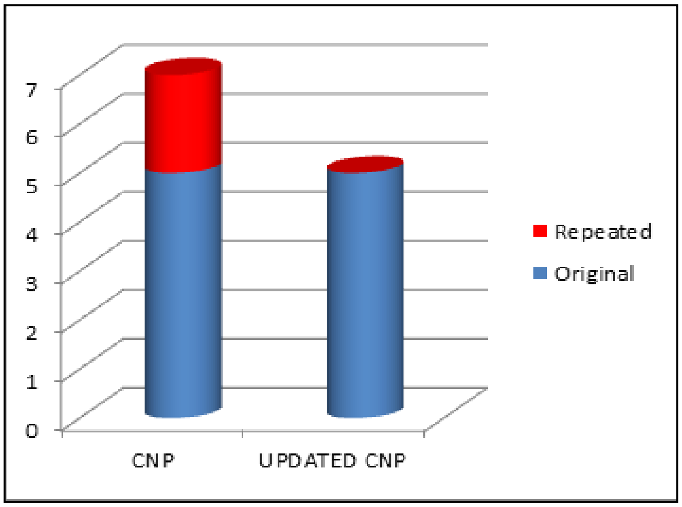


Fig. 3. Performance of CNP and updated CNP in terms of time for CNP execution for FIPA-ACL

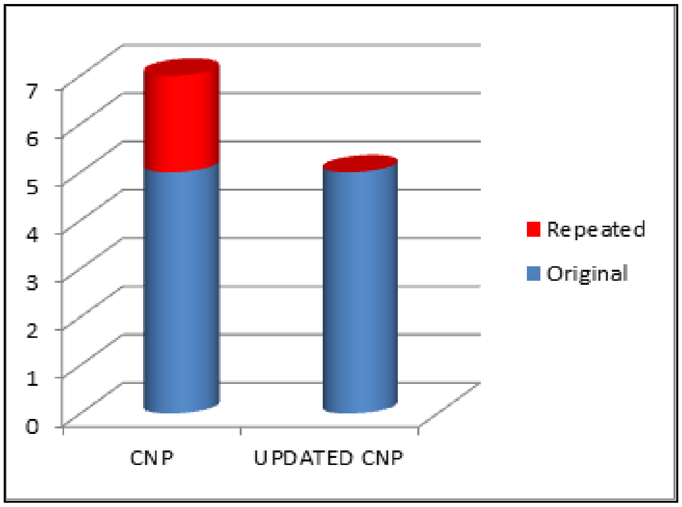


Fig. 4. Performance of CNP and updated CNP for task changes during CNP execution for KQML.

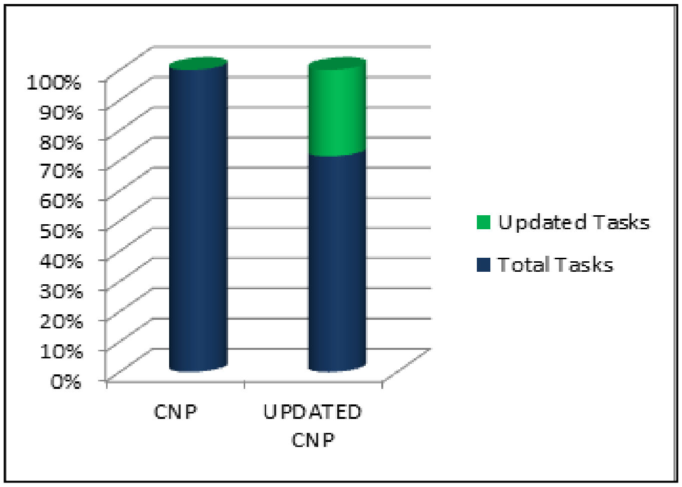


Fig. 5. Performance of CNP and updated CNP for task repetition during CNP execution for KQML

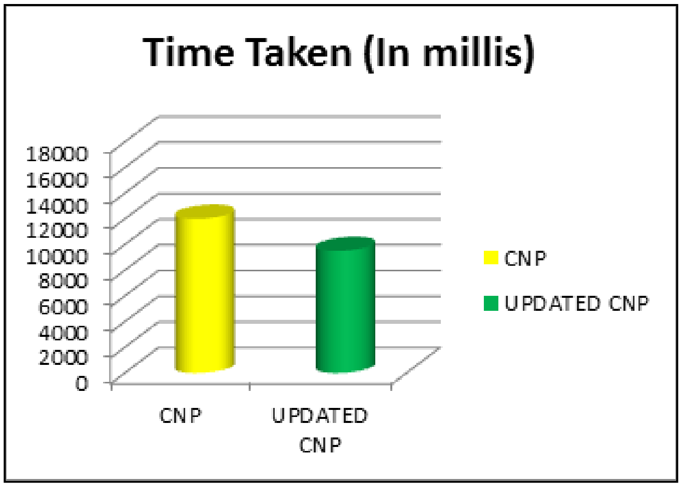


Fig. 6. Performance of CNP and updated CNP in terms of time for CNP execution for KQML

Here, in this paper, we have proposed and implemented an updated version of CNP, called updated-CNP. In updated-CNP, we have added an additional step into a conventional CNP, by the means of if any changes are to be made in task which is allocated to contractor agent by manager agent can be modified during the processing of the task before the final report of completion is send by contractor agent. This is save the overhead of restarting the process of CNP execution between a set of agents who wish to coordinate in order to achieve a certain objective. This will in turn improve the efficiency and effectiveness of protocol, in case there is frequent change in requirement set of manager agents, which, in turn requires task modification for contractor agents.

VI. Future Work

In future, this work can be extended for other agent communication languages like FLBC, UCL and DBCL to know the performance of CNP and updated CNP, in case of application specific case studies. Then, comparison can also be performed for all these languages to check the relative performance of CNP and updated CNP in these languages.

In addition to this, scalability, load balancing, responsiveness, fairness, utilization of resources, robustnessand other limitations can be worked upon this using updated-CNP which we have demonstrated in this paper for different case studies depending upon the use of CNP in that case study. Also, all of these cannot be improved at the same time by enhancing CNP, because this enhancement in itself will become a very complex problem. Therefore, incremental enhancement of CNP can be done by removing one limitation at a time.

References

[1] B. Draa, Chaib B. and F. Dignum, Trends in agent communication languages, Computational Intelligence, Volume18, No.2.

[2] S.Vaniya, B Lad, S Bhavsar. " A Survey on Agent Communication Languages" 2nd International Conference on Innovation, Management and Service (ICIMS) -Singapore, September 2011.

[3] T. P. Rew , Multi-Agent Systems Case Study: Effects of Inter-Agent Communication in the Predator-Prey Domain,Clarkson University, Dept. of Electrical and Computer Engineering, Potsdam, NY 13699 , unpublished.

[4] T.S. Bushby, and H. Michael, Standardizing EMCS Communication Protocols, ASHRAE Journal. Vol. 31, No. 1 pp. 33 - 36 ,1989.

[5] Agent Communication Languages, 1997 FIPA - Foundation for Intelligent Physical Agents Version 1.0 Part 2 Publication date: 10th October, 1997.

[6] D. Sun, J. Wu., Multi-agent Coordination Based on Contract Net Protocol, International symposium on Intelligent Ubiquitous Computing and Education 2009, IEEE Press, pp. 353 – 357,15-16 May 2009.

[7] B. Marzougui, K. Barkaoui, Interaction Protocols in Multi-Agent Systems based on Agent Petri Nets Model, IJACSA, Vol. 4, No.7, 2013.

[8] C. Sibertin-Blanc, C. Hanachi, J. Cardoso, Communication Protocols as First-class Components of Multiagent Systems, Proceedings of Fourth International Conference onMultiAgent Systems, ICMAS 2000, pp. 437-438.

[9] Y.Labrou, T.Finin, Y.Peng, The Current Landscape of ACL, Intelligent Systems,Vol.14,No.2,1999.

[10] Y.Labrou,T.Finin,Semantics and Conversation for an Agent Communication Languages 1997, IJCAI-97, Vol. 1, No. 1,pp.584-591.

[11] R. G. Smith, The Contract Net Protocol, IEEE Transactions on Computers, Vol C-29, No. 12, Dec 1980.

[12] [12] L. Xu, H. Weigand, The evolution of contract net protocol, WAIM 2001, KNCS 2118, Springer, pp. 257-264, 2001.

[13] S. Russell and A. Norvig, A Modern Approach to Artificial Intelligence, Chapter 2, 1995.

[14] Hyacinth S. Nwana, Software Agents : An overview, Vol. 11, No. 3, pp 1-40, Sept 1996, Cambridge University Press.

[15] N. V. Findler, R.M. Malyankar, Social Structures and the Problem of Coordination in Intelligent Agent Societies, 16-th IMACS World Congress 2000.

[16] F. Bergenti, A. Ricci, Three Approaches to the Coordination of Multiagent Systems,In Proceeding SAC '02, Proceedings of the 2002 ACM symposium on Applied computing, pp. 367-372,ACM New York, NY, USA .

[17] E. Bozdag, A Survey of Extensions to the Contract Net Protocol, Technical report, CiteSeerX - Scientic Literature Digital Library and Search Engine. [http://citeseerx.ist.psu.edu/oai2] (United States). 2008.

[18] F. Bellifemine, G. Caire, A. Poggi, G. Rimassa, JADE- Java Agent Development Frammework., A White Paper, exp - Volume 3 - n. 3 - September 2003.

[19] F. L. Bellifemine, G. Caire, D. Greenwood, Developing Multiagent Systems using JADE, , Wiley & Sons, February 2007.

[20] H. Jeon, C. Petrie, M. R. Cutkosky, JATlite: JATLite: A Java Agent Infrastructure with Message Routing, Stanford Center for Design Research (CDR) Stanford, CA.

[21] A. M. Elmahalawy, Applying a new Communication Technique among Agents in Hospital Management System, International Journal of Computer Applications, Volume 75, Number 5, 2013.

Blind Turing-Machines: Arbitrary Private Computations from Group Homomorphic Encryption

Stefan Rass
System Security Group, Institute of Applied Informatics
Alpen-Adria Universität Klagenfurt
Klagenfurt, Austria

***Abstract*—Secure function evaluation (SFE) is the process of computing a function (or running an algorithm) on some data, while keeping the input, output and intermediate results hidden from the environment in which the function is evaluated. This can be done using fully homomorphic encryption, Yao's garbled circuits or secure multiparty computation. Applications are manifold, most prominently the outsourcing of computations to cloud service providers, where data is to be manipulated and processed in full confidentiality. Today, one of the most intensively studied solutions to SFE is fully homomorphic encryption (FHE). Ever since the first such systems have been discovered in 2009, and despite much progress, FHE still remains inefficient and difficult to implement practically. Similar concerns apply to garbled circuits and (generic) multiparty computation protocols. In this work, we introduce the concept of a blind Turing-machine, which uses simple homomorphic encryption (an extension of ElGamal encryption) to process ciphertexts in the way as standard Turing-machines do, thus achieving computability of any function in total privacy. Remarkably, this shows that fully homomorphic encryption is indeed an overly strong primitive to do SFE, as group homomorphic encryption with equality check is already sufficient. Moreover, the technique is easy to implement and perhaps opens the door to efficient private computations on nowadays computing machinery, requiring only simple changes to well-established computer architectures.**

Keywords—secure function evaluation; homomorphic encryption; chosen ciphertext security; cloud computing

I. Introduction

Many security systems at some point employ trusted parties (e.g., trust-centers, smartcards) to perform computations on secret (confidential) information. Trying to get rid of such trusted instances in a security system is often difficult (if not impossible), and one possible solution is to emulate the trusted party by a collection of instances rather than a single one. Such distributed computations usually rely on secret-sharing techniques, capable of function evaluation such that only a permitted specified set of coalitions can learn any secret information or results of the computation. The work of Gennaro[1][2] and [3]made significant contributions to the theory in this area known as *secure multiparty computation* (SMC). Its practical usefulness, however, is somewhat limited, as it isoften tied to a vast communication effort and intricate additional security precautions (e.g., pairwise confidential channels, broadcast channels, etc.). Furthermore, it is a special case ofthe more general problem known as *secure function evaluation* (SFE), in which a single (potentially malicious) instance is made to compute some function on externally supplied (potentially encrypted) inputs. This is the area where this work falls into, and on which we will exclusively concentrate us in the following.Commoncomputational models upon which SFE is based are Turing-machines or circuits, where the appropriateness of each model depends on the details of the SFE technique. We will base our construction on Turing-machines, drawing strongly from circuit complexity models to ease life in cryptographic matters.

Related work: One famous approach to SFE, leaving the computations with a single not necessarily trusted entity,is provided by Yao's *garbled circuits* (GC) [4]. Here, the computational model are circuits, which are good for hardware-implementation (as well as theoretical treatment), yet somewhat difficult to apply in a generic fashion to handle inputs of arbitrary size. Despite much progress in this direction [5]-[7] as well as on applications of GC for SMC [8], [9], only uniform circuits can be set up effectively in practice, in which case they are essentially equivalent to Turing-machines. However, there is so far no analogous concept of a garbled Turing-machine.

Without doubt, the most powerful (and recent) solution to SFE is fully homomorphic encryption (FHE). In brief, this is (or can be) a trapdoor one-way automorphism$E: (R, +, \cdot) \times \mathcal{K} \to (R, +, \cdot)$ where R is a ring or a field, and $\mathcal{K}$ is the key-space. We denote the encryption $E(m, k)$ of m under the key k as $E_k(m)$ hereafter. The central property of FHE is its compatibility with arithmetic operations in the sense that for any two plaintexts m_1, m_2 and any key k, we get $E_k(m_1) + E_k(m_2) = E_k(m_1 + m_2)$ and $E_k(m_1 \cdot m_2) = E_k(m_1) \cdot E_k(m_2)$. That is, arithmetic manipulations done to ciphertexts identically apply to the underlying plaintexts. It is easy to imagine that such an encryption enables any kind of data processing given ciphertexts only, which is exactly what secure function evaluation means.While it is usually simple to get a homomorphic property w.r.t. addition *or* multiplication (e.g., standard encryptions such as RSA or ElGamal are multiplicatively homomorphic; in general *group homomorphic*), homomorphy w.r.t. both operations is intricate and has only recently been achieved [10]. Since this breakthrough, FHE has evolved into a major research branch of cryptography, with many interesting results [11]-[18].

Our contribution in this work is to show that despite the theoretical beauty of FHE, it is nevertheless an overly strong primitive for secure function evaluation. To this end, we investigate the weaker notion of *public-key encryption with equality check* (PKEET) [19], and show how the functionality of a basic (single two-way infinite tape) Turing-machine can be implemented with simple homomorphic encryption that allows equality checks. We call the resulting computing model a *blind Turing-machine*, since it works on encrypted tape content only, doing its transitions by virtue of equality checks, and manipulating the tape content using the homomorphic properties of the encryption in charge. Hence, the TM does not see (in plaintext) any of the content that it processes.

The rest of this article is organized as follows: we start from the PKEET system of [19], which is secure under adaptive chosen ciphertext attacks, and as such cannot be in any sense homomorphic. To restore the homomorphic property in the framework of [19], we describe a generic technique (based on [19]) to construct a *homomorphic public-key encryption with equality check* (hereafter abbreviated as HPKEET) from any additively homomorphic encryption. We prove it secure under (non-adaptive) chosen ciphertext attacks (stronger notions are provably unachievable for any homomorphic encryption in general). Section II formally introduces the respective models, with the construction of HPKEET and its security analysis to follow in section III. Blind Turing-machines (BTM) are introduced in section IV, based on a brief review of how conventional Turing-machines (TM) are formally defined. Security and complexity of computations on such blind TM are studied along a sequence of subsections.

In section VII we report on a practical implementation of the encryption. Remarks on future work and open problems follow in section VIII.

II. Definitions

We write $x \overset{r}{\leftarrow} X$ to denote a uniformly random draw of an element x from a set X. We let $|x|$ denote the length of x in bits (assuming a canonical string representation, if x is a group element). Our treatment in the following is non-uniform. That is, we consider the *complexity* of an algorithm as the *size*, i.e. the number of gates, of a circuit representing the algorithm. To handle inputs of varying length, a *circuit* (e.g., adversary) of complexity τ is thus to be understood as a sequence of circuits (circuit family) $(C_n)_{n\in\mathbb{N}}$, whose size is a function $\tau(n)$, whenever the circuit C_n has n input gates. Besides circuit complexity, section V will heavily rest on time-complexity considerations. To distinguish the two notions from each other, we will refer to *circuit complexity* simply as *complexity*, as opposed to *time-complexity*, always carrying the prefix „time". To further clarify which concept is in charge, we will speak of *circuits* to mean circuit families and circuit complexity, and *algorithms* when we consider time-complexities.

A public-key encryption scheme is a triple of circuits(G, E, D). The circuit G generates the secret and public key pair, denoted as $(pk, sk) \leftarrow G$. Forhomomorphy, assume that the encryption function $E: \mathbb{G}_m \times \mathcal{K} \to \mathbb{G}_c$ is defined on a cyclic plaintext group $(\mathbb{G}_m, +)$, keyspace $\mathcal{K} \subseteq \{0,1\}^*$ and cyclic ciphertext group $(\mathbb{G}_c, \cdot)$. Abbreviating the encryption of a plaintext m under the public key $pk$by $E_{pk}(m)$, we require group homomorphy under identical public keys, i.e., $E_{pk}(m_1) \cdot E_{pk}(m_2) = E_{pk}(m_1 + m_2)$ for all $m_1, m_2 \in \mathbb{G}_m$. The function $D: \mathbb{G}_c \times \mathcal{K} \to \mathbb{G}_m$ decrypts a ciphertext c upon given the secret key sk; denoted as $D_{sk}(c)$.

Security of an encryption is commonly defined in terms of indistinguishable ciphertexts under differently strong attack scenarios. However, an indistinguishability requirement is obviously useless once we endow an encryption with comparison facilities for plaintexts that work on ciphertexts only (as we attempt here). To fix this, we additionally introduce an *authorization* function (circuit) that outputs a (secret) *comparison key*, hereafter called a *token*, which enables comparisons, while any party not knowing the token will be unable to distinguish any two given ciphertexts. In that sense, we consider two different kinds of attacker (following the framework of [19]), both of which are given all system parameters and public keys:

Type 1 attacker: This one can do ciphertext comparisons, in which case we can only ensure the cipher to be one-way but not indistinguishable.

Type 2 attacker: This one does *not* have the authorization token to do comparisons, thus security against this (weaker) attacker can properly be defined in terms of indistinguishability.

Onewayness under chosen-ciphertext attacks is defined in the usual way by giving oracle access to $D_{sk}(\cdot)$ to the attacker $\mathcal{A}$, indicated as $\mathcal{A}^{D_{sk}(\cdot)}$, and engaging in the following experiment $\mathbf{Exp}_{\text{OW-CCA1}}^{\mathcal{A}}$ with the challenger.

Setup phase: the challenger creates$(pk, sk) \leftarrow G$.

Query phase: the attacker (adaptively) chooses a number of q ciphertexts $c_i \in \mathbb{G}_c$ and retrieves $m_i \leftarrow D_{sk}(c)$ from the challenger for $i = 1,2, \dots, q$.

Challenge phase: challenger chooses a plaintext m that has not been returned in the query phase, and submits $c^* \leftarrow E_{pk}(m)$ to the adversary.

Guess phase: attacker outputs a guess m^*.

The *advantage* of $\mathcal{A}^{D_{sk}(\cdot)}$in $\mathbf{Exp}_{\text{OW-CCA1}}$is

$$\mathbf{Adv}_{\text{OW-CCA1}}^{\mathcal{A}^{D_{sk}(\cdot)}} \coloneqq \Pr\left[D_{sk}(c^*) = m^* \middle| m^* \leftarrow \mathcal{A}^{D_{sk}(\cdot)}\right].$$

We call the encryption (τ, q, ε)*-OW-CCA1-secure*, if an adversary $\mathcal{A}^{D_{sk}(\cdot)}$ of complexity $\leq \tau$ and submitting no more than q queries has an advantage $\mathbf{Adv}_{\text{OW-CCA1}}^{\mathcal{A}^{D_{sk}(\cdot)}} < \varepsilon$.

Indistinguishability under chosen-ciphertext attacks is defined by the following experiment $\mathbf{Exp}_{\text{IND-CCA1}}^{\mathcal{A}}$. As before, we assume oracle access to decryptions under sk:

Setup phase: the challenger creates $(pk, sk) \leftarrow G$.

Query phase: the attacker (adaptively) chooses a number of q ciphertexts $c_i \in \mathbb{G}_c$ and retrieves $m_i \leftarrow D_{sk}(c)$ from the challenger for $i = 1,2, \dots, q$.

Challenge phase: the attacker generates two messages $m_0, m_1 \leftarrow \mathcal{A}^{D_{sk}(\cdot)}$, where $m_0 \neq m_1$ and $|m_0| = |m_1|$. The challenger receiversm_0, m_1, chooses $b \overset{r}{\leftarrow} \{0,1\}$, and returns $c_b \leftarrow E_{pk}(m_b)$.

Guess phase: attacker outputs a guess b^*.

The *advantage* of $\mathcal{A}^{D_{sk}(\cdot)}$in $\mathbf{Exp}_{\text{IND-CCA1}}$is

$$\mathbf{Adv}_{\text{IND-CCA1}}^{\mathcal{A}^{D_{sk}(\cdot)}} := \left|\Pr[b^* = b] - \frac{1}{2}\right|.$$

We call the encryption (τ, q, ε)*-IND-CCA1-secure*, if an adversary $\mathcal{A}^{D_{sk}(\cdot)}$ of complexity $\leq \tau$ and submitting no more than q queries has an advantage $\mathbf{Adv}_{\text{IND-CCA1}}^{\mathcal{A}^{D_{sk}(\cdot)}} < \varepsilon$.

Comparisons can be done by allowing decryptions of either the plaintext or a hash-value thereof (not revealing the plaintext as such). To this end, we define a commitment-like hash-function that acts on the same plaintext group as the encryption does. For security, we require the discrete logarithm problem to be difficult on this group. Formally, let $g \in \mathbb{G}_c$ generate the group $\mathbb{G}_c$. We call $\mathbb{G}_c$ a (τ, ε) -DL group, if $\Pr\left[x = x' \middle| x \overset{r}{\leftarrow} \mathbb{G}_m, y \leftarrow g^x, x' \leftarrow \mathcal{A}(y)\right] < \varepsilon$ for all circuits $\mathcal{A}$ of complexity $\leq \tau$.

A. Asymptotic Security

For generality, we give concrete security statements here in terms of the security parameters τ, q, ε, leaving their obvious respective asymptotic formulations aside. Throughout the rest of this work, we confine ourselves to stressing that all parameters, in the asymptotic formulation, would depend on a (common) security parameter $\kappa \in \mathbb{N}$ that usually controls key-sizes, group structures or similar (consequently, it goes as a parameter into the key-generation circuit G). As an example, the asymptotic version of (τ, q, ε)-OW-CCA1 security would read as follows: for every polynomials$\tau(\kappa), q(\kappa)$ there is a negligible function [1] $\varepsilon(\kappa)$ such that the encryption is $(\tau(\kappa), q(\kappa), \varepsilon(\kappa))$ -OW-CCA1 secure. All results and definitions to follow can be restated in a similar manner.

III. The Encryption Scheme

Our encryption scheme will allow comparisons by attaching a keyed hash of the inner plaintext to the ciphertext, where the key for the hash is also encrypted[2]. Comparisons then need the permission by the originator of the ciphertext, who must provide the decryption key to disclose the hash-key. This key is obtained by an *authorization function*`Aut`. The comparison procedure `com` then simply compares the "decrypted" hashes. To distinguish the components of our HPKEET-encryption scheme (`KeyGen`, `Enc`, `Dec`, `Aut`, `Com`) from that of the underlying OW-CCA1 and IND-CCA1 secure encryption (G, E, D), we use a different notation hereafter. Moreover, we assume that the plaintext group $\mathbb{G}_m$ is such that DL-commitments to $m \in \mathbb{G}_m$ are well-defined; that is, we can compute g^m for a generator g of $\mathbb{G}_c$ and some value $m \in \mathbb{G}_m$. This is trivially satisfied for prime order groups over the integers (say, if $\mathbb{G}_c \simeq \mathbb{G}_m \simeq \mathbb{Z}_p$ for some prime p), and requires only simple additional measures in elliptic curve settings.

`KeyGen`: Create $(pk_1, sk_1) \leftarrow G$and $(pk_2, sk_2) \leftarrow G$. Put $pk := (pk_1, pk_2)$and $sk := (sk_1, sk_2)$. Choose two (distinct) generators $g, h$of $\mathbb{G}_c$. The system parameters globally known to all instances are $\mathbb{G}_m, \mathbb{G}_c, g$and h.

`Enc`: Given the message $m \in \mathbb{Z}$, the encryption is `Enc`$: \mathbb{G}_m \to \mathbb{G}_c^3$ by choosing an integer $r < |\mathbb{G}_c|$ and returning $\texttt{Enc}_{pk}(m) := (E_{pk_1}(m), g^m h^r, E_{pk_2}(r))$.

`Dec`: If the given ciphertext c cannot be parsed as an element $(c_1, c_2, c_3) \in \mathbb{G}_c^3$, return $\perp$. Otherwise, put $m' \leftarrow D_{sk_1}(c_1), r' \leftarrow D_{sk_2}(c_3)$ and verify if $c_2 = g^{m'} h^{r'}$. Output m' upon a match, and $\perp$ otherwise.

`Aut`: To authorize a third party to do comparisons, `Aut` extracts and returns the *token* $t \leftarrow sk_2$ from the secret key $sk = (sk_1, sk_2)$.

`Com`: Given a token t and two (syntactically correct) ciphertexts $(c_1, c_2, c_3), (c_1', c_2', c_3') \in \mathbb{G}_c^3$, compute $r \leftarrow D_t(c_3), r' \leftarrow D_t(c_3')$ and output the result of the comparison $c_2 \cdot h^{-r} = c_2' \cdot h^{-r'}$ in $\mathbb{G}_c$.

Notice that a trivial instantiation of the above scheme by a symmetric (e.g., AES) or deterministic (e.g. plain RSA) encryption would be insecure. Even though comparisons are easy in that case (ciphertext equality implies plaintext equality), such a scheme would not be indistinguishable, and thus fail to achieve the security that we desire against a type 2 attacker.

A. Homomorphy

Let two ciphertexts $c_i = (E_{pk}(m_i), g^{m_i} h^{r_i}, E_{pk}(r_i))$ for $i = 1,2$ be given, and consider their component-wise product in $(\mathbb{G}_c, \cdot)$, which is

$$\begin{aligned} c_1 \cdot c_2 \quad &= \quad [E_{pk}(m_1) \cdot E_{pk}(m_2), g^{m_1} h^{r_1} g^{m_2} h^{r_2}, \\ &\qquad E_{pk}(r_1) \cdot E_{pk}(r_2)] \\ &= \quad (E_{pk}(m_1 + m_2), g^{m_1 + m_2} h^{r_1 + r_2}, E_{pk}(r_1 + r_2) \end{aligned}$$

This is a valid ciphertext if and only if the underlying encryption (G, E, D) is *additively homomorphic*. Unfortunately, we cannot instantiate (G, E, D) as a Paillier-encryption, since this works over a composite modulus $n = pq$ for which $\mathbb{Z}_n$ is not cyclic (in general). An "almost" compatible IND-CCA1 secure encryption, except for its multiplicative homomorphy, is found in Damgårds version of ElGamal encryption [20]. Changing the multiplicative homomorphic property of ElGamal encryption into an additive one is easy by encrypting commitments g^m instead of m, if the plaintext space is only of "tractably small size" (e.g., polynomial size in the security parameter) to let us recover m from g^m efficiently. While this requirement is easily met in our application to Turing-machines, we stress that care has to be taken in the encoding of m in order to avoid trial opening of commitments (and thus breaking the encryption) during an invocation of `Com`, if the token (secret key to decrypt the randomizer) is available (through an invocation of `Aut`). We take a closer look at this now.

[1] Negligibility of a function f is defined in the usual way, by requiring for every $\alpha > 0$ the existence of a constant $K_\alpha > 0$ so that $f(\kappa) < \frac{1}{\kappa^\alpha}$as soon as $\kappa > K_\alpha$.

[2] This proposal is as well found in [19], where it is instantiated in an insecure manner under an adaptive chosen ciphertext attack scenario (CCA2). We consider a similar instantiation (equally well not IND-CCA2 secure), but prove it secure in the weaker model of (non-adaptive) chosen-ciphertext attacks (IND-CCA1).

B. Security Analysis

We start with a (well-known) necessary condition for security to avoid brute-force plaintext search.

1) Offline Message Recovery

Plaintext discovery by trial encryptions and checking equality with the given ciphertext is essentially unavoidable, but can be made infeasible if the plaintexts have high min-entropy: recall that a random plaintext M over a set $\mathbb{G}_m$ has min-entropy $H_\infty(M)=k$, if k is the largest number such that $\Pr[M=m]\leq 2^{-k}$ for all $m\in\mathbb{G}_m$.

Lemma 1. If an encryption function E is such that for any circuit A of complexity τ, we have $\Pr[M=m^*|m^*\leftarrow\mathcal{A}(c)]<\varepsilon$ for any given ciphertext $c=E_{pk}(M)$, then the plaintext M has min-entropy

$$H_\infty(M)>\log_2\left[1-(1-\varepsilon)^{t_{ED}+\frac{t_{comp}}{\tau}}\right], \tag{1}$$

Where$t_{ED}$is the complexity of computing an encryption, and t_{comp} measures how much circuitry is required to string-compare two ciphertexts.

*Proof.*If the lemma were wrong, then a circuit can do encryptions (of complexity t_{ED}) and comparisons (of complexity t_{comp}) to determine the correct plaintext. From the geometric distribution, it is easy to obtain the number of trials until the success probability becomes $\geq\varepsilon$. Constraining this number to be less than $\tau/(t_{ED}+t_{comp})$ (assuming the circuitry to be divided equally into blocks that do encryptions and comparisons), gives the stated min-entropy bound. □

Lemma 1 is a necessary yet insufficient condition for security. Its asymptotic counterpart (i.e., $H_\infty(M)\in\omega(\kappa)$ when κ is the security parameter) is a standard requirement for security of deterministic or searchable encryption (cf. [21]) against polynomial time-bounded attackers. We establish security of the encryption as such in the next section, and postpone a discussion on how to practically assure condition (1) until section IV.B.

2) Chosen Ciphertext Security

As the encryption comes with comparison facilities, we modify the OW-CCA1 and IND-CCA1 games appropriately, by letting the attacker submit Aut-queries besides decryption requests. To distinguish the experiments concerning HPKEET from that on the underlying cipher (G,E,D), we denote these extended versions as $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$ and $\mathbf{Exp}^{\mathcal{A}}_{\text{IND-CCAE1}}$, i.e., security under chosen ciphertexts and equality checks. The definition of $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$ is the same as that of $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCA1}}$, except for a slight modification in the query phase:

$\mathbf{Exp}_{\text{OW-CCAE1}}$ **query phase**: the attacker submits no more than q queries of the giving $m_i\leftarrow D_{sk}(c_i)$ for (adaptively) chosen ciphertextsc_i or $t\leftarrow$ Aut, for an authorization query.

Obviously, we cannot apply the same change to $\mathbf{Exp}_{\text{IND-CCAE1}}$, so we define this experiment *exactly identical* to $\mathbf{Exp}_{\text{IND-CCA1}}$.

By construction, our encryption is a humble application of E on two stochastically independent quantities $m$and r, along with a product of two commitments thereof. Hence, the reductions establish only a slight advantage over that in breaking (G,E,D). Formally, we have

Lemma 2. Let (G,E,D) be defined over an $(\tau+t_{ED}+t_{eim},\varepsilon_{DL})$-DL-group $\mathbb{G}_m$ of plaintexts, where t_{ED} is the maximum complexity of an encryption or decryption, and t_{eim} is the total complexity of one exponentiation with inversion and multiplication in $\mathbb{G}_c$. If (G,E,D) is $(\tau+(q+1)(t_{ED}+t_{eim}),q,\varepsilon)$ -OW-CCA1-secure, then the corresponding HPKEET scheme is $(\tau,q,\varepsilon+\varepsilon_{DL})$-OW-CCA1-secure.

*Proof.*Suppose the existence of an attacker $\mathcal{A}$ with advantage $\mathbf{Adv}^{\mathcal{A}^{\mathtt{Dec}_{sk}(\cdot)}}_{\text{OW-CCAE1}}>\varepsilon+\varepsilon_{DL}$ and complexity $\leq\tau$ and making q queries. We construct an attacker $\mathcal{A}'$ that wins $\mathbf{Exp}^{\mathcal{A}'}_{\text{OW-CCA1}}$as follows: given (pk,sk) from $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCA1}}$, $\mathcal{A}'$ sets $(pk_1,sk_1):=(pk,sk)$ and obtains $(pk_2,sk_2)\leftarrow G$ on its own. It then simulates $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$for $\mathcal{A}$, answering the i-th query (for $i=1,2,\ldots,q$) as follows:

- Dec-queries on an incoming HPKEET ciphertext$c_i=(c_{i,1},c_{i,2},c_{i,3})$ are forwarded as decryption challenges $c_i=c_{i,1}$ to the OW-CCA1 challenger, which returns $m'\leftarrow\mathtt{Dec}_{sk}(c_i)=D_{sk_1}(c_{i,1})$. Then, $\mathcal{A}'$ goes on by decrypting $c_{i,3}$ using its own secret key sk_2 into $r'\leftarrow\mathtt{Dec}_{sk_2}(c_{i,3})$, and returns m' if $g^{m'}h^{r'}=c_{i,2}$ in $\mathbb{G}_c$, and $\perp$ otherwise. We stress that the keypairs (pk_1,sk_1) and (pk_2,sk_2) for encrypting the payload and the randomizer are in any case chosen stochastically independent. Hence, $\mathcal{A}'$ actually acts properly if it generates (pk_2,sk_2) by itself, and receives the other pair (pk_1,sk_1) from an external source (the OW-CCA1 challenger).
- Aut-queries are answered faithfully by responding with sk_2.

In the challenge phase, the complexity of $\mathcal{A}$ is thus dominated by simulations of half of the decryption of challenges from$\mathcal{A}$, which is $\leq q\cdot(t_{ED}+t_{eim})$.

To ease notation, let us incorporate all information from the query phase of $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$ into the circuit $\mathcal{A}$, which in the guess phase of $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$ computes its output upon a given ciphertext $(c_1^*,c_2^*,c_3^*)=(E_{pk_1}(m),g^mh^r,E_{pk_2}(r))$. Observe that c_3^*, in an information-theoretic sense, does not provide any information on c_1^* , and uniquely determines g^m from c_2^* . Therefore, in any $\mathbf{Exp}^{\mathcal{A}}_{\text{OW-CCAE1}}$ execution in which at least one Aut-query has been submitted,

$$\begin{aligned}&\mathbf{Adv}^{\mathcal{A}^{\mathtt{Dec}_{sk}(\cdot)}}_{\text{OW-CCAE1}}\\&\quad=\Pr\left[\mathtt{Dec}_{sk_1}(c^*)=m^*|m^*\leftarrow\mathcal{A}(E_{pk_1}(m),g^mh^r,E_{pk_2}(r))\right]\\&\quad=\Pr\left[\mathtt{Dec}_{sk_1}(c^*)=m^*|m^*\leftarrow\hat{\mathcal{A}}\big(E_{pk_1}(m),g^m\big)\right]\end{aligned}$$

for some circuit $\hat{\mathcal{A}}$. Obviously, one could convert from the inputs $(E_{pk_1}(m), g^m)$ to $(E_{pk_1}(m), g^m h^r, E_{pk_2}(r))$ (and back) by choosing (or decrypting) the randomizer r and doing (or inverting) the remaining operations. Hence, the complexity of $\hat{\mathcal{A}}$ is bounded from above by $\tau + t_{ED} + t_{eim}$, where t_{ED} and t_{eim} are the complexities of an encryption/decryption (maximum thereof), and an exponentiation with inversion and multiplication in $\mathbb{G}_c$. The advantage of $\hat{\mathcal{A}}$ is the probability of guessing m^* correctly either from $E_{pk_1}(m)$ or g^m alone, or from both. From the union bound and by assuming that $\mathbb{G}_c$is an $(\tau + t_{ED} + t_{eim}, \varepsilon_{DL})$-DL-group, we get

$$\begin{aligned} \varepsilon + \varepsilon_{DL} &< \mathbf{Adv}^{\hat{\mathcal{A}}}_{\text{OW-CCAE1}} \\ &\leq \Pr\left[m = m^* | m^* \text{extracted from} E_{pk_1}(m)\right] \\ &\quad + \Pr[m = m^* | m^* \text{extracted from} g^m] \\ &= \mathbf{Adv}^{\mathcal{A}'}_{\text{OW-CCA1}} + \varepsilon_{DL}. \end{aligned}$$

The complexity of $\mathcal{A}'$ is thus $q \cdot (t_{ED} + t_{eim}) + \tau + t_{ED} + t_{eim} = \tau + (q+1)(t_{ED} + t_{eim})$.

In the challenge phase of $\mathbf{Exp}^{\mathcal{A}'}_{\text{OW-CCA1}}$, upon incoming of $c^* = E_{pk}(m)$, $\mathcal{A}'$ can therefore run $\hat{\mathcal{A}}$ in place of $\mathcal{A}$, to discover m from the OW-CCA1 challenge $c^* = E_{pk}(m)$, with an advantage $\mathbf{Adv}^{\mathcal{A}'}_{\text{OW-CCA1}} > \varepsilon$ contradicting the security of (G, E, D). □

Likewise, we establish IND-CCA1-security of HPKEET by virtue of the following well-known concrete result on how indistinguishability implies semantic security.

Lemma 3. If (G, E, D) has (τ, ε) -indistinguishable encryptions, and E_{pk} has complexity $\leq t_{ED}$, then (G, E, D) is $(\tau - \ell_f, t_{ED}, \varepsilon)$ -semantically secure where (t_1, t_2, δ) - semantic security is defined as follows: for every distribution X over messages, every functions $I: \mathbb{G}_m \to \{0,1\}^*, F: \mathbb{G}_m \to \{0,1\}^{\ell_f}$ (of arbitrary complexity) and every circuit A of complexity $\leq t_1$, there is another circuit A^* with complexity $\leq t_1 + t_2$ so that

$$\left| \Pr\left[A\left(E_{pk}(m), I(m)\right) = F(m)\right] - \Pr\left[A^*\left(I(m)\right) = F(m)\right] \right| < \varepsilon.$$

Lemma 4. Let (G, E, D) be defined over a group $\mathbb{G}_m$ of plaintexts, where t_{ED} bounds the complexity of an encryption or decryption, and t_{eim} is the complexity of one exponentiation with inversion and multiplication in $\mathbb{G}_c$. If (G, E, D) is $(\tau + q \cdot (t_{ED} + t_{eim}) + t_{ED} - 1, q, \varepsilon)$-IND-CCA1-secure, then the corresponding HPKEET scheme is $(\tau, q, 2\varepsilon)$-IND-CCA1-secure.

*Proof.*Besides a few modifications that we describe now, the line of arguments is completely analogous as in the proof of Lemma 2, except for the important difference that the adversary is not allowed to issue `Aut`-queries in $\mathbf{Exp}^{\mathcal{A}}_{\text{IND-CCAE1}}$. Assume an attacker $\mathcal{A}$ with 2ε-advantage in $\mathbf{Adv}^{\mathcal{A}}_{\text{IND-CCAE1}}$. The complexity of $\mathcal{A}$ during the challenge phase is (as before) $q \cdot (t_{ED} + t_{eim})$. Upon the incoming challenge $c_b = E_{pk}(m_b)$ in $\mathbf{Exp}^{\mathcal{A}}_{\text{IND-CCA1}}$, $\mathcal{A}$ embeds it in a HPKEET ciphertext $c_b^* = (c_b, r, r')$, for $r, r' \overset{r}{\leftarrow} \mathbb{G}_m$. Observe that a unique value$r'' \in \mathbb{G}_m$exists for which $g^{m_b} h^{r''} = r$. For c_b^* to be a valid HPKEET ciphertext, r' should equal $E_{pk_2}(r')$, which is most likely not the case. We can fix this by exploiting the indistinguishability of encryptions under E as follows: as E is $(\tau + q \cdot (t_{ED} + t_{eim}), q, \varepsilon)$ -IND-CCA1-secure,Lemma 3lets us replace $\mathcal{A}$ by another circuit $\mathcal{A}^*$ that has complexity $\tau + q \cdot (t_{ED} + t_{eim}) + t_{ED} - 1$ and delivers the decision $f(c_b) = b' \in \{0,1\}^{\ell_f = 1}$ so that

$$\left| \Pr\left[\mathcal{A}\left(E_{pk}(m_b), g^{m_b} h^{r'}, E_{pk}(r')\right) = b\right] - \Pr\left[\mathcal{A}^*\left(E_{pk}(m_b), g^{m_b} h^{r'}\right) = b\right] \right| \leq \varepsilon \qquad (2)$$

Observe that $g^{m_b} h^{r'} = g^{m_0} h^{r''}$ for some random r'', which means that this second parameter to $\mathcal{A}^*$ – in an information-theoretic sense – does not provide any additional information on b. So, there is another circuit $\mathcal{A}^{**}$, no more complex than $\mathcal{A}^*$, such that$\Pr\left[\mathcal{A}^*(E_{pk}(m_b), g^{m_b} h^{r'}) = b\right] = \Pr\left[\mathcal{A}^{**}(E_{pk}(m_b)) = b\right]$.Now, we can construct an attacker $\mathcal{A}'$that wins the IND-CCA1 game as follows: $\mathcal{A}'$ invokes $\mathcal{A}^{**}$ on input of the IND-CCA1 challenge c_b, and output whatever $\mathcal{A}^{**}$ guesses. Inequality (2) tells that the result of $\mathcal{A}^{**}$ differs from that of $\mathcal{A}$ (on a syntactically correct input) with a probability of less than ε. Moreover, $\mathcal{A}$ would by assumption guess correctly with an advantage of at least 2ε. So by the second triangle inequality, and with the abbreviation $I = (E_{pk}(m_b), g^{m_b} h^{r'}, E_{pk}(r'))$, we get

$$\begin{aligned} &\left| \Pr[\mathcal{A}^{**}(c_b) = b] - \frac{1}{2} \right| \\ &\geq \big| |\Pr[\mathcal{A}(I) = b] - 1/2| \\ &\qquad - |\Pr[\mathcal{A}(I) = b] - \Pr[\mathcal{A}^{**}(c_b) = b]| \big| \\ &\geq 2\varepsilon - \varepsilon = \varepsilon, \end{aligned}$$

where $\mathcal{A}^{**}$ has complexity $\tau + q \cdot (t_{ED} + t_{eim}) + t_{ED} - 1$,taking at most q queries, which contradicts the assumed IND-CCA1-security of E.

IV. Blind Turing-Machines

Informally, a blind Turing-machine (BTM) is a normal TM, having its tape alphabet and transition function encrypted under a homomorphic public-key encryption with plaintext equality checking. The transition between states is made by homomorphic manipulations, and the choice of the current transition is made upon plaintext comparisons. We describe the construction over a sequence of subsections to follow.

A. Definitions

We start with a standard two-way infinite tape Turing-machine $M = (Z, \Sigma, \delta, s)$, working over a tape alphabet Σ with Z being its state-space (including the halting state), and $s \in Z$ being the initial state. The mapping δ describes the state transitions in terms of transforming configurations of the TM to one another. A *configuration* is a tuple $(q, w_L, \sigma, w_R) \in Z \times \Sigma^* \times \Sigma \times \Sigma^* =: Conf_M$, describing the fact that the machine is currently in state $q \in Z$, with symbol $\sigma \in \Sigma$ under its head, and with $w_L, w_R \in \Sigma^*$ being the words to the left- and right of the head. The transition function $\delta: Conf_M \to Conf_M$ is a finite set of transformations $\delta(q, w_L, \sigma, w_R) = (p, w_L{}', \sigma', w_R{}')$, i.e., M moves to state p and modifies the tape content toward $w_L{}', w_R{}'$ and σ'. Without loss of generality, we restrict our attention to *deterministic* TM here, as there is no conceptual difference in the nondeterministic case, except that we work on a transition *relation* rather than a function (as will become clear below, the necessary changes to define blind nondeterministic TM are all obvious).

We abbreviate configurations as χ and write $\chi_1 \vdash \chi_2$ as a shorthand of $\delta(\chi_1) = \chi_2$. A *computation* of M on an initial configuration χ_0 is a finite sequence of configurations $\chi_0 \vdash \chi_1 \vdash \chi_2 \vdash \cdots \vdash \chi_\tau = (h, \ldots)$ that ends in a halting state $h \in Z$ and output configuration χ_τ. The number τ of steps is called the machine's *time-complexity*, which normally depends on the size of the input (polynomial mostly, if we are after efficient algorithms).

Notice that for our purposes, we do not distinguish moving steps (where only the head is relocated) from substitution steps (in which the current symbol on the tape is replaced by something else). Also note that it is difficult to hide the head movements from the execution environment of the TM (e.g., a universal TM), yet it is necessary to "decorrelate" the head movement pattern from the tape content to achieve confidentiality of the overall computation. Otherwise, the movement of the TM discloses the tape content length and perhaps even reveals the current action that is been carried out (by virtue of some characteristic moving sequences, as would perhaps be the case for pen-and-paper multiplication or division by repeated subtraction which reveals the quotient via counting the number of subtractions, regardless of whether or not they are encrypted).

In section V, we will look at necessary precautions to prevent leakage of information from the Turing-machines head movements alone (quasi as a side-channel to the data as such). Note that similar concerns may apply to garbled circuits as well, as the way in which circuit gates (whether or not they are garbled) are interconnected may already leak partial information about the circuit's potential functionality. Still, we emphasize that our main goal in this work is to protect the data being processed. Hiding the algorithm itself from the execution environment is subject of future considerations and outside the scope of this current work.

B. Encoding of States and Tape-Symbols

Take a conventional TM $M = (Z, \Sigma, \delta, s)$. Let HPKEET operate on the plaintext space $\mathbb{G}_m$ and ciphertext space $\mathbb{G}_c$, and fix an (invertible) encoding $C: Z \times \Sigma \to \mathbb{G}_m$, so that we can encrypt both, the state *and* current tape symbol.

Computations are usually done over relatively small alphabet, say bits ($\Sigma = \{0,1\}$) or radix-10 numbers ($\Sigma = \{0, \ldots, 9\}$). Moreover, the number of states can be expected to be feasibly small as well (otherwise, the representation of M could not be handled by the universal TM in feasible time). Hence, if $|\mathbb{G}_m| = 2^{O(\kappa)}$ for some security parameter κ, then high min-entropy in the sense of (1) can be assured by sufficiently large κ and by assigning random and unique representatives from $\mathbb{G}_m$ to each element of $\Sigma \times Z$, in order to thwart trial decryptions succeeding in polynomial time.

C. Construction

Our blind TM works over ciphertexts only, and does its transitions using a proper „encryption" of the original state transition function in M. To this end, we extend M toward $\widehat{M} = (\mathbb{G}_c, \mathbb{G}_c, \hat{\delta}, \hat{s})$ and define the blind TM as the pair $BTM = (\widehat{M}$,HPKEET$)$. Here, the ∧-accent is used to denote the „encrypted"counterparts of the respective elements in M's description. We stress that the description is technically complete but to this extent insecure, as the head movement pattern may leak information about the tape content. For the sake of a complete description at this point, however, we postpone the necessary details on security to section V. A blind TM works exactly as a normal TM, but employs HPKEET to do transitions over encrypted configurations as follows:

1) *Encrypted configurations: given a configuration $\chi = (q, w_L, \sigma, w_R)$ of M, the respective encrypted configuration $\hat{\chi}$ (under the public key pk, which we omit in the following to simplify our notation), is defined as*

$$\chi := (\mathtt{Enc}_{\mathrm{pk}}(\mathrm{p}), \mathrm{w_L}', \mathtt{Enc}_{\mathrm{pk}}(\sigma), \mathrm{w_R}'),$$

Where$w_L{}', w_R{}'$ are the encryptions of the tape content under $\mathtt{Enc}_{pk}$ in electronic codebook mode.

2) *Transition functions: for each pair of consecutive configurations $\chi = (q, w_L, \sigma, w_R') \vdash \chi' = (p, w_L{}', \sigma, w_R{}')$ of M, the transition function $\hat{\delta}$ for the blind TM is created from δ as*

$$\hat{\delta} = \{\hat{\chi} \mapsto \left(\mathtt{Enc}_{pk}(\Delta_p), w_L', \mathtt{Enc}_{pk}(\Delta_\sigma), w_R'\right) |\, \delta(\chi) = \chi'\}$$

Where$\Delta_p = p - q$and $\Delta_\sigma = \sigma' - \sigma$, both computed in $(\mathbb{G}_m, +)$. So, unlike the transition of the TM, a blind TM encrypts only the "difference" between the current and next configuration, in order to enforce re-randomization via homomorphic manipulations on the ciphertexts. Hence, actually doing a transition is now a two-step process:

a) We invoke $\mathtt{Com}$ with the token $t \leftarrow \mathtt{Aut}(sk)$ on the current (encrypted) configuration $\hat{\chi}$ of the BTM to match the states and symbols, and retrieve $\hat{\delta}$.

b) We create the new configuration χ' from the current one $\chi = (\mathit{Enc}_{pk}(q), w_L, \mathit{Enc}_{pk}(\sigma), w_R)$ by computing in $(\mathbb{G}_c, \cdot)$,

$$\begin{aligned}\hat{\chi}' \; = \; (\mathtt{Enc}_{pk}(q) \cdot \mathtt{Enc}_{pk}(\mathit{\Delta}_p), w_L', \mathtt{Enc}_{pk}(\sigma) \\ \cdot \mathtt{Enc}_{pk}(\mathit{\Delta}_\sigma), w_R{}')\end{aligned}$$

so as to resemble M's original move via the homomorphic properties of $\mathtt{Enc}$, which is easily verified to be

$$\begin{aligned}\hat{\chi}' \;&=\; (\mathtt{Enc}_{pk}(q + \Delta_p), w_L', \mathtt{Enc}_{pk}(\sigma + \Delta_\sigma), w_R') \\ &=\; \left(\mathtt{Enc}_{pk}(p), w_L', \mathtt{Enc}_{pk}(\sigma'), w_R'\right) = \widehat{\chi'}\end{aligned}$$

Doing tape manipulations by other means than homomorphic transformations for the sake of stronger IND-CCA2 security is potentially insecure, as we will discuss in a little more detail in section V.A.

Based on this construction, it is a trivial matter to decrypt and recover the tape content by virtue of $\mathtt{Dec}_{sk}$, and we omit the details here. However, note that like in a setting of fully homomorphic encryption, the state transitions require the token as an "evaluation key".

V. Security Of Blind Computations

By construction, an execution of a BTM produces a sequence of encrypted configurations, enjoying a one-to-one correspondence to the respective sequence of configurations arising from an execution of M . However, to retain indistinguishability in experiment $\mathbf{Exp}_{\text{IND-CCAE1}}^{\mathcal{A}^{\mathtt{Dec}_{sk}(\cdot)}}$, we ought to equalize the length of computations on inputs of equal length, *and* make the head movements indistinguishable over different inputs. To this end, we must transform the given Turing-machine accordingly before turning it into a blind TM.

To equalize the length of computations, we restrict the time-complexity bound of M to time-constructible functions. We say that a function $T: \mathbb{N} \to \mathbb{N}$ is *time-constructible*, if there is a Turing-machine M_T, which for every input w of length $|w|$ takes exactly $T(|w|)$ steps for its computation on w (not necessarily saying that it computes anything useful). For example, every polynomial function is time-constructible (but also exponential functions, sums, products and compositions of time-constructible functions retain this property).

Furthermore, we must logically decouple the tape content from the head movement pattern to avoid leakage of information via tracking what the head of the blind TM does. Turing-machines whose head movements are a function of the time only (hence independent of the tape content) are called *oblivious* TM. Besides theoretical interest in these for the sake of constructing circuits, the following well-known theorem will help to establish security of blind computations:

Theorem 1. (Pippenger and Fischer [22]) Any Turing-machine that runs in time τ can be simulated by an oblivious Turing-machine in time $O(\tau \log \tau)$.

A naive yet constructive approach to create an oblivious TM from a given one is to mark where the head of the tape is and then scan the tape to locate the head marker in each step. This yields a suboptimal time bound of $O(\tau^2)$ for a running time of τ on the original TM, and Theorem 1 gives in fact the optimal bound.

So, given a Turing-machine M whose time-complexity is a time-constructible function, we first transform M into an oblivious Turing-machine M', running in time τ_M, which is again time-constructible by some Turing-machine M_T. Then, we let our blind TM run M_T in parallel to M' on a second tape, so as to equalize the length of its computation, while running the oblivious TM M' to do the actual computation with head movements that are independent of the data. This proves the following (intermediate nevertheless important) statement:

Theorem 2. Let M be a Turing-machine, whose time complexity T is a time-constructible function. Then there exists a functionally equivalent Turing-machine M' with the following property: given any two input words $w_1 \neq w_2$ of the same length $|w_1| = |w_2|$, a computation of M' takes identical head movements on both, w_1 and w_2.

We can now turn to the task of lifting security assurances that hold for HPKEET towards security for an entire computation on a blind oblivious Turing-machine. Notice that so far, we considered security only for *one* message to be deciphered (as in $\mathbf{Exp}_{\text{OW-CCA(E)1}}^{\mathcal{A}}$) or recognized (from two given ones, as in $\mathbf{Exp}_{\text{IND-CCA(E)1}}^{\mathcal{A}}$). Security of a computation of a blind TM, however, requires a slight change to the experiments, in the sense that the challenge-phase in both games now itself is repeated a number of times that equals the time-complexity[3] τ_M of the underlying TM M. Omitting the obvious details on the changes to the experiments here for brevity, let us directly turn to the respective security conclusions about HPKEET under τ_M many encryptions (each one of which arises along the emulation of M by a blind TM).

Lemma 5. If HPKEET is (τ, q, ε)-IND-CCA1 secure for a single encryption, then it is $(\tau - n \cdot t_{\mathtt{Enc}}, q, n \cdot \varepsilon)$-IND-CCA1-secure for n encryptions, when $t_{\mathtt{Enc}}$ bounds the complexity of an encryption using HPKEET. Given the additional hypothesis that all random encrypted plaintexts have high min-entropy in the sense of Lemma 1, then the system is also $(\tau - n \cdot t_{\mathtt{Enc}}, q, n \cdot \varepsilon)$-OW-CCA1-secure for n encryptions.

*Proof (sketch).*Indistinguishability is shown by assuming the existence of a pair of (with probability $\geq n \cdot \varepsilon$) distinguishable n-tuples, and constructing hybrids to infer distinguishability in the single-message case with probability ε. To this end, the distinguisher must emulate encryptions of no more than n (fixed) input messages, which enlarges the circuits (and yields the modified complexity-bounds).

Onewayness is analyzed in a similar fashion, but is slightly simpler in the details: if a circuit $\mathcal{A}$ exists that upon input of n ciphertexts ouputs one of the underlying plaintexts with probability $\geq n\varepsilon$, then a new circuit $\mathcal{A}'$ can be constructed to correctly answer a single challenge in $\mathbf{Exp}_{\text{OW-CCA(E)1}}^{\mathcal{A}'}$ as follows: $\mathcal{A}'$ randomly constructs $n-1$ challenges $\tilde{c}_1^*, \dots, \tilde{c}_{n-1}^*$ (adding the complexity $\leq n \cdot t_{\mathtt{Enc}}$), and invokes $\mathcal{A}$ on these challenges along with the given challenge $\tilde{c}_n = c^*$. With probability $\geq n \cdot \varepsilon$, $\mathcal{A}$ outputs one plaintext $\widetilde{m}_i$ from the n ciphertexts. The chances for this to be m^* are $\frac{1}{n} n\varepsilon = \varepsilon$, contradicting the presumed OW-CCA1-security of HPKEET.

We stress that onewayness is analyzed under the assumption that plaintext comparisons are possible. Therefore, we must assume high min-entropy of plaintexts, but cannot – and in fact do not – rest on a secret encoding (as described in section IV.B, as this would prevent the constructed machines from emulating proper encryptions (for the same reason, security of multiple encryptions fails in the secret key paradigm). For one-wayness to be assured, however (and fortunately), we do not need the encoding function in the technical arguments, since it is easy to generate random plaintexts whose min-entropy is high, even if these do not lie in the image of the encoding function that the honest party

[3] Note that we do not need the space-complexity here, as we only need to count (bound) the number of modifications on the tape, which is bounded by the number of transitions, which is the time-complexity.

potentially uses. Under the additional min-entropy assumption, trial decryptions under a complexity bound τ or better ($\leq \tau - n \cdot t_{\texttt{Enc}}$) are ruled out. □

Since a BTM basically produces a sequence of ciphertexts rather than a single one, it is a simple matter to instantiate the concrete security parameters of HPKEET based on Lemma 5. Theorem 3 assumes an oblivious Turing-machine to be available, which is assured by theorem Theorem 2.

Theorem 3. Let M be an oblivious deterministic Turing-machine with time-complexity τ_M. If HPKEET is (τ, q, ε)-OW-CCA1/IND-CCA1 secure, then the execution of the BTM constructed from M is $(\tau - \tau_M \cdot t_{\texttt{Enc}}, q, \tau_M \cdot \varepsilon)$-OW-CCA1 secure. Furthermore, if τ_M is time-constructible by another Turing-machine M_τ, and if the oblivious BTM emulates (on two parallel tracks on its tape) the executions of both, M and M_τ, stopping not before both executions terminate, then the execution is also $(\tau - \tau_M \cdot t_{\texttt{Enc}}, q, \tau_M \cdot \varepsilon)$-IND-CCA1 secure.

Note that the apparently awkward mix of time-complexities and circuit complexities that appears in the above statement is actually meaningful, as the time-complexity merely determines how many ciphertexts an execution of an algorithm will provide to the cryptanalytic circuit (being an adversary of type 1). Hence, the circuit complexity is somewhat proportional to the time-complexity.

A. (In)security of Non-Homomorphic Transitions

Encryptions with equality checks have been designed earlier [19] under the stronger notion of security against adaptive chosen ciphertexts (OW/IND-CCA2), which makes the encryption necessarily no longer homomorphic. Doing a transition by a humble replacement of the current tape ciphertext (encrypted symbol) by another is possible, yet removes the indistinguishability property of computations, because an (encrypted) symbol will always and necessarily be replaced by the *same* symbol, even if the computation itself is different. As a consequence, distinct plaintexts $m_1 \neq m_2$, even if they are equally long, can be distinguished by an external instance upon different sequences of configurations. This can be done without the `Aut`-token, so that the computation would be insecure in our modified IND-CCA1 game (where the challenge phase includes multiple ciphertexts), and hence be insecure in an adaptive chosen-ciphertext scenario too.

VI. Putting A Blind TM to Work

With the ground prepared in previous sections, we now give the complete picture of how a blind TM is created and envisioned to work in a potentially hostile environment. Let Alice be the honest party who wishes to have her data processed externally by a service provider (SP), having a public key pk_{SP}. Alice has her own secret/public key pair (pk_A, sk_A). For the sake of practicality, let us assume that Alice uses Damgårds Version of ElGamal encryption for (G, E, D), which is multiplicatively homomorphic and known to be CCA1-secure [20]. To change the multiplicative homomorphic property into an additive one, Alice encrypts commitments g^m instead of m, so that the HPKEET ciphertexts now take the form $(E_{pk_1}(g^m), g^m h^r, E_{pk_2}(r))$. Assuming that the tape alphabet and number of states of her TM is feasibly small, recovering m from g^m is doable via lookup-tables. This adds an additional commit/decommit stage– shown dashed – in Fig. 1, where the overall process is sketched, including locations of type 1 and type 2 adversaries.

To have the SP process her data using a Turing-machine M, while not learning anything about it, Alice performs the steps below.

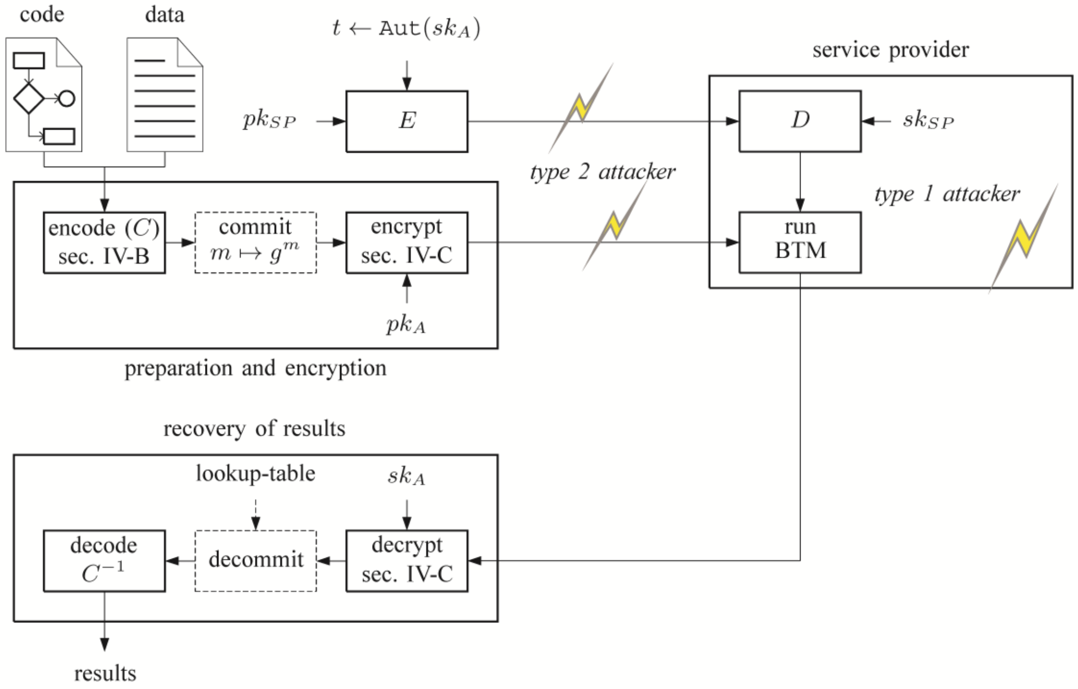


Fig. 1. Usage scheme of a blind Turing machine

1) She constructs an oblivious Turing-machine M' that simulates M and on a second track/tape (obliviously) runs the machine M_T that takes exactly $\tau_M(|w|)$ steps to terminate for an input w. This is to equalize the length of computations and head movements, regardless of the actual input. Call the resulting Turing-machine M (again, for simplicity).

2) She constructs a blind Turing-machine $\widehat{M}$ (code) from M as described in section IV. In doing so, she prepares the tape content (data) in a three steps:

a) Encode each tape symbol and state by the function C (to assure high min-entropy).

b) Compute a commitment to each encoded state and symbol (to make the multiplicatively homomorphic encryption additively homomorphic; this step can be omitted if (G, E, D) is additively homomorphic already, hence is shown dashed in Fig. 1).

c) Encrypt the commitment under the public key pk_A.

3) She then sends all information to the service provider, potentially under the eyes of a type 2 attacker (cf. section II), against which Theorem 3 assures IND-CCA1 security.

4) She submits the authorization token $t \leftarrow \mathtt{Aut}(sk_A)$ in encrypted form (under the public key pk_{SP}) to the service provider. Observe that the encrypted authorization token plays the role of something like a "license" to execute the given Turing-program, which would otherwise not be possible.

The service provider executes the (encrypted) code, i.e., runs the blind Turing-machine on the encrypted tape content, and returns the encrypted tape content. While doing so, the SP may attempt to learn information from the execution of the BTM or the intermediate tape contents, in which case the SP becomes a type 1 attacker (cf. section II), against which Theorem 3assures OW-CCA1 security.

The decryption of the ouput tape content is done as the encryption, only in reverse order, and by virtue of lookup tables to „decommit" the decrypted commitments g^m, if there has been a commitment stage during the data preparation. The results are finally available after decoding (function C^{-1}).

TABLE I. COMPLEXITIES (IMPLEMENTATION-RELATED)

Object/Action		Complexity
HPKEET	KeyGen	$2 \times G$
	$\mathtt{Enc}(m)$	$2 \times E + 2\mathtt{e} + \mathtt{m}$
	$g^m \leftarrow \mathtt{Dec}$	$2 \times D + \mathtt{e} + \mathtt{m}$
	Aut	$O(1)$
	Com	$2(D + \mathtt{e} + \mathtt{i} + \mathtt{m}$
blind TM	transition selection	$D + O(1)$
	Tape manipulation	$3\mathtt{m}$

It is as well imaginable to let the program come from a different instance (entity in the system) as the data, given that both instances have agreed on a common encoding. This scenario would be, for example, useful when a software is provided by some vendor V, and shall be executed on data that the customer A owns, while protecting the intellectual property of the software vendor. If the execution of the software shall be outsourced to an SP, then V and A both submit their authorization tokens to the SP, while A and V agree on some common encoding to have the data compatible with the code. The SP then acts as usual to compute the results in privacy. The customer A can in that case only receive and decrypt the results, while being itself unable to execute the program as Alacks V's authorization (token).

Another variation could be A encrypting the data under someone else's public key, to make the results available to another (third or fourth) party, which sees neither the input data nor the code.

VII. PERFORMANCE AND PRACTICAL ASPECTS

Assuming that the universal blind TM can select the proper transition based on the comparison facility of HPKEET, there would be no change in the asymptotic complexity of any function, whether it is computed on a conventional or blind Turing-Machine. More concretely, however, if a function $f: \{0,1\}^* \to \{0,1\}^*$ on a value $x \in \{0,1\}^n$ can be computed in time $T(n)$, then a blind TM can compute the same result in time $\leq \tau \cdot T(n)$, where τ is a constant time bound needed to manipulate a state (via homomorphic operations on the ciphertext). A similar argument can be made for the change in the space-complexity, since tape symbols are now encoded as group elements, thus multiplying the required space as well by a concrete constant factor.

Practically, a blind TM will need more time to complete its computation than a conventional TM since it has to find the proper transition based on invocations of Com. However, this can easily be accelerated if the selection is done by a hash-table taking the commitment g^m of the current tape symbol (being a HPKEET ciphertext $(E_{pk_1}(m), g^m h^r, E_{pk_2}(r))$) as the key for hashing. The transition can then be obtained from the hash-table in roughly $O(1)$ steps, as the commitments can reasonably well be assumed as being uniformly distributed (hence ideal for hashing).

Table I shows an overview of the actions involved when computing on ciphertexts, including actions that refer to HPKEET alone, taking into account that tape symbol commitments are encrypted, decrypted and compared (with the obvious changes to the formal descriptions given in section III). Here, the symbols e, i and m stand for *exponentiation*, *inversion* and *multiplication* inside the group $\mathbb{G}_c$. The notation "$k \times X$" where $X \in \{G, E, D\}$ refers to k executions of the respective algorithm implementation of the encryption (G, E, D) underlying HPKEET.

A (not very much optimized) Java implementation of our HPKEETcryptosystem based on Damgård's version of ElGamal encryption brought up some runtime estimates on a 3.6 GHz computer with 8 GB RAM and 64 Bit Windows 7, as shown in Table II. The numbers are based on an average of 100 invocations of Enc, Dec and Com for key lengths of $\kappa \in \{256, 512, 2048\}$ bit (according to current recommendations of the NIST and other bodies). The value for a transition selection and tape manipulations give a rough estimate on how much slower a blind TM will run compared to a conventional TM (i.e., the factor τ from the first paragraph).

VIII. Outlook And Open Problems

A practical topic of future work is the implementation of the concept within a practical computer architecture including assembler code and hardware. Challenges in such a practical implementation may concern the realization of other arithmetic operations such as integer divisions with remainder or logical manipulations. Results on this will be reported in companion and subsequent work.

TABLE II. Benchmark Results (For (G, E, D) Being Damgård-Elgamal Encryption)

Key size		256 bit	512 bit	2048 bit
Running time [ms]	Enc	0.93	4.07	174.52
	Dec	0.32	2.35	108.84
	Com	1.08	3.27	130.11
	BTM transition selection	0.16	1.25	43.64
	BTM tape manipulation	< 0.05	< 0.05	0.16

The central contribution here is the insight that (only) additively homomorphic encryption can be used to construct Turing-machines that work on encrypted information only, by virtue of public-key encryption with equality check. Hence, this work adds a fourth alternative to existing approaches to secure function evaluation, besides fully homomorphic encryption, garbled circuits or secure multiparty computation. Unfortunately, the necessary ingredient of additively homomorphic encryption that is secure against chosen ciphertext attacks is surprisingly rare, while non-homomorphic encryptions under stronger security notions are better known. Taking a closer look at why we require homomorphy to do state transitions reveals that the weaker requirement of re-randomization of ciphertexts is actually sufficient to invalidate the arguments of section V.A. An interesting open problem is thus finding encryptions that allow re-randomization of ciphertexts but are still CCA2-secure (if such encryptions exist).

References

[1] R. Gennaro, M. O. Rabin, and T. Rabin, "Simplified VSS and fast-track multiparty computations with applications to threshold cryptography," in Proceedings of the 17th annual ACM symposium on Principles of distributed computing. New York, NY, USA: ACM, 1998, pp. 101–111.

[2] R. Gennaro, C. Gentry, and B. Parno, "Non-interactive verifiable computing: outsourcing computation to untrusted workers," in: CRYPTO'10. Springer, 2010, pp. 465–482.

[3] M. Hirt, "Multi-party computation: Efficient protocols, general adversaries, and voting," Ph.D. dissertation, ETH Zurich, 2001.

[4] A. C.-C. Yao, "How to Generate and Exchange Secrets (Extended Abstract)," in FOCS. IEEE, 1986, pp. 162–167.

[5] V. Kolesnikov and T. Schneider, "Improved garbled circuit: Free XOR gates and applications," in Automata, Languages and Programming, ser. Lecture Notes in Computer Science, Springer, 2008, vol. 5126, pp. 486–498.

[6] T. Schneider and M. Zohner, "Gmw vs. yao? efficient secure two-party computation with low depth circuits," in Financial Cryptography and Data Security, ser. Lecture Notes in Computer Science, A.-R. Sadeghi, Ed. Springer, 2013, vol. 7859, pp. 275–292.

[7] W. Melicher, S. Zahur, and D. Evans, "An intermediate language for garbled circuits," in IEEE Symposium on Security and Privacy, 2012.

[8] Y. Huang, D. Evans, J. Katz, and L. Malka, "Faster Secure Two-Party Computation Using Garbled Circuits," in 20th USENIX Security Symposium. USENIX Association, 2011.

[9] Y. Lindell, B. Pinkas, and N. P. Smart, "Implementing two-party computation efficiently with security against malicious adversaries," in Proceedings of the 6th international conference on Security and Cryptography for Networks, ser. SCN '08. Springer, 2008, pp. 2–20.

[10] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proceedings of the 41st annual ACM symposium on Theory of computing, ser. STOC '09. New York, NY, USA: ACM, 2009, pp. 169–178.

[11] M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully Homomorphic Encryption Over the Integers," in EUROCRYPT, ser. LNCS, vol. 6110. Springer, 2010, pp. 24–43.

[12] D. Stehlé and R. Steinfeld, "Faster Fully Homomorphic Encryption," in ASIACRYPT, ser. LNCS, vol. 6477. Springer, 2010, pp. 377–394.

[13] N. P. Smart and F. Vercauteren, "Fully Homomorphic Encryption with Relatively Small Key and Ciphertext Sizes," in PKC, ser. LNCS, vol. 6056. Springer, 2010, pp. 420–443.

[14] Z. Brakerski and V. Vaikuntanathan, "Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages," in CRYPTO, ser. LNCS, vol. 6841, 2011, pp. 505–524.

[15] Z. Brakerski and V. Vaikuntanathan, "Efficient Fully Homomorphic Encryption from (Standard) LWE," in FOCS 2011. IEEE, 2011, pp. 97–106.

[16] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(Leveled) Fully Homomorphic Encryption without Bootstrapping," in ITCS. ACM, 2012, pp. 309–325.

[17] C. Gentry, S. Halevi, and N. P. Smart, "Fully Homomorphic Encryption with Polylog Overhead," in EUROCRYPT, ser. LNCS, vol. 7237. Springer, 2012, pp. 465–482.

[18] Z. Brakerski, "Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP," in CRYPTO, ser. LNCS, vol. 7417. Springer, 2012, pp. 868–886.

[19] Q. Tang, "Public key encryption supporting plaintext equality test and user-specified authorization," Security and Communication Networks, vol. 5, no. 12, pp. 1351–1362, 2012.

[20] H. Lipmaa, "On the CCA1-Security of Elgamal and Damgård's Elgamal," in ISC, ser. LNCS, vol. 6584. Springer, 2010, pp. 18–35.

[21] M. Bellare, A. Boldyreva, and A. O'Neill, "Deterministic and efficiently searchable encryption," in Advances in Cryptology CRYPTO, ser. LNCS, vol. 4622. Springer, 2007, pp. 535–552.

[22] N. Pippenger and M. J. Fischer, "Relations among complexity measures," Journal of the ACM, vol. 26, no. 2, pp. 361–381, Apr. 1979.

An Exploratory study of proposed factorsto Adopt e-government Services

Saudi Arabia as a case study

Sulaiman A. Alateyah
Electronics and Computer Science
University of Southampton
Southampton, UK

Richard M Crowder
Electronics and Computer Science
University of Southampton
Southampton, UK

Gary B Wills
Electronics and Computer Science
University of Southampton
Southampton, UK

***Abstract*—this paper discusses e-government, in particular the challenges that face adoption of e-government in Saudi Arabia. In this research we define e-government as a matrix of stakeholders: governments to governments, governments to business and governments to citizens, using information and communications technology to deliver and consume services. e-government services still face many challenges in their implementation and general adoption in many countries including Saudi Arabia. In addition, the background and the discussion identify the influential factors that affect the citizens' intention to adopt e-government services in Saudi Arabia. Consequently, these factors have been defined and categorized followed by an exploratory study to examine the importance of these factors. Therefore, this research has identified factors that determine if the citizen will adopt e-government services and thereby aiding governments in accessing what is required to increase adoption.**

Keywords—adoption; citizens' intention;e-government;e-government models; influential factors; Saudi Arabia

I. INTRODUCTION

As information and communication technologies (ICT) rapidly develop, coupled with considerable improvements in digital connectivity, government departments are reassessing the way they work and interact both with internal departments and external organisations [1]. This technology has encouraged the government's organisations and affiliations to reconsider their internal and external relations and transactions.

Therefore, in order to succeed and build for the future, the administrative processes of government are being transferred to electronic systems. Governments worldwide are considering establishing an electronic approach (e-government) to government organisations and agencies in order to provide and facilitate many services to people anywhere and at any time, and to replace traditional routine procedures. Within the paradigm of human and social development, the United Nations [2] has a conceptual framework for e-government programmes. In the United Nations context, e-government is achieved when a state uses ICT to improve the availability of information to its citizens. In order to achieve this, the capacity and readiness of the public sector have to increase in the areas of a country's technological and telecommunications infrastructure and the level of its human resources development [3].

A. The Saudi'Arabiane-government Program

The Saudi government launched the YESSER Program, the country's first national e- government strategy, in 2005 [4]. The aim of this initiative is to create user-centric electronic initiatives that focus on improving government services to the public sector. In addition, the vision of the Kingdom of Saudi Arabia is to adopt and activate communication and IT systems which led to realize an IT community and a digital economy [5]. The government of Saudi Arabia has taken steps to develop business process and disseminate the concept of e-services in various government agencies in order to realize their vision [5]. Furthermore, it has been announced by Lean, et al. [5] that to achieve the objectives, a set of promising ambitious plans and strategies have been adopted by the Saudi Arabian government. The plans for developing and implementing the e-government program has been set and have two actions, which is the first plan has took a place from 2006 to 2010, and the second is progressing from 2012 to 2016 [5].

B. Adopting new technology

The success of the implementation of the e-government is dependent not only on the government support, but also on willingness to accept and adopt e-government services by the citizens [6]. Although the government decision makers are keen on providing services using traditional approaches, they also need to understand the factors that would encourage their citizens to use the electronic service delivery channels [6]. In fact, the research on exploring factors that would encourage citizen to adopt e-government services in developing countries is not enough [6]. Therefore, one theaim, of this research, is to identify the factors that affect the citizens' intention to adopt e-government services.

C. The paper's structure

The structure of this paper is as follows: the followingsection discusses the background of e-government, the adoption of the e-government, and various models of citizens' adoption are reviewed. Section III discusses the challenges facing the implementing and developing e-government in Saudi Arabia specifically, followed by factors that would influence the Saudi Arabian citizens to adopt e-government. A new integrated model is proposed in Section IV. In section V, the approaches that have been used to validate the proposed model are presented, leading to Section VI that presents the results of the study. The paper concludes withSection VII.

II. Background

As Saudi Arabia has been chosen as the case study in this research, the related e-government drivers and barriers within the country areidentified and discussed. The review primarily concentrates on the adoption of e-government services by citizen, and different approaches that have been used to influence citizen to adopt e-government services.

A. e-government

To define e-government from a single perspective is relatively easy, but defining e-government in a way that suits everyone's view or needs is a significant challenge. Based on the work by Meng Seng, et al. [7], it has been noted that although -e-government as a term has become known across the world, there is evidence of insufficient consensus on its meaning, particularly regarding the main features of e-government. e-government can be defined in different ways. For instance, it can mean everything from just looking up information to using an online service, such as renewing a passport [2]. In addition, the use of information technology to enable and increase efficiency is key to e-government, while providing services and information to citizens, employees, businesses and government agencies [8]. A different approach is to define e-government as using the Internet as a tool for information and communications technology (ICT) to accomplish better government [9, 10]. Lu, et al. [11] define e-government as an ICT application that interacts efficiently, effectively, transparently and accountably with stakeholders. Lu, et al. [11] also identify three different transactional exchanges: government to government (G2G), government to business (G2B) and government to citizens (G2C), while Ndou [12] stated that e-government is simply providing information and engaging in digital transitions, which can be achieved through a simple web portal.

A wide range of different definitions from researchers have been identified; while everyone has a different point of view and requirements, most of them share the view that e-government incorporates ICT as one of its major elements.

From the definitions abovewe can conclude thate-government can be defined in terms of isa matrix of stakeholders: government to government, government to business and government to citizens, using information and communications technology to deliver and consume services. e-government has the objective of saving money, time and effort with increased efficiency, with due consideration for information security and privacy.

B. Challenges facing e-government

Developing any framework, such as e-government, that is capable of benefiting private and public organisations, results in a number of challenges for the different stakeholders, both internal and external, of the organisations [1]. Furthermore, in order to build e-government, there are some barriers facing implementation that should be kept in mind. Therefore, it is important to find out what these challenges and barriers are and how we can solve or avoid them. Later inthis section, we will discuss the most common challenges in general that other researchers have found, and present their solutions.

C. Adoption

Adoption is an important aspect for the success of e-government initiatives in developing countries [13]. However, growing interest in e-government raises the question of how governments can increase citizen adoption and use of their online government services [14]. To date, there has been little research exploring factors that determine the adoption of e-government services by citizens in developing countries, especially in the Arab world [6, 15]. Moreover, Dong, et al. [16] point out that e-government researchers often do not consider the adoption of e-government. They also make the point that, although there is enormous potential for online government services, citizens are not adopting them [14]. Furthermore, Carter and Belanger [8] agreed with other researchers that, although numerous studies have analysed user adoption of electronic commerce [17-19], to date, no study has identified the core factors that influence citizen adoption of e-government initiatives.

According to Colesca [20], many studies focused on the citizen adoption of e-government services suggest that trust [21], security [22] and transparency [23] are major issues for e-government adoption. Based on Margetts [24], cited by Yonazi, et al. [13], high adoption of these initiatives increases the chance that e-government will facilitate social and economic benefits to citizens. In Kuwait the increasing use of ICT by government departments resulted in the creation of an IT infrastructure capable of supporting e-government services [15]. User acceptance of IT is deemed a necessary condition for the effective implementation of any IT project [6, 25]. Adoption comes after direct experience with the technology and after an individual has decided to accept the technology [6, 26]. A number of studies have investigated the adoption of e-government services in developed countries [6, 27], whereas relatively little has been undertaken in developing countries [6, 15]. Successful implementation of adoptable e-government initiatives in that context requires complex customization between the technology and implementation context in developing countries [13, 28]; the result in designing citizen-adoptable e-government initiatives is still a challenge to many developing countries' governments [13]. AlAwadhi and Morris [6] conducted a study in Kuwait to explore factors that affect the adoption of e-government services, and concluded the main factors that could influence citizens to adopt e-government which includes, usefulness of e-government services, ease of learning and use, cultural and social influences, remove face-to-face interaction, gender issues, technical issues, lack of awareness, trust in the Internet, and cultural differences

However these factors influence Kuwaiti citizens to adopt e-government services, these factors might influence Saudi citizens since the culture in Kuwait and Saudi Arabia is almost identical. In addition,Alshehri, et al. [29] has identified somefactors the might influence the intention of the Saudi Arabian citizen to adopt e-government services. Therefore, in order to determine which of these factors can influence Saudi citizens and whether there are other factors that have not been mentioned, an investigation is going to be carried out among citizens of Saudi Arabia and selected Saudi organisations.

D. Previous Models used to measure adoption of new technologies

Many researchers have introduced models of citizen adoption. These models are constructed based on three widely used models, which are used to measure the acceptance of a new technology. The three models are the Technology Adoption Model, the Diffusion of Innovations Model, and the Unified Theory of Acceptance and Use of Technology. Trustworthiness is a fourth model that has been addressed by [30] to measure citizens' intention to use a new system. Trustworthiness has been used by Carter and Belanger [8] in their proposed model (TAM, DOI and Trust). Trustworthiness is referred to as the citizens' perception of the reliability and integrity of the electronic marketer [8, 30].

1) Technology Adoption Model and Diffusion of Innovations Model

Davis [31] proposed a model that can measure how far people can accept or reject a new technology. Technology's adoption depends on two basic attributes: Perceived Usefulness (PU) and Perceived Ease of Use (PEU) [31, 32]. Davis [31] defines perceived usefulness as "the degree to which a person believes that using a particular system would enhance his or her job performance". In contrast, perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of effort" [31]. The intention to use the system is determined by perceived usefulness and perceived ease of use [31, 32]. On the other hand, Rogers [33] has addressed a theory called Diffusion of innovations (DOI). The diffusion of innovations model is used to explain user adoption of new technologies in Information System research [8]. DOI consist of relative advantage, complexity, compatibility, trialability and observability.

2) Unified Theory of Acceptance and Use of Technology

The Unified Theory of Acceptance and Use of Technology (UTAUT) has been presented by [26]. It consists of five main constructs including performance expectancy, effort expectancy, social influence, facilitating conditions and behavioural intention, which play an important role as direct determinants of usage behaviour and user acceptance [26]. According to Venkatesh, et al. [26], these constructs are influenced by gender, age, voluntariness and experience.

E. Models proposed for influencing citizen to adopt e-government

To introduce the research model, in this section different researchers' models and contributions are going to be presented. First, Carter and Belanger [8] proposed a research model based on Technology Adoption Model (TAM), Diffusion of Innovations Model (DOI) and Trustworthiness. In addition, Carter and Belanger [8] proposed a research model that contains attributes from TAM, DOI and Trustworthiness model. Compatibility, Relative advantage and Complexity have been adopted from DOI, while Trialability and Observability have been excluded and replaced by image [8]. Image refers to "the degree to which the use of the innovation is seen as enhancing to an individual's image or social status" [8]. Carter and Belanger [8] have adopted Perceived Usefulness and Perceived Ease of Use from TAM. Trustworthiness has been adopted and included in the author's research model. Second, AlNuaimi, et al. [32] presented another research model for citizen adoption that is based on TAM, DOI and Unified Theory of Acceptance and Use of Technology (UTAUT). The model has been created to have attributes that have been adopted from TAM, DOI and UTAUT models with some modifications to suit use within the United Arab Emirates. The model has 11 independent variables and has been used to examine the impacts of these variables on the use of e-government services [32]. Rehman and Esichaikul [34] delivered a third model of citizen adoption based on integrated models adapted from TAM, DOI and UTAUT. Rehman and Esichaikul [34] defined factors that influence the citizens' intention to adopt e-government services in Pakistan and categorize them in their proposed model. As Pakistan is a developing country as Saudi Arabia, and also is using to implement and develop e-government, these proposed factors might affect the Saudi Arabia citizen.

There are some factors that have been mentioned by other researchers as influencing people to use e-government services. Cultural and social influences, including connection (Wasta[1]), face-to-face interaction, cultural differences and gender issues have an impact on the intention to use e-government services [35]. Privacy is another issue that influences citizen to adopt e-government services [1]. In addition, web usability and accessibility are also critical factors that affect the intention to use e-government services [36, 37].

III. Research Discussion

Based on the background research, the following discussion will consider two aspects of e-government in order to answer the following key questions:

a) What are the challenges or barriers to implement and develop e-government in Saudi Arabia?,

b) How can citizens adopt e-government? and

c) What are the influential factors to be integrated in a model for implementing and developing e-government in order to be adopted by citizen?

These aspects are the challenges facing e-government implementation and development in Saudi Arabia; and also considerthe factors that influence citizens' intention to adopt e-government services.

A. e-government challenges and barriers in Saudi Arabia

In section II we noted that many researchers have discussed challenges and barriers that face e-government implementation and development in many countries. Some of these challenges are common, such as security, privacy and trust, while there are other challenges that vary from country to country and from city to city, or even from department to department in one organisation. The Saudi Arabian e-government, for instance, has both general and specific challenges and barriers. The core question is: What are the challenges or barriers to implement and develop e-government in Saudi Arabia? To answer this question, some of the challenges which are mentioned by other researchers, are going to be investigated including accessibility,

[1] [1] It is Arabic word which means being served because you know someone in the organization otherwise you will not get these services if you don't know anyone, for instance, jumping the queue.

availability, citizen expectations, computer and information literacy, cost of Internet usage, culture, privacy and security, technical infrastructure, and trust. However, not all of these challenges face the Saudi Arabian government with its plan to introduce e-government.

B. Factors influencing citizens' intention to adopt e-government services in Saudi Arabia

The initial question for this research and investigation is: How can the Saudi government overcome challenges to help its citizens adopt e-government? To answer this question and to help people adopt e-government services, there are some factors that should be credited to government requirements. Therefore, in Table I the influential factors have been identified and grouped based on the Background.

IV. Research Model

In the previous sections we have developed an understanding of the requirements for the introduction of e-government in to Saudi Arabia. In this section we develop a suitable model. The model will be developed by adapting and integrating the critical factors that have been identified in the previous sections. Figure 1 shows the high level view of the proposed new model, which combines the intention to use e-government services and e-Readiness.

These two main blocks, "Intention to use e-government services" and "e-Readiness", have factors that affect the adoption of e-government services. The intention to use e-government services, which has been classified as citizens' concerns, includes Trust, Privacy, Security, Culture and Website design while e-Readiness has Quality Services, DOI, Skills and knowledge, Culture, Lack of Awareness, Technical Infrastructure and Security, and it is classified as government's responsibility. In this paper we will only present the breakdowns of the e-readiness block which is shown in Figure1.

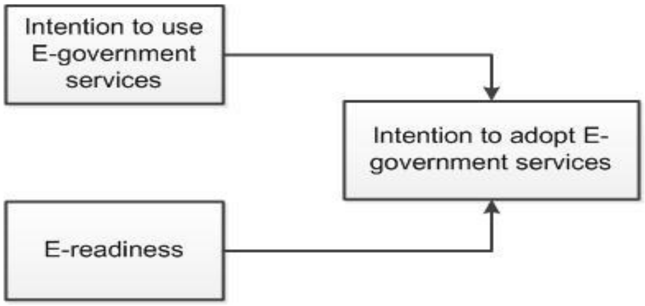

Intention to use E-government services
Intention to adopt E-government services
E-readiness

Fig. 1. A high level overview of an Integrated Model for Citizen Adoption of e-government Services in Saudi Arabia

TABLE I. Categories, Together With The Factors That Influences Citizens To Adopt E-Government Services.

Categories	Factors
Infrastructure Issues	Technical Infrastructure
Skills and Knowledge	• Computer and Information Literacy • Education • Age. • Gender.
Security Issues	• Information Security.
Quality of Service	• Service Quality. • Speed of Delivery. • Reliability. • Information Quality. • Availability.
DOI	• Compatibility. • Image. • Complexity. • Relative Advantage
Website Design	• Usability. • Multi-lingual Website. • Accessibility.
Awareness Issues	• Lack of Awareness
Culture Issues	• Culture

A. Quality of service

Quality of service has been suggested to play an important role in online services [34]. To encourage citizens to adopt e-government services, it is important for the government to provide high quality service and high quality information with the objective of speed of delivery, with due consideration of information reliability and availability [34].

- Service Quality: Service quality refers to the assessment done by the consumer for the overall excellence of the online provided service [38]. The government website should be designed carefully to address customers' needs because the face-to-face interaction is lacking in online service [39].

- Reliability: One critical issue regarding building an integral e-government to provide online services is making it reliable. Liu and Arnett [39] state that in customer online services, reliability is required. A system could be reliable when it has a quick error recovery [39], whereas service quality would be reliable when delivering services to the customers as promised [40]. Moreover, reliability is defined as the capability of a system to accomplish its intended function [41].

- Availability: It is important to the customers to use online services whenever they want. Therefore, system availability is an influential factor for the citizens' adoption of e-government services [34]. System availability refers to the probability of the system to be ready to provide responses at a specific time [42]. In addition, Lin and Chang [41] defined system availability as the expectation of a system to be available for operating tasks.

- Speed of Delivery Consumers of services or products are concerned about the speed of receiving their orders. Rehman and Esichaikul [34] identified speed of delivery as a critical factor of the quality service that influences citizens' intention to adopt e-government services. When a government increases the delivery speed of their online services, it would help the citizens to use and adopt the new services [40]. Furthermore, speed of delivery refers to the elapsed time between customers requesting services and receiving them [40].

- Information Quality The assessment of the government's website quality lists information quality as a key element [43]. Additionally, prior research employed various measures of IS success that result in the importance of the information quality for a website to success [39]. Bock, et al. [44] state that the degree to which the information on the website possesses the elements of content, usefulness, timeliness and accuracy is referred to as information quality.

B. *Skills and Knowledge*

Literacy as applied to ICT is defined as "whatever a person needs to be able to use (and know about) computers" [45], while "the ability to use information, or possibly the possession of knowledge of information is information literacy" [45]. The computer and information literacy are affected by the citizen's level of education, age and gender [1], which all bar the citizen to adopt e-government services [32]. Additionally, researchers have stated that the age of a person and the level of education can positively or negatively influence the intention to use e-government services [26, 32,34]. People, who have grown up among educated family and have got use to technology, have a highly chance to adopt a new technology e.g. e-government. Furthermore, Gender has played critical roles in influencing citizens' intention to use the e-government services [34]. It has been stated that people who are forty and below are more likely to welcome the usage of e-government services than older [20].

C. *Culture*

Culture impacts citizens' intentions to use e-government services, including cultural influences, culture awareness and national culture [35, 46]. Culture has been defined as "values, beliefs, norms and behavioural patterns of a group – people in a society for national culture, staff of an organisation for organisational culture, specific professions for professional" [47]. Akkaya, et al. [46] state that many researchers recognize the importance of considering cultural characteristics in the development and use of online services.

D. *Lake of Awareness*

Awareness refers to how a person understands the activities of others, which provides a context for his own activity [48]. To encourage citizens to adopt e-government services, the government should increase citizens' awareness. It has been found that awareness is one of the barriers that affect the adoption of e-government services [15, 35]. According to Baker and Bellordre [49], a major concern related to the deployment and use of new technologies is a lack of awareness that a given technology exists, or that the citizen could benefit from using the new technology.

E. *Technical Infrastructure*

Technical infrastructure can be defined as: "design and installation of LAN local area network, determination of cooperation scope in the corporate WAN network (Internet, Intranet), technical parameter specification of computers used as workstations and servers, selection of operational system environment and database platform" [50]. A study by AlAwadhi and Morris [35] found that most of the participants were worried about the technical issues. AlAwadhi and Morris [35] state that the findings give a clear view that technical infrastructure is important for encouraging citizens to adopt e-government services. In addition, Al-Sobhi, et al. [1] state that reliable and integrated technical infrastructure could be the difficult part facing the government, especially in developing countries, in obtaining a higher level of e-government services that can influence citizens to adopt e-government services. Also, Al-Sobhi, et al. [1] suggest that governments should provide a budget to build a strong technical infrastructure in order to encourage citizens to adopt e-government services.

F. *Diffusion of innovation and Website design*

This element of the DOI model is based on Rogers [33] model of Diffusion of Innovation, as discussed in the background Section E.1. Subsequently, Carter and Belanger [8] have made a modification by adopting compatibility, relative advantage and complexity, and excluding trialability and observability to replace them with image. Furthermore, as it is known that e-government and e-commerce are almost identical and both use online services, one of the key components of the online marketing strategy is the website; this means that good website design is required to serve the target market effectively and efficiently [51]. It is mentioned that a consideration of elements such as ease of navigation, accessibility, and features such as personalisation, customisation and multiple languages are required [51]. Combining these elements will directly influence users' experiences and encourage them to adopt the services [51]. In addition, researchers have suggested that the design of an e-government website may encourage citizens to use the services and make a good impression to increase citizens' repeated usage [34, 52]. Website design, includingusability, accessibility and multiple languages are the main factors that governments should focus on to influence citizens to adopt and use e-government services [34, 36].

- Usability: Website usability is a key aspect of website functionality [53]. Usability is defined as the ease with which users can access and navigate information in a portal with the objective of learning to manage the system and become familiar with basic functions [53].

Well-designed portals are easy to use and have pleasant, consistent interfaces [53]. Nielsen [54] states that improving the ease-of-use of a website during the design process by using methods known as usability. Also, usability refers to the quality attributes that measure how easy it is to use a user-interface, which includes five factors: learnability,efficiency, memorability, errors and satisfaction [54].

- Accessibility: Accessibility of a website is an essential factor that may affect citizens' intentions to use e-government services [37]. Website accessibility is defined as the degree to which citizens and automatic tools can access web information [36].
- Multi-Lingual Website and disabilities: Rehman and Esichaikul [34] suggest that building an e-government website with multi-lingual web support will positively influence the citizens' intention to adopt e-government services. Multi-lingual web support includes the official language with one or more additional well-known languages and output for disabled users, which allows citizens to access and navigate the information easily [53].

G. *Security*

It is mentioned that citizens concerned with information privacy have an impact on the consumers of electronic services [30]. According to Akkaya, et al. [46], citizens are sensitive towards storage of their personal data, which has a negative influence on the intention to adopt and continue e-government services. In Addition, security is defined as the protection of information or systems from unsanctioned intrusions or outflows [55]. Lack of security is one of the main factors that affect the intention to adopt e-government services that have been identified in most studies [55].

- Information Security: Information security is defined as "the subjective probability with which consumers believe that their personal information will not be viewed, stored or manipulated during transit or storage by inappropriate parties, in a manner consistent with their confident expectations" [56].

V. APPROACHES TO VALIDATE THIS MODEL

Since the factors have been integrated in the addressed model, it is essential to find out the importance of these factors on the citizens' intentions to adopt the e-government services. In this paper, these identified factors are going to be validated and confirmed using the Triangulation method. Triangulation is used to increase precision in empirical research [57].

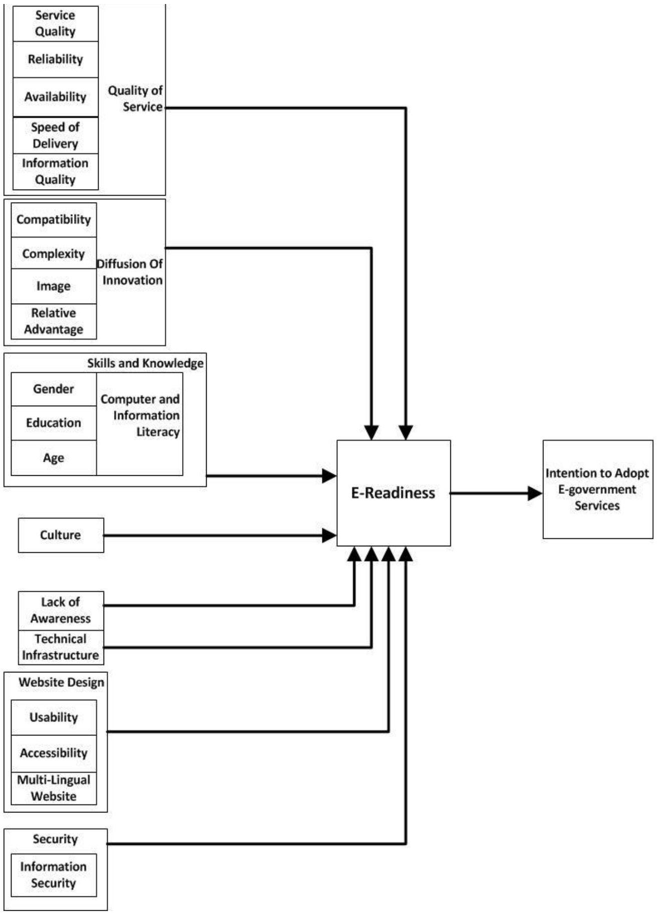


Fig. 2. A breakdown of the high level overview of the e-readiness of the Integrated Model for Citizen Adoption of e-government Services figure

According to Runeson and Höst [57], using the triangulation method by taking different angles towards the studied object will provide a broader picture. Runeson and Höst [57] also defined four different types of triangulation as follows:

- Data (source) triangulation—using more than one data source or collecting the same data at different occasions.
- Observer triangulation—using more than one observer in the study.
- Methodological triangulation—combining different types of data collection methods, e.g. qualitative and quantitative methods.
- Theory triangulation—using alternative theories or viewpoints.

In order to validate the proposed factors using triangulation methods, three main activities were undertaken. Firstly, a detailed background literature review has been conducted, which produced a summarized table of reviews from expert authors in this field. Secondly, questionnaires were distributed among Saudis' government employee. Finally, interviewsand questionnaires were conducted among government staff and leadership. However, the questionnaires are used as an exploratory study since there is no basis model for Saudi's e-government. The exploratory study gave a clear picture of the important factors affecting the adoption of e-government in Saudi Arabia.

VI. Exploratory Study And Results

Since the factors influencing Saudi's citizens are still argued, a study will be used to clarify the importance of the discussed factors and validating the proposed model. A study has been undertaken by developing a questionnaire for Saudi Arabiangovernment employee and an interview of experts in the information technology field. This study used mixed techniques, including questionnaires (structured questions) and interviews (semi-structured questions) to clarify the importance of the integrated factors and validate the proposed model. The surveys included a number of objectives which wereidentify challenges and factors valid from the employee's view, and identify challenges and factors valid from the expert's view.

A. Questionnaire applies to government employees

The questionnaires that wereansweredby employees who work at any government organisations have been designed to include open- and closed-ended questions. The closed-ended questions gather the opinions about the whether the proposed factors are important for adopting e-government services, whereas the open-ended questions asked for their opinions based on their experience and whether there were any missed factors. The government staff answered a questionnaire including twenty-three questions grouped under eight categories, which are: quality of service, culture, security, skills and knowledge, website design, lack of awareness, technical infrastructure and diffusion of innovation. The experts were asked for their opinion about all the proposed factors as closed-ended questions, while having open-ended questions to give their opinion about any missed factors.

B. Interviewingexperts in Saudi's e-government project

Similar to the government employees' questionnaires, the expert interviews were also designed to include open- and closed-ended questions. The closed-ended questions gather the opinions about the whether the proposed factors are important for adopting e-government services, whereas the open-ended questions ask for their opinions based on their experience and whether or not there were any missed factors. The experts were asked for their opinion about all the proposed factors as closed-ended questions, while having open-ended questions to give their opinion about any other missed factors.

C. Data analysis of the questionnaires

The government's employee survey wasdesigned to be conducted in person. Two government organisations werechosen to be part of this study. These organisations were selected because one has launched its website and started to provide basic online services, while the other is trying to implement an e-service to serve citizens. Thirty-five questionnaires were handed out and thirty-one responses were received, which was enough for this exploratorystudy. On the other hand, the interview survey wasdesigned to have all the proposed factors that have been discussed in the proposed model, and eightexperts were interviewed. This survey will get information from experts who work on the Saudi's e-government project. As the respondents for all surveys could respond between 1 (strongly disagreed) and 5 (strongly agreed), the results wasevaluatedusing aone-sample t-testwith the important value wasdefined as 3.5. The 'important value', is used to test if the collected data is statistically different. This is an exploratory study and the scale that has been chosen has 5 possible answers. Therefore3.5 has been identified as the lower value of importance.

1) Summarising the collected government employees' comments and suggestions

The government employees were asked to give feedback about the questioned factors and comments based on their experience. Respondents provided valuable comments and suggestions. Additionally, the majorities agreed with most of the questioned factors and believe these factors have an impact on the citizens' intentions to adopt the e-government services. However, some employees gave helpful comments that might be useful to be kept in mind for future investigation. The comments included:

- "Government should advertise in the media and launch a campaign to increase the citizens' awareness".
- "The media in Saudi Arabia has a negative impact on the citizens' intention to adopt e-government by announcing hacking crime and ignoring the developments and achievements in information security by the government".

2) Experts' comments

The interviewed experts' emphasised increasing the citizens' awareness of trust, privacy, security and gaining benefits from using e-government services. An expert suggests that introducing a demo about how to use the online services would make the use of the e-government website easier. Additionally, putting e-services machines and mobile kiosks would help citizens to become familiar with how to use the services and the benefits as well," the expert says.

D. Discussion

As it is mentioned previously, the collected data has been analysed using a one sample t-test. The accepted value to be statistically important has been identified as 3.5 and above. The test showed that the result of most of the identified factors have been seen adequate Table II. The result of data analysis of the factor Multi-Lingual has been accepted by the result of the data analysis for the questionnaires applied to citizen and government's employee as well as from the background, even though the experts' result is seen not adequate. The reliability of the results is givenin Table III, using Cronbach's alpha. Therefore, the identified factors have been accepted and being integrated in the proposed model in order to be used in future.

TABLE II. THE RESULT OF THE ONE SAMPLE T-TEST OF THE FULL STUDY

Factors	Government Employees	Experts	Result
Information Security	.016	<.001	Accepted
Culture	.006	.004	Accepted
Multi-Lingual	.008	.052	Accepted[2]
Usability	<.001	.007	Accepted
Accessibility	<.001	.004	Accepted
Relative Advantage	<.001	.013	Accepted
Compatibility	<.001	.031	Accepted
Image	.001	.020	Accepted
Complexity	.014	.013	Accepted
Computer and Information Literacy	<.001	.033	Accepted
Gender	.009	.013	Accepted
Education	<.001	.013	Accepted
Age	.005	.020	Accepted
Technical Infrastructure	.001	.013	Accepted
Lack of Awareness	<.001	.048	Accepted
Service Quality	<.001	.013	Accepted
Reliability	<.001	.007	Accepted
Availability	<.001	.007	Accepted
Speed of delivery	.036	.007	Accepted
Information quality	.016	.013	Accepted

TABLE III. RELIABILITY ANALYSIS FOR THE QUESTIONNAIRES, AS MEASURED BY CRONBACH ALPHA

Questionnaire	Government's employees	Saudi Arabian experts on e-government
Cronbach's α	0.846	0.664

From the survey's analysis, the proposed factors have been accepted to be integrated in the proposed model. Furthermore, since one of the author is a Saudi citizen and experienced how the services are delivered in Saudi Arabia, he found that there is a problem with delivering services due to the postal infrastructure. The Saudi citizens do not have a clear address, such as house number or a clear street name, or secure postal address which usually available in front of their house or a box beside their doors. Although there are post office Boxes available for the citizen rent, not many Saudis have one, which makes the communication between the citizens and the government difficult. Therefore, the mail service system in Saudi Arabia needs to be developed and improved to become a certified contact. The postal address should be introduced as the main contact between government and citizen.

VII. CONCLUSION

Since the rapid development of information and communication technologies (ICT) and the significant improvements in digital connectivity, adoption of e-government services by citizens is the concern of many governments. This paper considers how to encourage citizens to adopt e-government services and the challenges facing implementation and development of e-government.

Initially, it is important to know how Electronic Government (e-government) is defined. e-government can be defined based on an existing set of requirements, since there is no unique definition. e-government has been developed and implemented for a considerable period of time in developed countries, while it is still being implemented and developed in most developing countries. This results in many benefits that e-government services have addressed to governments, businesses and citizens. In addition, many researchers have found and discussed challenges that face the implementation and adoption of e-government. There are common challenges such as privacy, security, trust, culture, skills and knowledge, and IT infrastructure.

There are also many other more specific challenges, including authentication, digital divide and funding shortage, facing some countries. Adoption is a critical issue to governments that want to implement and develop e-government. However, governments can find aspects of the process can influence and encourage citizens to adopt e-government services. Nevertheless, challenges and barriers can be overcome by investigating various approaches to adopting e-government services and presenting an appropriate model that can suit most similar countries, including Gulf States. Additionally, the core questions of this research are (i) What are the challenges or barriers to implementing and developing e-government in Saudi Arabia?, (ii) How can the Saudi Arabian government overcome these challenges?, (iii) How can citizens adopt e-government? and (iv) What are the influential factors to be integrated in a model for implementing and developing e-government in order to be adopted by citizens? A discussion and investigation has been conducted to answer these questions. The result is presented in the previous sections which show the factors that would influence the Saudi citizens to adopt e-government services. However, it is proposed that a large survey will be conduct in the near future; this will lead to a far better model. Supported by SEM.

ACKNOWLEDGMENT

The authors acknowledge the award of a King Abdullah scholarship to SulaimanAlateyah to allow this research to be undertaken.

REFERENCES

[1] F. Al-Sobhi, V. Weerakkody, and M. M. Kamal, "An exploratory study on the role of intermediaries in delivering public services in madinah city: Case of saudiarabia," Transforming Government: People, Process and Policy, vol. 4, pp. 14-36, 2010.

[2] United Nations (2010, 05/07). The united nations e-government development database. Available: http://www2.unpan.org/egovkb/about/index.htm

[2] Based on the literature review and the employees' result.

[3] United Nations "United nations e-government survey : Leveraging e-government at a time of financial and economic crisis," New York2010 2010.

[4] S. Sahraoui, G. Gharaibeh, and A. Al-Jboori, "E-government in saudiarabia: Can it overcome its challenges?," eGovernment Workshop ’06, 2006.

[5] O. K. Lean, S. Zailani, T. Ramayah, and Y. Fernando, "Factors influencing intention to use e-government services among citizens in malaysia," International Journal of Information Management, vol. 29, pp. 458-475, 2009.

[6] S. AlAwadhi and A. Morris, "The use of the utaut model in the adoption of e-government services in kuwait," presented at the Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences, 2008.

[7] W. Meng Seng, N. Hideki, and P. George, "The use of importance-performance analysis (ipa) in evaluating japan's e-government services," Journal of Theoretical and Applied Electronic Commerce Research, vol. 6, pp. 17-30, Aug. 2011.

[8] L. Carter and F. Belanger, "Citizen adoption of electronic government initiatives," in System Sciences 2004. , Proceedings of the 37th Annual Hawaii International Conference, 2004, p. 10

[9] OECD, E-government studies: The e-government imperative: OECD Publishing, 2003.

[10] I. Alghamdi, R. Goodwin, and G. Rampersad, "E-government readiness assessment for government organizations in developing countries," Computer and Information Science, vol. 4, pp. 3-17, May 2011 2011.

[11] L. Lu, G. Zhu, and J. Chen, "An infrastructure for e-government based on semantic web services," in International Conference on Services Computing, 2004.

[12] V. Ndou, "E – government for developing countries: Opportunities and challenges," EJISDC, vol. 18, pp. 1-24, 2004.

[13] J. Yonazi, H. Sol, and A. Boonstra, "Exploring issues underlying citizen adoption of egovernment initiatives in developing countries: The case of tanzania.," Electronic Journal of e-Government, vol. 8, pp. 176-188, Dec 2010.

[14] M. Warkentin, D. Gefen, P. A. Pavlou, and G. M. Rose, "Encouraging citizen adoption of e-government by building trust," Electronic Markets, vol. 12, pp. 157-162, 2002.

[15] H. AlShihi, "E-government development and adoption dilemma: Oman case study," in The 6th International We-B (Working for eBusiness) Conference, Victoria University, Melbourne, Australia 2005.

[16] X. Dong, L. Xiong, and W. Wang, "How adoption is g2c model e-government? &#x2014; evidence from xi' an and nan jing," in E -Business and E -Government (ICEE), 2011 International Conference on, 2011, pp. 1-4.

[17] P. Pavlou, "Integrating trust in electronic commerce with the technology acceptance model: Model development and validation," in Americas Conference on Information Systems 2001, p. 159.

[18] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," Information Systems Research, vol. 13, pp. 334-359, 2002.

[19] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and tam in online shopping: An integrated model," MIS Quarterly, vol. 27, pp. 51-90, 2003.

[20] S. E. Colesca, "Increasing e-trust: A solution to minimize risk in e-government adoption," JOURNAL OF APPLIED QUANTITATIVE METHODS, vol. 4, pp. 31-44, 2009.

[21] S. C. Srivastava and T. S. H. Teo, "Citizen trust development for e-government adoption: Case of singapore," in “Proocedings of Pacific Asia Conference on Information Systems, 2005, pp. 721-724.

[22] S. Colesca, "The main factors of on-line trust," Economia. Seria Management, vol. 10, pp. 27-37, 2007.

[23] S. Marche and J. D. McNiven, "E-government and e-governance: The future isn't what it used to be," Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration, vol. 20, pp. 74-86, 2003.

[24] H. Margetts, "E-government in britain—a decade on," Parliamentary Affairs, vol. 59, pp. 250-265, 2006.

[25] J. K. Pinto and S. J. Mantel, Jr., "The causes of project failure," IEEE Transactions on Engineering Management, vol. 37, pp. 269-276, 1990.

[26] V. Venkatesh, M. G. Morris, B. D. Gordon, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, vol. 27, pp. 425-478, 2003.

[27] R. Titah and H. Barki, "E-government adoption and acceptance: A literature review," International Journal of Electronic Government Research (IJEGR), vol. 2, pp. 23-57, 2006.

[28] R. Heeks, Implementing and managing egovernment: An international text. London: SAGE Publications Ltd 2006.

[29] M. Alshehri, S. Drew, and O. Alfarraj, "A comprehensive analysis of e-government services adoption in saudiarabia: Obstacles and challenges," International Journal of Advanced Computer Science and Applications, vol. 3, pp. 1-6, 2012.

[30] F. Belanger, J. S. Hiller, and W. J. Smith, "Trustworthiness in electronic commerce: The role of privacy, security, and site attributes," The Journal of Strategic Information Systems, vol. 11, pp. 245-270, 2002.

[31] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, vol. 13, pp. 319-340, 1989.

[32] M. AlNuaimi, K. Shaalan, M. Alnuaimi, and K. Alnuaimi, "Barriers to electronic government citizens' adoption: A case of municipal sector in the emirate of abudhabi," in Developments in E-systems Engineering (DeSE), 2011, 2011, pp. 398-403.

[33] E. M. Rogers, Diffusion of innovations, 4th ed. New York: Free Press, 1995.

[34] M. Rehman and V. Esichaikul, "Factors influencing the adoption of e-government in pakistan," in E -Business and E -Government (ICEE), 2011 International Conference on, 2011, pp. 1-4.

[35] S. AlAwadhi and A. Morris, "Factors influencing the adoption of e-government services," Journal of Software, vol. 4, pp. 584-590, 2009.

[36] A. Abanumy, A. Al-Badi, and P. Mayhew, "E-government website accessibility: In-depth evaluation of saudiarabia and oman," The Electronic Journal of e-Government, vol. 3, pp. 99-106, 2005.

[37] M. K. Alomari, P. Woods, and K. Sandhu, "Predictors for e-government adoption in jordan: Deployment of an empirical evaluation based on a citizen-centric approach," Information Technology & People, vol. 25, pp. 4-4, 2012.

[38] R. N. Bolton and J. H. Drew, "A multistage model of customers' assessments of service quality and value," Journal of Consumer Research, vol. 17, pp. 375-384, 1991.

[39] C. Liu and K. P. Arnett, "Exploring the factors associated with web site success in the context of electronic commerce," Information & Management, vol. 38, pp. 23-33, 2000.

[40] A. Trentin, E. Perin, and C. Forza, "Overcoming the customization-responsiveness squeeze by using product configurators: Beyond anecdotal evidence," Computers in Industry, vol. 62, pp. 260-268, 2011.

[41] Y.-K. Lin and P.-C. Chang, "Evaluation of system reliability for a cloud computing system with imperfect nodes," Systems Engineering, vol. 15, pp. 83-94, 2012.

[42] T. Walkowiak, "Web systems availability analysis by monte-carlo simulation," Computer Modelling and New Technologies, vol. 15, pp. 37-48, 2011.

[43] V. McKinney and K. Yoon, "The measurement of web-customer satisfaction: An expectation and disconfirmation approach," Information Systems Research, vol. 13, pp. 296-315, 2002.

[44] G.-W. Bock, J. Lee, H.-H. Kuan, and J.-H. Kim, "The progression of online trust in the multi-channel retailer context and the role of product uncertainty," Decision Support Systems, vol. 53, pp. 97-107, 2012.

[45] I. J. Cole and A. Kelsey, "Computer and information literacy in post-qualifying education," Nurse Education in Practice, vol. 4, pp. 190-199, 2004.

[46] C. Akkaya, P. Wolf, and H. Krcmar, "Factors influencing citizen adoption of e-government services: A cross-cultural comparison (research in progress)," in System Science (HICSS), 2012 45th Hawaii International Conference on, 2012, pp. 2531-2540.

[47] M. Ali, V. Weerakkody, and R. El-Haddadeh, "The impact of national culture on e-government implementation: A comparison case study.," in Proceedings of the Fifteenth Americas Conference on Information Systems., San Francisco, California, 2009, pp. 1-13.

[48] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces," presented at the Proceedings of the 1992 ACM conference on Computer-supported cooperative work, Toronto, Ontario, Canada, 1992.

[49] P. M. A. Baker and C. Bellordre, "Adoption of information and communication technologies: Key policy issues, barriers and opportunities for people with disabilities," in System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, 2004, p. 10 pp.

[50] A. Kaminski, "Computer integrated enterprise in the mrp/erp software implementation," foundations of management, vol. 4, p. 25, 2010.

[51] V. Kumar, B. Mukerji, I. Butt, and A. Persaud, "Factors for successful e-government adoption: A conceptual framework," Electronic Journal of e-Government, vol. 5, pp. 63-76, 2007.

[52] S. M. Pi, H. L. Liao, and H. M. Chen, "Factors that affect consumers' trust and continuous adoption of online financial services," International Journal of Business and Management, vol. 7, pp. 108-119, 2012.

[53] J. P. Gant and D. B. Gant, "Web portal functionality and state government e-service," in System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on, 2002, pp. 1627-1636.

[54] J. Nielsen, "Usability 101: Introduction to usability," JakobNielse'sAlertbox, August 25 2003.

[55] D. Berdykhanova, A. Dehghantanha, and K. Hariraj, "Trust challenges and issues of e-government: E-tax prospective," in Information Technology (ITSim), 2010 International Symposium in, 2010, pp. 1015-1019.

[56] K. C. Ramnath and A. P. Paul, "Perceived information security, financial liability and consumer trust in electronic commerce transactions," Logistics Information Management, vol. 15, pp. 358-368, 2002.

[57] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," Empirical Software Engineering, vol. 14, pp. 131-164, 2009.

Grouping-based Scheduling with Load Balancing for Fine-Grained Jobs in Grid Computing

Rabab Mohamed Ezzat
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Amal Elsayed Aboutabl
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Mostafa Sami Mostafa
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—Grid computing is characterized by the existence of a collection of heterogeneous geographicallydistributed resources that are connected over high speed networks.Job scheduling and resource management have been a great challenge to researchers in the area of grid computing.Very often, there are applications having alarge number of fine-grainedjobs.Sending these fine-grained jobsindividually to be executed on grid resources that have high processing power reduces resource utilization and is thus uneconomical. This paper presents efficient grouping-based scheduling models that group fine-grained jobs to form coarse-grained jobs which are sent for execution on grid resources. Our groupingstrategy is based on the processing capability of resources and the processing requirements of grouped jobs.A load balancing approach is also presented to achieve efficient utilization of resources. Simulation experiments were conducted using the Gridsim toolkit. Results show that thetotal simulation time and the cost are improved by grouping.Furthermore, our load balancing approach enhances resource utilization and achieves load balancing among resources.**

Keywords—grid computing;job scheduling; job grouping; load balancing; resource utilization; Gridsim

I. INTRODUCTION

Grid Computing is a computing paradigm that emerged in the late 1990's [10]. The emergence of this paradigm was mainly due tothe spread of powerful computers that have high computing power at low cost in addition to the popularity of the internet and availability of high speed networks [1]. Grid computing allows sharing and using of geographically distributed resources including supercomputers, data sources and specialized devices that are owned by different organizations [2].There are large scale compute-intensive problems in different fields such as engineering, science and economics that need high computing power to be solved. Grid computing enables sharing resources that are connected through the internet for solving these problems.

Resource management and job scheduling have been a cause of great challenge to researchers in the field of grid computing[7]. Grid scheduling is a complex process which differs from scheduling in traditional distributed systems because of the characteristics of grid computing environment:

- Resources are geographicallydistributed over different multiple administrative domains.
- Resources are not under central control.
- Resources are heterogeneous; different in architecture and management policies.
- Jobsin the grid are from different usershaving different requirements.

Many applications consist of a large number of fine-grained jobs having small scale processing requirements. Sending these jobs individually to be executed on grid resources that have high processing power reduces resource utilization and is thus uneconomical. Moreover, the total communication time for transmitting each fine-grained job to the resource may exceed the total computation time of that job on the resource.In grid computing, and for such type of applications, having coarse-grained jobs is more efficient and cost-effective than fine-grained jobs [12]. Therefore, instead of sending such jobs individually, coarse-grained jobs can be created by collecting a suitable number of jobs [3]. Grouping fine-grained jobs together to form coarse-grained jobs reduces the transmission time and increases resources utilization [4]. The total processing time needed for each fine-grained job includes:scheduling time that is the time spent to schedule the job, Sending time that is thetime spent to send the job to a resource, execution time that is the time spent to execute the jobandreceiving time that is the time spent to receive the job from a resource after execution.

Scheduling is the process of assigning or mapping jobs to suitable resources that execute jobs achieving the following goals [4]:

- Minimizing the processing time.
- Minimizing processing cost.
- Achieving load balancing among resources.

A *grid computing scheduler* is responsible for selecting the most suitable machine or computing resource for processingeach job to achieve maximum system throughput. In case of having fine-grained jobs requesting service on the grid, these fine-grained jobs are grouped to form coarse-grained jobswhich are then handled by the scheduler in such a way that achieves loadbalancing [7].

Load balancing is a mapping strategy that distributes applications load among resources so that there will be efficient utilization of resources and hence the performance of the system is improved [1] [6]. Load balancing algorithms of traditional parallel and distributed systems cannot be used in grid computing because of the special characteristics of grid environments. In grid computing environment, resources differ in their computational power. Efficient load balancing algorithms are neededto maximize resources utilization and prevent the condition where some resources may be overloaded and other resources may be idle [2].

The paper is organizes as follows. Section II discusses previous works in the area of grouping-based grid scheduling.Section III presents our proposed models. Section IV provides a detailed description of our simulation, experiments and results. Finally, Section V provides a conclusion of this work.

II. Previous Work

Grid scheduling is a complexprocess that has been a challenge for researchers due to the heterogeneity of the grid environment. There have been a number of attempts in the area of grid scheduling in the literature since the grid computing paradigm emerged. In particular, we focus here onmodels that were developed to group and schedule fine-grained jobsin grid environments.

Constraint-based job and resource scheduling in grid computing [15] is a model in which resources are arranged in a hierarchical manner so that the resource with the highest computation power can be found using tree heap sort while jobs are grouped according to the processing capability, bandwidth and memory size of resources. A job scheduling model based on grouping was developed in [10]. In this model, resources are sorted in ascending order of their processing capability and then jobs are grouped according to the processing capability, bandwidth and memory size of the selected resources.

Two other grouping-based models were developed in [12] and [13] where resources are sorted according to bottleneck bandwidth and group jobs according to processing capability and bandwidth of the resources. In both the two models using the bandwidth strategy is not efficient to transfer the jobs. In [13] grouping strategy does not utilize the resources sufficiently. Fine-grained jobs are scheduled in [11] according to processing capability and bandwidth of selected resources.A dynamic job scheduling approach which is based on grouping is proposed in [14] for deploying applications with fine-grained tasks on global grids and is based on the processing capability of resources. This model reduces the processing time and communication time but does not utilize the resources sufficiently.A time minimization dynamic job grouping based scheduling is proposed in [5]. Resources are sorted in descending order based ontheir processing capability then fine-grained jobs are grouped according to processing capability of the selected resource by taking one job from the front of the sorted job list and one job from the end.An agent-based dynamic resource scheduling model with FCFS job grouping strategy is presented in [9]. Another algorithm which is based on grouping and takes into consideration both the memory requirements and execution time of jobs is presented in [7].

All of the previous researches depend on grouping fine-grained jobs to obtain coarse-grained jobs which are then sent to selected resources to be executed. The focus in previous researcheshas mainly been reducing processing and communication time. However, increasing the efficiency of resource utilization bybalancing load among resources has not been sufficiently addressed.Two load balancing approaches for computational grids were presented in [1] and [6] where a job is sent to the resource that has minimum queue length. Load balancing is achieved in the grid but at the expense of high overall execution time caused by increased communication time incurred by sending fine-grained jobs individually.

III. Proposed Models

Our proposed model consists of two parts; a grouping strategy and a scheduling model. Our grouping strategy groups fine-grained jobs to form a smaller number of coarse-grained jobs depending on the processing capability of the selected resource and the jobs' processing requirements. The UFF (User-Finished-First) scheduling model and URS (Users-Resources-Sharing) scheduling model are two different proposed models that group fine grained jobs and schedule these jobs in two different ways. RMQ (Resource with Minimum Queue Length) scheduling model is another proposed model that group fine grained jobs and schedule these jobs to the resource with minimum number of waiting jobs. This model is a load balancing approach based on the queue length of the available resource.

A. Grouping Strategy

The job scheduler groupsfine-grained jobsbased on both the processing requirements of each job and the processing capability of each resource.First, the scheduler selectsa resource from the ResourcesList and computes the product of MIPS (million instructions per second) which is used to define a resource's processing capabilityand G.T (granularity time)that is a user defined parameterwhich is used to measure the total number of jobs that can be completed within a specified period of time. In UFF scheduling model and URS scheduling model the resources in the ResourcesList are sorted in ascending order based on processing capability of each resource. In RMQ scheduling model select the resource that have minimum queue length (minimum number of waiting jobs).Second, the scheduler selectsfine-grained jobs from GridletList one after the other. Collect the MI of the selected fine-grained jobs. Each job's MI (million instructions) defines the computational power needed to execute the job. The grouping step ensures that the total required computational power of grouped jobsdoes not exceed the processing capability of the resource.The term gridlet is used here to refer to a job that can run independently and sequentially on a grid resource.

The detailed steps of the grouping strategy are listed below.

1) Availablegridletsare sent to the job scheduler for scheduling.

2) Grid resources register their information atthe Grid Information Service (GIS).

3) The job scheduler requestsresources information from GIS. GIS sends the information of available resources to the scheduler.

4) Sort resources in ResourcesList in ascending order based on resource processing capability (MIPS).

5) Sort gridlets of each user in a separate GridletsListin ascending order based on gridlet length (MI).

6) Get the first user.

7) Initialize indices of GridletsList, ResourcesListand Grouped_Gridlets, named I, X and J respectively, all initialized to 0.

8) Selectthe resource specified by ResourcesList [X].

9) Select job specified by GridletList [I] of current user.

10) Groupedjob_length=0.

11) MI of ResourcesList[X]= MIPS of ResourcesList[X]* Granularity time.

12) For(I=0 to GridletsList size-1)

13) {

14) If(Groupedjob_length< MI of ResourcesList[X]))

15) Groupedjob_length= Groupedjob_length +length of GridletsList [I]

16) Else

17) {

18) Groupedjob_length= Groupedjob_length - length of GridletsList[I]

19) Create a new job,Grouped_Gridlet [J],whose length is equal to Groupedjob_length.

20) Increment J.

21) Break;

22) }

23) }

Starting from line 24, two different sequences of steps are presented reflecting two different scheduling models.

B. UFF *Scheduling*

After forming the grouped jobs, the question is whether to map the grouped jobs of a certain user to the available resources or let a number of users share the resources. In the *UFF (User-Finished-First)*scheduling model, the grouped jobs of a user are assigned to the available resources in parallel before proceeding with the next user. Accordingly, the first user sends grouped job 0 to resource 0 then grouped job 1 to resource 1 and so on till the last resource is reached. The next user is selected and the same steps are repeated. In lines 24-29, Total_Resources refers to the number of resources.

UFF (User-Finished-First) Scheduling Model

24) Submit Grouped_Gridlet [J] to ResourcesList[X].
25) X++.
26) If (X== Total-Resources)
27) { X==0
28) Get Next_User}
29) Go to step 8.

C. URS *Scheduling*

In this model, grouped jobs of different users are assigned to available resources in parallel. Therefore, grouped job0 of n different users are sent to the first n resources in the list. If the number of users is less than the number of resources, repeat starting from the first user with grouped job 1 and so on. When the last resource is reached, start from the first resource.

URS (Users-Resources_Sharing) Scheduling Model

24) Submit Grouped_Gridlet [J] to ResourcesList [x].
25) Get Next_User.
26) X++.
27) If (X== Total-Resources)
28) X==0.
29) Go to step 8.

D. RMQ (Resource with Minimum Queue Length) Scheduling Model

In this model, the scheduler selects the resource that has minimum queue length (minimum number of waiting jobs). Then, the scheduler uses the grouping strategy that based on the processing capability of the resource in addition to the processing requirements of the jobsto group fine-grained jobs and form-grouped jobs. Then send these grouped jobs to the resources to be executed. This model reduces the processing time and cost and achieves load balancing among resources. The following steps show the grouping strategy together with the load balancing approach.

1) Available gridlets are sent to the job scheduler for scheduling.

2) Grid resources register their information at the Grid Information Service (GIS).

3) The job scheduler requests resources information from GIS. GIS sends the information of available resources to the scheduler.

4) Sort gridlets of each user in a separate GridletsList in ascending order based on gridlet length (MI).

5) Get first user.

6) Initialize indices of GridletsList, ResourcesListand Grouped_Gridlets, named I, X and J respectively, all initialized to 0.

7) Select ResourceList [X] such that it is the resource with minimum QLength.

8) Select job specified byGridletList [I] of current user.

9) Groupedjob_length=0

10) MI of ResourceList[X]= MIPS of ResourceList[X]* Granularity time

11) For(I=0 to GridletList size-1)

12) {

13) If(Groupedjob_length< MI of ResourceList[X])

14) Groupedjob_length= Groupedjob_length +length of GridletList [I]

15) Else

16) {

17) Groupedjob_length= Groupedjob_length - length of GridletList[I]

18) Create new job Grouped_Gridlet [J] with length equal to Groupedjob_length

19) Increment J.

20) Break

```
21) }
22) }
23) Submit Grouped_Gridlet [J] to ResourceList [X] (resource with minimum QLength)
24) QLength[X]= QLength[X]+1
25) Get Next_User
26) Go to step 7
```

IV. Experimental Work And Results

Grid computing environment is a dynamic environment so it is extremely difficult to perform repeated experiments and studies on this environment in practice.Using simulation in such studies helps in performing a large number of experiments with various parameters. In this work, Gridsim was used for simulating the grid resources, jobs as well as our grouping strategy together with the proposed scheduling and load balancing models.

Most of the previous researches in the area of grid scheduling and resource allocation are based on a single user. Inthis work, multiple users are assumed. Each user has a number of independent jobs (Gridlets) that will be scheduled and then executed on heterogeneous resources taking into consideration resources load balancing.

A. Gridsim Simulation Environment

Gridsim is a java based discrete event grid simulator toolkit that allows modeling and simulation of grid computing system entities: resources, gridlets, scheduler, grid information service and users.Gridsimis also used to test scheduling and load balancing models [2]. Gridsim users are able to model and simulate the characteristics of grid resources and networks with different configurations. It, therefore, allows researchers to study grids and test new algorithms and strategies in a controlled environment.In Gridsim terminology,Gridlets are jobs that could run independently and sequentially on grid resources.

A grid environment is built using the Gridsim5-2 toolkit. After installing the Gridsim5-2 toolkit, the Gridsim package is imported. The grid environment is simulated using Jcreator by writing java code and implementingour grid entities:

- Create grid user(s):Multiple users are allowed. Each user in Gridsim must have a unique id.
- Create grid resources: Resources in Gridsimare defined by *resource name*, *communication speed*, *resource characteristics* (operating system, architecture, and cost), *and number of machines*.Each machine may consist of a *number of processing elements* each processing element is defined by a*unique id* and *processing capability*in*MIPS* (millions of instructions per second).
- Create gridlets:A gridlet is defined by *gridletlength*;*input file size* and*output file size*.

B. Simulation Input and Output

A number of simulation parameters are fed into the simulator:

- Gridlets: the number of gridlets.
- A_MI: average gridlet length in MI reflecting the processing requirementsof the job.Based on a gridlet's MI, the resource that hasa suitable processing capabilityis selected to execute this gridlet.
- Deviate%:MI deviation percentagewhich is used to create different number of gridlets that have different lengths.
- G_Time: Granularity time (expected job processing time). It is a measure of thenumber of jobs that can be completed within a certain time on a particular resource [14].
- OH_Time: Gridlet overhead time. In real environments, overhead time for each job depends on the current network load and speed. In our simulation, the overhead time of each gridletis an input value.
- Resources: Resources to be used in a simulation experiment are selected from the resources list.

After simulation input parameters have been defined, we conduct oursimulation experiments with andwithout grouping. This allows us to compare between scheduling fine-grained jobs with grouping and scheduling fine-grained jobs without grouping. This also allows us to study the effect of grouping the input gridlets on the overall performance. Two performancemetrics are used in this respect: *total processing time* and *total processing cost*. Total processing time is computed based on:

- Gridlet overhead processing time.
- Time taken to perform grouping.
- Time taken for sending gridlets to the resources.
- Time of processing the gridlets at the resources.
- Time taken for receiving the processed gridlets.

The total processing cost is computed based on:

- The time taken for computing the gridlets at the grid resource
- The cost specified at the grid resource.

C. Results

Tables 1, 2 and 3 show three different sets of simulation input parameters denoted by SI1, SI2 and SI3. Three users are assumed where each user is defined by a number of gridlets, MI and Deviate % of these gridlets as explained before. SI1, SI2 and SI3 assume 4, 7 and 5 resources respectively. Resources MIPS are varied to simulate resources' heterogeneity. Granularity and overhead time are also input.

TABLE I. Simulation Inputs Si1

User	Gridlets	MI	Deviate %
User1	100	10	10
User2	50	20	20
User3	150	30	30
Resources: R7,R5,R3,R1			

Resources_MIPS: 66,60,39,20
G_Time =5, **OH_Time**=5

TABLE II. SIMULATION INPUTS SI2

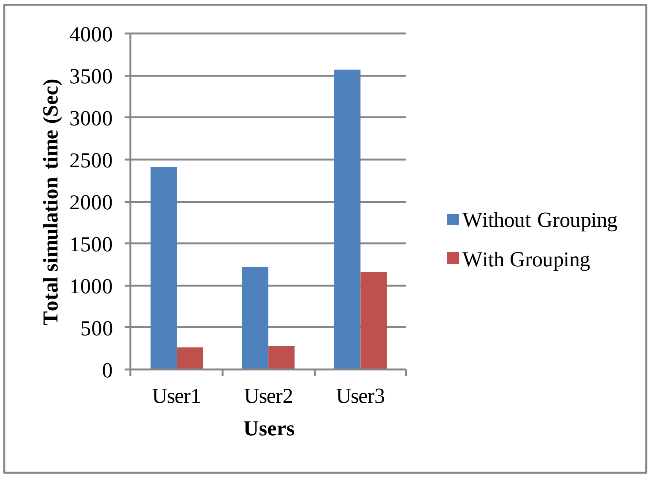
User	Gridlets	A_MI	Deviate %
User1	100	10	10
User2	50	20	20
User3	150	30	30
Resources: R4,R6,R7,R5,R3,R2,R1			
Resources_MIPS: 120,72,66,60,39,24,20			
G_Time =5, **OH_Time**=5			

TABLE III. SIMULATION INPUTS SI3

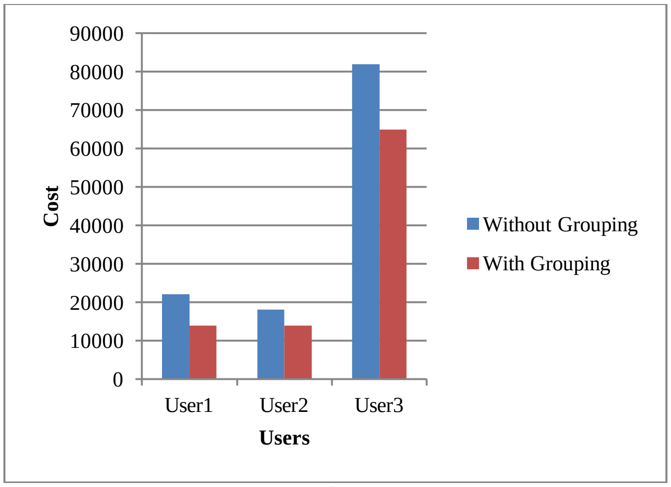
User	Gridlets	MI	Deviate %
User1	100	10	10
User2	150	20	20
User3	200	30	30
Resources:R4, R5,R3,R2,R1			
Resources_MIPS:120,60,39,24,20			
G_Time =5, **OH_Time**=5			

Figures 1, 2 and 3 show the results of total simulation time and cost with and without grouping in case of using simulation input parameters in table 1 for the three proposed modelsrespectively. Simulation time and cost are improved by using the grouping strategy.

This is due to the fact that total communication time is higher in case of scheduling the fine-grained jobs individually without grouping. On the other hand,when the grouping strategy is used, fine-grained jobs are grouped into a fewer number of coarse-grained jobs thus reducing the overall communication time.Similar results are obtained using simulation input parameters SI2 in table 2 as in Figures 4, 5 and 6.

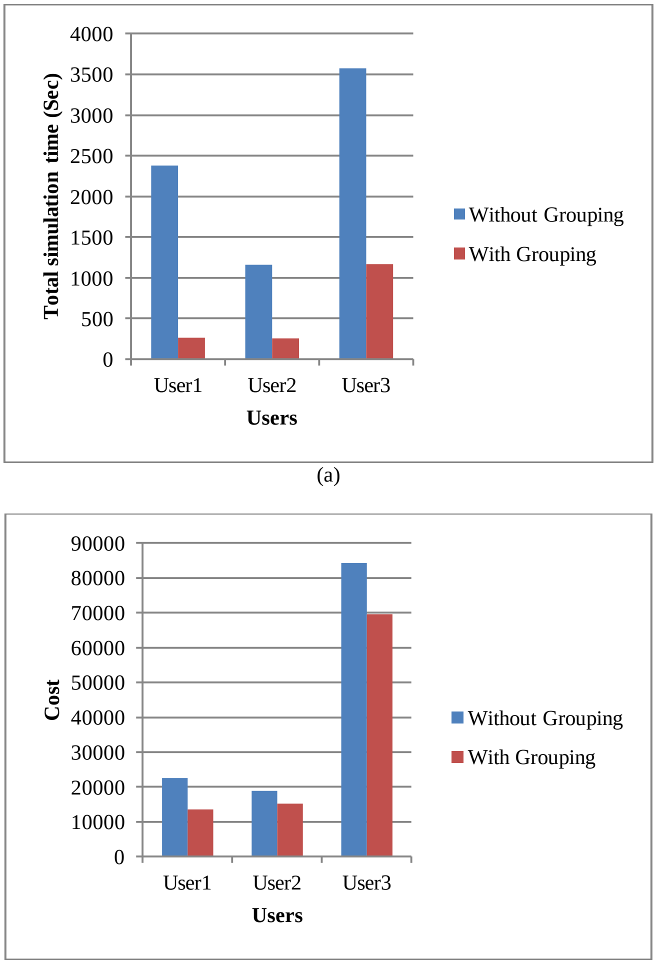


(a)



(b)

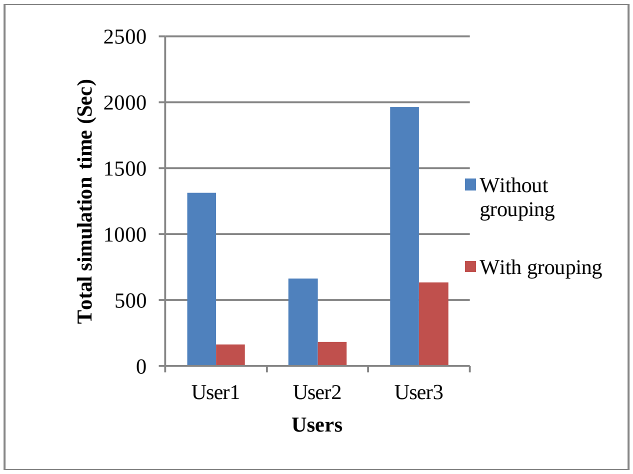
Fig. 1. Simulation time and processing cost using simulation inputs parameters SI1 in table 1 using UFF.



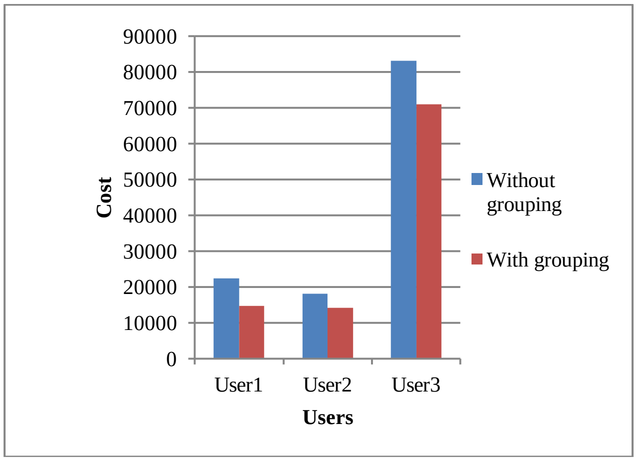
(a)

(b)

Fig. 2. Simulation time and processing cost using simulation input parameters SI1 in table 1 using URS.

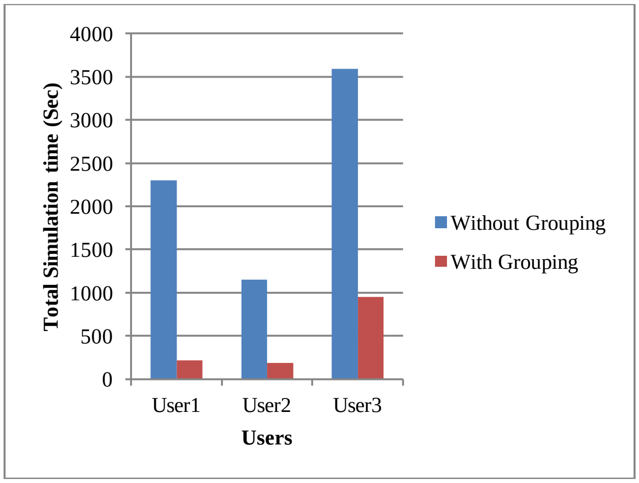


(a)

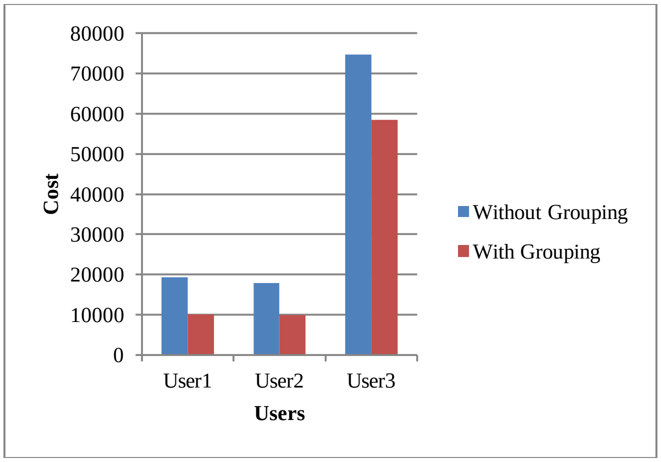


(b)

Fig. 3. Simulation time and processing cost using simulation inputs parameters SI1 in table 1 using RMQ.

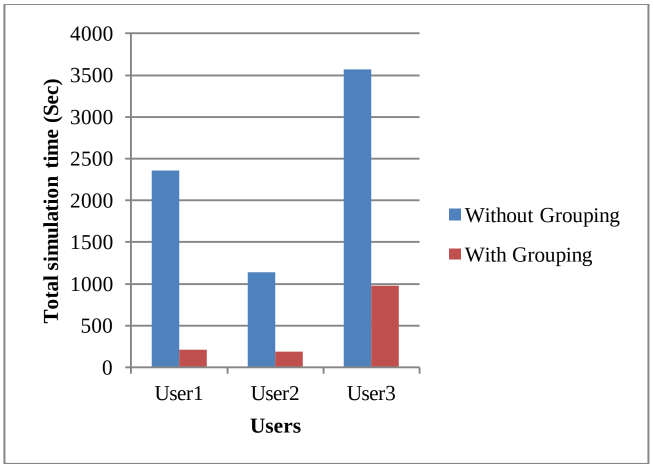


(a)

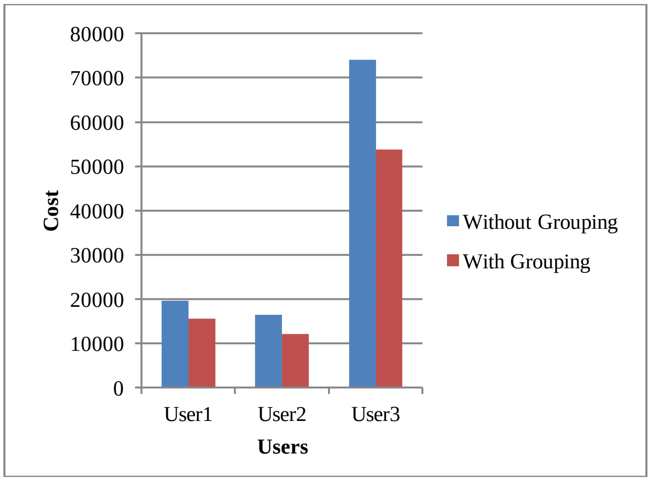


(b)

Fig. 4. Simulation time and processing cost using simulation input parameters SI2 in table 2 using UFF.

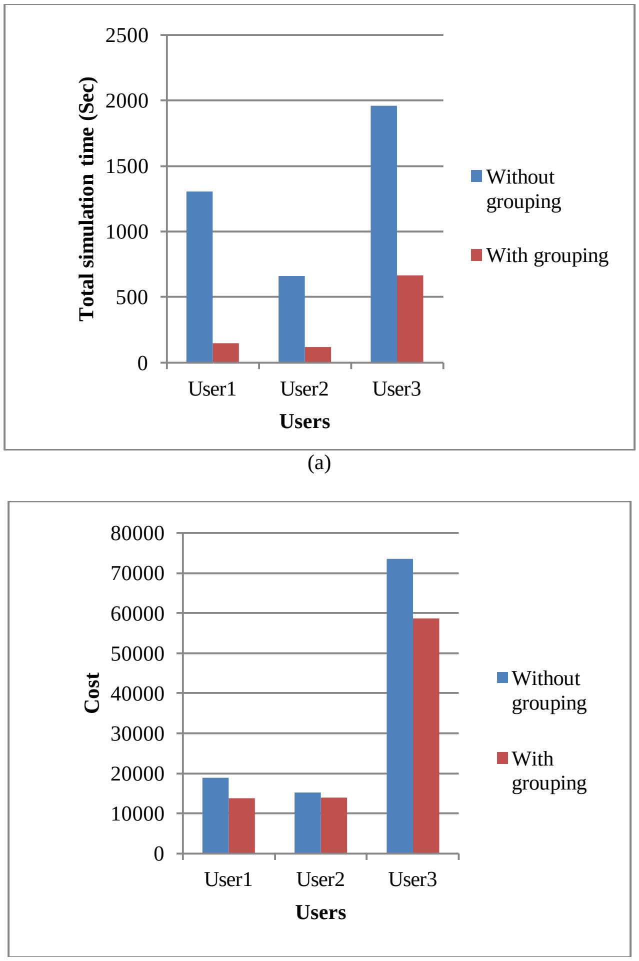


(a)



(b)

Fig. 5. Simulation time and processing cost using simulation input parameters SI2 in table 2 using URS



(a)

(b)

Fig. 6. Simulation time and processing cost using simulation input parameters SI2 in table 2 using RMQ.

The effect of varying the number of resources on the total simulation time is shown in Figures 7 and 8 using UFF and URS models. The simulation time in case of using sevenresources is less than the Simulation time when using four resources.

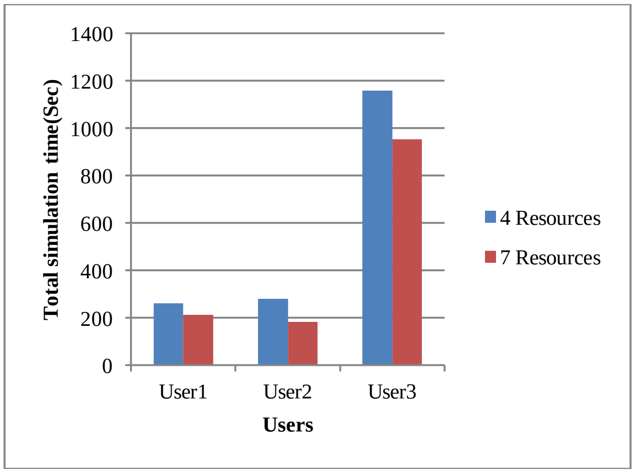


Fig. 7. Effect of varying the number of resources on total simulation time of jobs for different users using UFF.

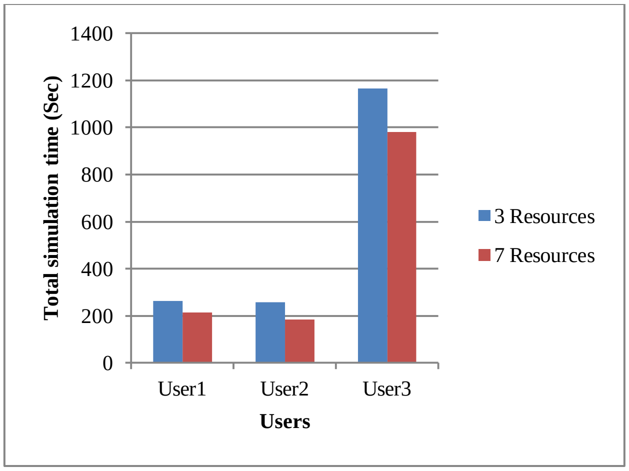


Fig. 8. Effect of varying the number of resources on total simulation time of jobs for different users using URS.

Figure 9 shows the results of total simulation time in case of using simulation input parameters SI3in table 3. The figure shows the simulation time of executing different number of gridlets of various users using 5 resources and different granularity time.Before starting simulation, thegiven granularity time is multiplied by the resource processing capability (MIPS). The result is the total (MI) that the resource can process within the given granularity time. Hence, higher granularity time means that the total (MI) that the resource can process will be also higher. Results shown are for experiments using simulation input parameters SI3 in table 3 and different values for granularity time: 5, 10, 15, and 20.The results show that the total simulation time for granularity time of 20 seconds is less than the total simulation time for granularity time of 15, 10 and 5 seconds.The 100 gridlets of user1 are grouped in three groups when granularity time is 5 seconds and are grouped in one group when granularity time is 10, 15, and 20. When the granularity time is equal to5, the product of granularity time and the resource's MIPS is equal tothe total MI that the resource can process within 5 seconds which is less than the total MI that the resource can process in 10, 15 and 20 seconds.When granularity time is less, more resources are needed to process the given gridlets within the same granularity time.

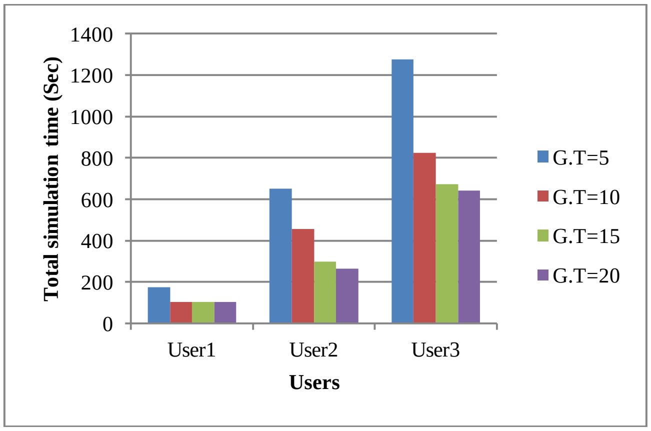


Fig. 9. Effect of varying the granularity time on total simulation time of jobs for different users usingsimulation input parameters SI3 in table 3.

Comparing the results of experiments with and without grouping, it is found that the total simulation time is reduced when grouping is applied reaching 9% to 33% of the total simulation time obtained without grouping.

The actual percentage depends on the number of gridlets and the processing requirements of each gridlet. Total cost in case of using grouping strategy is reduced reaching 52% to 91% of the total cost obtained without using grouping strategy.

Tables 4, 5 and 6 show the load distribution of gridlets on resources using the three proposed models.

TABLE IV. LOAD DISTRIBUTION ON GRID RESOURCES FOR USER1 USING SIMULATION INPUT PARAMETERS SI1 AND THE UFF SCHEDULING MODEL.

Grouped-Gridlet ID	Gridlets	Resource Name
0	0-32	R7
1	33-62	R5
2	63-81	R3
3	82-90	R1
4	91-99	R7

TABLE V. LOAD DISTRIBUTION ON GRID RESOURCES FOR USER1 USING SIMULATION INPUT PARAMETERS SI1 AND THE URS SCHEDULING MODEL.

Grouped-Gridlet ID	Gridlets	Resource Name
0	0-29	R1
1	30-38	R7
2	39-57	R3
3	58-88	R5
4	89-99	R7

TABLE VI. LOAD DISTRIBUTION ON GRID RESOURCES FOR USER1 USING SIMULATION INPUT PARAMETERS SI1 AND THE RMQ SCHEDULING MODEL.

Grouped-Gridlet ID	Gridlets	Resource Name
0	0-8	R1
1	9-39	R7
2	40-69	R5
3	70-89	R3
4	90-99	R1

V. CONCLUSION AND FUTURE WORK

Three models for scheduling fine-grained jobs in a grid computing environment are presented. First, the jobs are grouped to reduce the overall communication time incurred by sending individual jobs to grid resources. Then, grouped jobs are mapped to resources based on the proposed UFF, URS and RMQ scheduling models.

Our experiments were conducted using the Gridsim toolkit.Results indicate that total simulation time and cost are improved by grouping fine-grained jobs. Total simulation time with grouping is 9%-33% of that without grouping. Total cost in case of grouping reaches 52%-91% of that without grouping.

Furthermore, grouping enhancesutilization of resources processing capability. The grouping strategy is based on both the processing requirements of individual jobs and the processing capability of resources.

A load balancing scheduling model is also presented. The queue length of a resource is taken into consideration when a resource is selected. Granularity time has been used to indicate the number of gridlets that can be processed by a resource within a specified time. Results show that the total simulation time decreases as granularity time increases.

Future workin this area will be directed towards developing a grouping model based onthe bandwidth of resources in addition to their processing capability.

REFERENCES

[1] Er.Sourabh Budhiraja, “A Dynamic Load Balancing Approach in Grid Environment”, International Journal of Engineering Research and Technology (IJERT), Vol.1, Issue 9, November 2012.

[2] Dinesh S.Gawande, Rajesh C.Dharmik and Chanda Panse, “A Load Balancing in Grid Environment”, International Journal of Engineering Research and Applications (IJERA), Vol.2, Issue 2, PP.445-450, March-April 2012.

[3] Simrat Kaur and Sarbjeet Singh, “Comparative Analysis of Job Grouping based Scheduling Strategies in Grid Computing”, International Journal of Computer Applications, Vol.43, No.15, April 2012.

[4] P.Suresh and P.Balasubramanie, “Grouping Based User Demand Aware Job Scheduling Approach For Computational Grid”, International Journal of Engineering Science and Technology, Vol.4, No.12, December 2012.

[5] Manoj Kumar Mishra, Prithviraj Mohanty, G.B.Mund, “A Time-minimization Dynamic Job Grouping-based Scheduling in Grid Computing”, International Journal of Computer Applications, Vol.40, No.16, February 2012.

[6] Manpreet Singh, Sandip Kumar Goyal and Vishal Gupta, “An Adaptive Load Balancing Algorithm for Computational Grid”, Journal of Engineering and Technology, Vol.1, Issue 2, July-December 2011.

[7] S.Gomathi and D.Manimegalai, “An Analysis of MIPS Group Based Job Scheduling Algorithm with other Algorithms in Grid Computing”, International Journal of Computer Science Issues (IJCSI), Vol.8, Issue 6, No.3, November 2011.

[8] Simarjit Kaur and Sukhjit Singh, “Role Based Access Control For Grid Environment Using Gridsim”, Journal of Engineering Research and Studies (JERS), Vol. I, Issue I, PP.111-117, July-September 2010.

[9] Raksha Sharma, Vishnu Kant Soni, Manoj Kumar Mishra, Prachet Bhuyan and Utpal Chandra Dey, "An Agent Based Dynamic Resource Scheduling Model with FCFS-Job Grouping Strategy in Grid Computing", World Academy of Science, Engineering and Technology Journal, vol. 64, PP.467-471, 2010.

[10] Vishnu Kant Soni, Raksha Sharma, and Manoj Kumar Mishra, "Grouping-Based Job Scheduling Model In Grid Computing", World Academy of Science, Engineering and Technology Journal, vol. 65, PP.781-784, 2010.

[11] Yeqing Liao and Quan Liu, “Research on Fine-grained Job Scheduling in Grid Computing”, Modern Education and Computer Science journal, PP.9-16, 2009.

[12] T.F Ang, W.K.Ng, T.C Ling, L.Y. Por and C.S. Liew, “A Bandwidth-Aware Job Grouping-Based Scheduling on Grid Environment” Information Technology Journal, Vol.8, NO.3, pp. 372-377, 2009.

[13] Ng Wai Keat, Ang Tan Fong, Ling Teck Chaw and Liew Chee Sun "Scheduling Framework For Bandwidth-Aware Job Grouping-Based Scheduling In Grid Computing", Malaysian Journal of Computer Science, Vol.19, No. 2, pp. 117-126, 2006.

[14] Nithiapidary Muthuvelu, Junyang Liu, Nay Lin Soe, Srikumar Venugopal, Anthony Sulistio and Rajkumar Buyya, "A Dynamic Job Grouping-Based Scheduling for Deploying Applications with Fine-Grained Tasks on Global Grids", in Proc of Australasian workshop on grid computing, vol. 4, pp. 41–48, 2005.

[15] Vishnu Kant Soni, Raksha Sharma, Manoj Kumar Mishra and Sarita Das, "Constraint-Based Job and Resource scheduling in Grid Computing", 3rd International Conference On Computer Science and Information Technology, IEEE, 2010.

AUTHORS PROFILE

Rabab Mohamed Ezzat is currently a Masters Student at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from Helwan University, Cairo, Egypt. She worked as Teaching Assistant in Modern Sciences and Arts University in Egypt for three years. Her current research interests include parallel computing, computer networks and human computer interaction.

Amal Elsayed Aboutabl is currently an Assistant Professor at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA. Her current research interests include parallel computing, image processing and software engineering.

MostafaSami M. Mostafais currently a Professor of computer science, Faculty of Computers and Information, Helwan University, Cairo, Egypt. He worked as an Ex-Dean of faculty of Computers and Information Technology, MUST, Cairo. He worked also as an Ex-Dean of student affairs and Ex-Head of Computer Science Department, faculty of Computers and Information, Helwan University, Cairo, Egypt. He is a Computer Engineer graduated 1967, MTC, Cairo, Egypt. He received his MSC 1977 and his PhD 1980 from University of Paul Sabatier, Toulouse, France. His research activities are in Software Engineering and Computer Networking. He is awarded supervising more than 80 Masters of Sc. and 18 PhDs in system modeling and design, software testing, middleware system development, real-time systems, computer graphics and animation, virtual reality, network security, wireless sensor networks and biomedical engineering.

Performance Evaluation for the DIPDAM schemeon the OLSR and the AODV MANETs Routing Protocols

Ahmad Almazeed ,Ahmed Mohamed Abdalla
Electronics Department, College of Technological Studies,
The Public Authority for Applied Education and Training ,
P.O.Box 42325,Shuwaikh 70654, Kuwait

Abstract—**the DIPDAM scheme is a fully-distributed message exchange framework designed to overcome the challenges caused by the decentralized and dynamic characteristics of mobile ad-hoc networks. The DIPDAM mechanism is based on three parts Path Validation Message (PVM) enables E2E feedback loop between the source and the destination, Attacker Finder Message (AFM) to detect attacker node through the routing path, and Attacker Isolation Message (AIM) to isolate the attacker from routing path and update the black list for each node then trigger to neighbors with updated information. The DIPDAM scheme was fully tested on the OLSR routing protocol. In order to prove the efficiency of DIPDAM scheme on detection and isolation packet dropping attackers, DIPDAM is applied to another routing protocol category, AODV. AODV represents different concepts in routing path calculation and it is widely adopted. The comparison between the two routing protocol is tested onsmart attackers. The goal from this comparison is to prove that the DIPDAM scheme can be applied to a different routing protocols category.**

Keywords—Ad hoc networks; AODV; Computer network management; IDS;MANETS; OLSR

I. Introduction

A mobile ad-hoc network (MANET) is categorized under infrastructure less network where a number of mobile nodes communicate with each other without any fixed infrastructure between them. Furthermore, all the transmission links are established through wireless medium [1].

The DIPDAM scheme [2, 3, 4] is a fully-distributed message exchange framework designed to overcome the challenges caused by the decentralized and dynamic characteristics of MANETs.

The collaboration of a group of neighbor nodes is used to make accurate decisions. Eliminating misbehavior node(s) enables the source to select another trusted path to its destination. In order to lower message exchange overhead aswell as to achieve scalability, message exchange is triggered only when new detection is observed, and only occurs with local neighbors.

DIPD AM scheme enables routing protocols to detect packet dropping frauds. In fact, source nodes in the network independently monitor the behavior of their own data when transferring through routing path, however, they need to collaborate in order to identify and isolate the intruders. This scheme is based on the reputation concept.

In this paper the DIPDAM scheme is tested on two different MANETs routing protocols, the OLSR and the AODV. The scheme is evaluated using four different performance metrics. Furthermore, the detection accuracy and false positive rate are calculated for the two routing protocols.

II. Previous Work

For Mobile Ad-hoc Networks, the general function of anIntrusion Detection Systems (IDS) is detecting misbehaviors by observing the networks traffic in a Mobile Ad-hoc [5]. Most of recent researches focused on providing preventive schemes to secure routing in MANETs [6-10]. Key distribution and an establishment of a line of defense defined in [6], [6] based on mechanism in which nodes are either trusted or not and if trusted they are not compromised. Also contribution in [8], [10] considers the compromise of trusted nodes. It assumed a public key infrastructure (PKI) and a timestamp algorithm are in place. However, the above approaches cannot prevent attacks from a node who owns a legitimate key.

It is necessary to understand how malicious nodes can attack the MANETs. A model to address the Black Hole Search problem algorithm and the number of agents that are necessary to locate the black hole without the knowledge of incoming link developed in [11]. In [12] a survey of different network layer attacks on MANET was provided and compared the existing solutions to combat single or cooperative black hole attack.

A feedback mechanism to secure OLSR against the link spoofing attacks was provided in [13, 14]. The solution assesses the integrity of control messages by correlating local routing data with additional feedback messages called CPM sent by the receivers of the control messages.

The proactive protocols are Table-Driven protocols in which each node maintains up-to-date routing information about every other node in a routing table and routes are quickly established without any delay [15].

Researchers in [16, 17] describes an explicit security issue on AODV Routing Protocol Suffering from Black Hole Attack. Source node sends the routing information to the nasty node which essentially cannot have a path to destination node in its own routing table. It thinks that fake route reply and it ignores the message without passing to destination. Authors also include the exact method to overcome the Black Hole Attack by providing a new method called Secured AODV (SAODV). It provides an additional procedure to AODV algorithm by requesting source node to broadcast the Secured Route Request along with random sequence number to destination. Destination checks whether source request sequence number from two or more path are the same.

III. COMPARING AODV AND OLSR PROTOCOLS

AODV and OLSR protocols are compared with respect to resource usage, mobility, and route discovery delay. Being a proactive protocol, OLSR imposes large control traffic overhead on the network. Maintaining up-to-date routing table for the entire network calls for excessive communication between the nodes, as periodic and triggered updates are flooded throughout the network. The use of MPR's reduces this control traffic overhead, but for small networks, the gain is minimal. The traffic overhead also consumes bandwidth. The creativeness of AODV is more sensitive to resource usage than OLSR. As control traffic is only emitted during route discovery, most of the resource and bandwidth consumption is related to actual data traffic.

A. Resource usage

Since information about the entire network needs to be maintained at all times, OLSR requires relatively much storage complexity and usage. Hence, there is a greater demand for storage capacity of nodes in such networks.

Also, the control overhead adds to the necessary processing in each node, hence increasing the battery depletion time. Another downside to OLSR is that it must maintain information about routes that may never be used.

AODV, on the other hand, only stores information about active routes at a node, which considerably simplifies the storage complexity and reduces energy consumption. The processing overhead is also less than OLSR, as little or no useless routing information is maintained.

B. Mobility

OLSR and AODV have different strengths and weaknesses when it comes to node mobility in MANETs. Unlike wired networks, the topology in wireless ad-hoc networks may be highly dynamic, causing frequent path breaks to ongoing sessions. When a path break occurs, new routes need to be found. As OLSR always have up-to-date topology information at hand, new routes can be calculated immediately when a path break is reported. In comparison, since AODV is a reactive protocol, this immediate new route calculation is not possible, so a route discovery must be initiated. In situations where the network traffic is sporadic, OLSR offers less routing overhead due to having found the routes proactively. AODV, on the other hand, will need to discover a route before the actual information can be transmitted. This calls for extensive control overhead per packet. In cases where the network traffic is more or less static (i.e., the traffic has a long duration), however, AODV may perform better, as the amount of control overhead per packet decreases.

TABLE I. AODV VS. OLSR ROUTING PROTOCOLSCOMPARISON.

Parameters	AODV routing protocols	OLSR routing protocols
Availability of routing	Available as required	Always available
Periodic route updates	Not required	Required
Dealing with Link	Use route discovery	Propagate information to neighbors to maintain consistent routing table
Routing overload	Increases with mobility of nodes	Independent of traffic and mostly greater than On-demand protocols

C. Route discovery delay

When a node in a network running the OLSR protocol needs to find the route to a host, it is only required to do a routing table lookup, whereas in a AODV network, a route discovery process need to be initialized unless no valid route is cached.

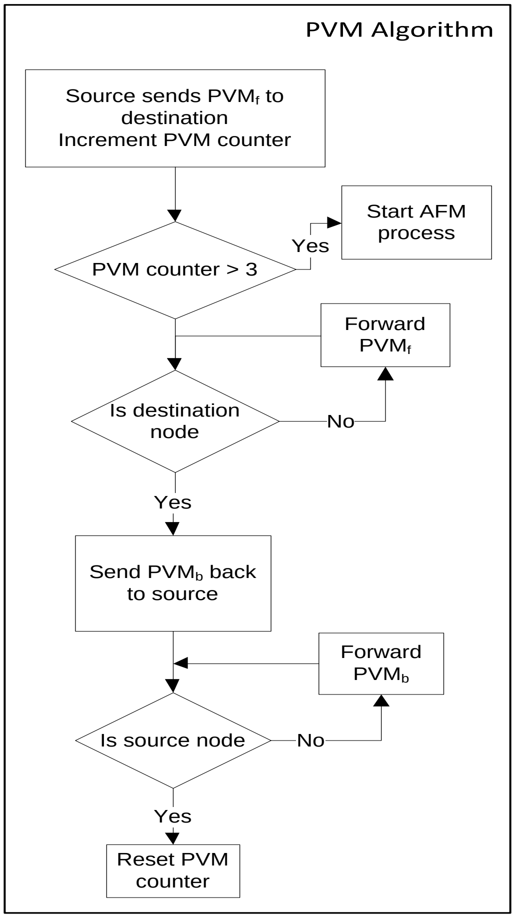


Fig. 1. Flow chart for Path Validation Message (PVM) algorithm

```
Source sends $AFM_f$ to Destination and starts a waiting time
If receiver node = destination then
    Send $AFM_b$ back to source
Else
    Forward $AFM_f$ to destination
    Send $AFM_b$ back to Source with information about
        next-node-to-destination(NNTD) and availability of route
    to destination in the routing table
End if
If Source received $AFM_b$ came from Destination then
    No attacker detected, start advanced detection
    Cancel AFM wait timer
    Send PVM to each node in path to D
        If Source receive PVM from intermediate node then
            Node is trusted
    Else
            Malicious node of type-N2 is detected.
            Add to blacklist table and end AFM process
    End if
Else
    Last NNTD known by S is suspected as type-N1 attacker
    Send PVM to NNTD
    If PVM received then
            NNTD is a trusted node
    Else
            NNTD is confirmed as an attacker
    End if
End if
```

Fig. 2. Attacker Finder Message (AFM) algorithm.

It goes without saying that a table-lookup takes less time than flooding the network, making the OLSR protocol performance better in delay-sensitive networks. Table 1 summarizes basic differences between the two protocols classes.

IV. Detection And Isolation Of Packet Dropped Attackers In Manets (DIPDAM)

New solutions for detecting data packet dropping in ad-hoc networks work by monitoring individual nodes. Other solutions used so far for protecting these networks are authentication and encryption [18]. Most of these mechanisms are not considerably appropriate for MANETs resource constraints, i.e., bandwidth limitation and battery power, since they result in heavy traffic load for exchanging and verification of keys.

In DIPDAM scheme, each source node in the network monitors its own packets (data packets or routing packets)by using a Path Validation Message (PVM) as shown in Fig. 1. If a misbehavior node is detected, the other neighboring nodes are informed in order to help them in protecting themselves. Each source node monitors the behavior of its neighborhood instead of making each node in the networking doing this job which consumes nodes resources.

A failure to get a reply for an N PVM messages sent (N is set to 3 in the flow chart), DIPDAM algorithm will trigger an Attacker Finder Message (AFM) algorithm shown in Fig. 2.

The detector node needs to share the information about the detected attacker with other nodes in the network. This is accomplished by flooding the network with Attacker Isolation Messages (AIMs) [2].It is noticed that nodes can be incorrectly detected as attackers due to network malfunction during a certain period. Such nodes would be wrongly isolated for the lifetime of the whole network. A verification step is added to ensure that nodes are correctly detected and isolated. The process is illustrated in Fig. 3. Fig. 4 shows a flow chart for the AIM algorithm.

To evaluate the robustness of DIPDAM scheme we tested MANETs under different attacker types [19].

N1 nodes take contribution in the route discovery and route maintenance processes but refuses to forward data packets to protect its resources. This attack type can reduce network throughput, but does not affect any of the network traffic unless it is routed through selfish nodes, selfish nodes refuse to forward or drop data packets, this attacker type will be named as smart attacker.

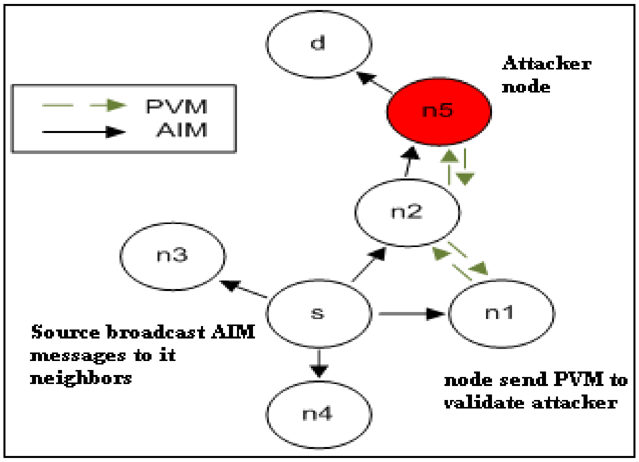

d
Attacker node
PVM
AIM
n5
n2
n3
s
n1
n4
Source broadcast AIM messages to it neighbors
node send PVM to validate attacker

Fig. 3. Attacker Isolation Message (AIM) process.

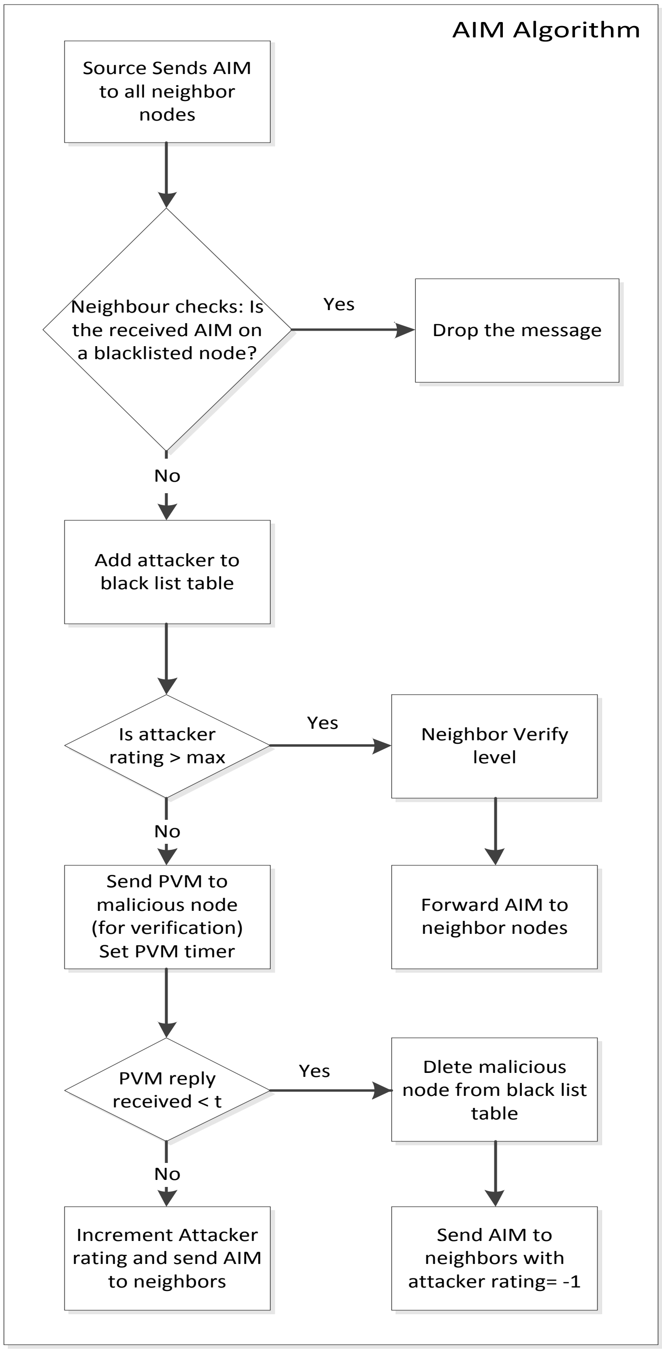

AIM Algorithm
Source Sends AIM to all neighbor nodes
Neighbour checks: Is the received AIM on a blacklisted node?
Yes
Drop the message
No
Add attacker to black list table
Is attacker rating > max
Yes
Neighbor Verify level
No
Send PVM to malicious node (for verification) Set PVM timer
Forward AIM to neighbor nodes
PVM reply received < t
Yes
Dlete malicious node from black list table
No
Increment Attacker rating and send AIM to neighbors
Send AIM to neighbors with attacker rating= -1

Fig. 4. Flow chart for AIM algorithm.

N2 nodes neither contribute to the route discovery processes nor data-forwarding processes. Instead they use their resources only for transmissions of their own packets which are called selfish nodes. An attacker with this criterion will be named normal attacker.

N3 nodes behave properly if its energy level lies between full energy-level and certain threshold T1. They behave like node of type N2 if energy level lies between threshold T1 and another threshold T2 and if energy level falls below T2, they behave like node of type N1.

N1, N2, and N3 nodes are risky to routing protocols. These nodes suspend the data flow by either dropping or refusing to forward the data packets thus forcing routing protocol to select an alternative available route which it may again contain some malicious nodes, resulting in the new route also to fail. This process form a loop which enforce source to conclude that data cannot be further transferred.

The proposed work is designed to detect and isolate N1typeand N2 type. N3type selfish nodes will be detected only when they behave similar to N1or N2type nodes.

Dropping any packets affects the network performance by causing the retransmission of data packets many times. Furthermore, it can prevent the end-to-end communications between nodes.

Network Simulator program

The NS-2 simulation tool [20-21] consists of two kinds of scenarios; topology scenario and traffic generation pattern. The topology scenario defines the simulation area and the mobility model of randomly distributed mobile nodes over the simulation time. The traffic pattern defines the characteristics of data communications, data packet size, packet type, packet transmission rate and number of traffic flows. Each node is assumed to be equipped with a wireless transceiver operating on 802.11 wireless standards. The physical radio frequency characteristics of each wireless transceiver such as transmit power, the antenna gain, and signal to noise and interference ratio, are chosen with a bit rate of 2Mb/sec and a transmission range of 250 meters with an omni-directional antenna.

The simulation scenarios consist of two different settings. First, the impact of network density or size is assessed by varying the number of mobile nodes placed on an area of a fixed size of 1500m x 300m. The second simulation scenario investigates the effects of node mobility on the performance of route discovery by varying the maximum speed of mobile nodes placed on a fixed area of 1500m x 300m.

Each node participating in the network is transmitting within the 250m transmission range, and each simulation runs for a period of 900sec. The above settings could represent a MANET scenario in real life; like a University campus. Note that the number of mobile nodes could be larger than the one presented in these scenarios and the operational time could be longer; the values chosen are to keep the simulation running time manageable while still generating enough traces for analysis. Flows of Constant Bit Rate (CBR) unicast data packets, each with size 512 bytes.

In this study, mobile nodes move according to the widely used random way point mobility model where each node at the beginning of the simulation remains stationary for pause time seconds, then chooses a random destination and starts moving towards it with a speed selected from a uniform distribution [0, V max].Other simulation parameters used in this research study have been widely adopted in existing performance evaluation studies of MANETs and are summarized below in Table 2.

TABLE II. SYSTEM PARAMETERS USED IN THE SIMULATION EXPERIMENTS.

Simulation Parameter	Value
Simulator	NS-2 (v.2.31)
Transmitter range	250 meter
Bandwidth	2 Mbps
Traffic type	CBR
Number of Nodes	30
Topology size	1500m x 300m
Packet size	512 bytes
Simulation time	900 sec

V. PERFORMANCE METRICS

In order to evaluate the performance of our proposed Intrusion Detection System DIPDAM, we will focus mainly on evaluating four performance metrics:-

Average overhead:

The average overhead is defined as the total number of data packet and routing control packets normalized by the total number of received data packets.

Average Packet Delivery Ratio (Rating):

It is the ratio of the number of packets received successfully to the total number of packets transmitted. Average Packet dropping:

The average packet dropping is the average percentage of data packet dropped to all data and control packets sent from the sources to the destinations.

Average end-to-end delay:

The end-to-end-delay is the average overall delay measured from the sources to the destinations.

A. Percentage of Overhead

The first performance metric used in comparison is the percentage of average overhead. Fig..5illustrates the percentage of average overhead in both routing protocol (OLSR & AODV) versus the number of attackers.

From Fig.5, it is clear that when the attacker numbers is relatively small AODV protocol achieve better average overhead than OLSR.

Increasing the number of attackers leads to an increasing in the average overhead in AODV, with the rate higher than OLSR. When the number of attackers is increased more, the OLSR achieves better percentage of average overhead than AODV.

The increase of the attacker numbers leads to the increase of lost links, then AODV will produce more control messages (like RREQ and RREP). These control messages will be broadcasted throughout the network nodes to create an alternative routing path causing the overall overhead to increase rapidly. These results are expected as OLSR is more stable than AODV and it is less affected with network changes than AODV.

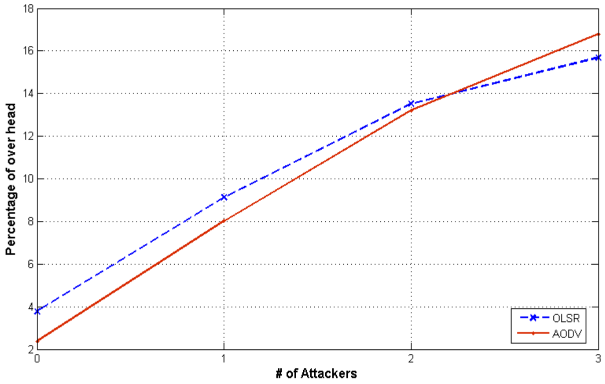


Fig. 5. Percentage of overhead vs. number of attackers.

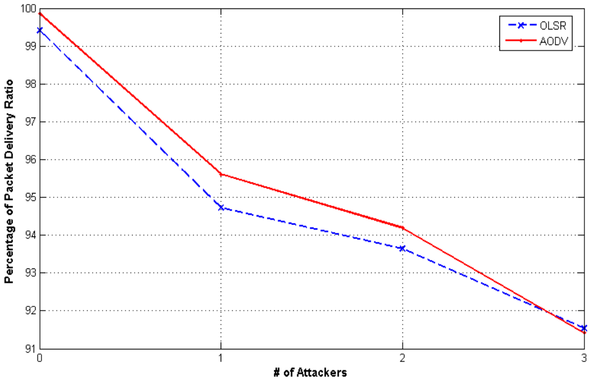


Fig. 6. Percentage ofPacket Delivery Ratio vs. number of attackers.

B. Percentage of Packet Delivery Ratio

The average packet delivery ratio performance metric in both routing protocol (OLSR & AODV) is showed in Fig..6. The results presented in the figure show that the AODV routing protocol achieves better average packet delivery ratio than OLSR routing protocol, especially when the number of attackers are relatively small. As the number of attackers increase, the average packet delivery ratio in AODV decreases at a rate higher than OLSR. When a certain number of attackers is reached (about 10% from the total nodes) the OLSR will perform better than the AODV.

The average packet delivery ratio in OLSR is slightly higher than that in AODV when the number of attackers is large. AODV needs to recalculate the routing path because the routing path expires if it is not used for a certain time or if the path is broken. During the recalculating process, the source node will not be able to send its data. The higher the number of attackers makes the recalculation process take more time, which affects the average packet delivery ratio.

C. Percentage of Dropped packets

The percentage of average dropped packet performance parameter in both routing protocols (OLSR & AODV) is plotted against the number of attackers as shown in Fig.7.

Fig. 7 results show that the value of the percentage of average dropped packets recorded is remarkably small when no attacker is found in the networks. The percentage of average dropped packets in the OLSR protocol increaseslinearly with the number of attacker, but the increasing is nonlinear in the AODV protocol .

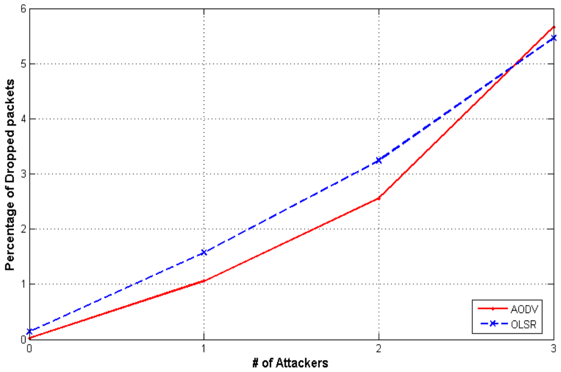


Fig. 7. Percentage of Dropped packets vs. number of attackers.

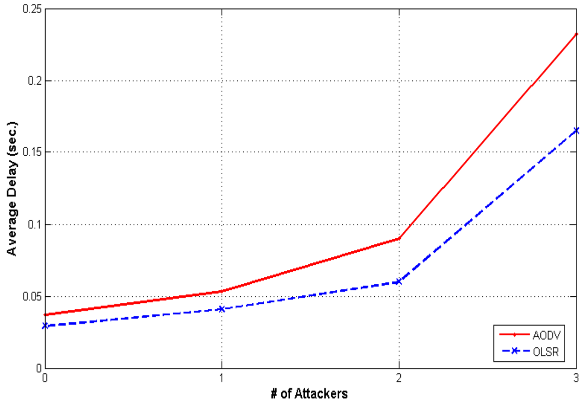


Fig. 8. Average End to End delay (sec.) vs. number of attackers.

Fig.7 results show that the value of the percentage of average dropped packets recorded is remarkably small when no attacker is found in the networks. The percentage of average dropped packets in the OLSR protocol increases linearly with the number of attacker, but the increasing is nonlinear in the AODV protocol .

The results obtained from Fig.7 show that the percentage of dropped packets in AODV is always less than OLSR when the number of attackers is relatively less than 10% of the total network nodes .

When the number of attackers approaches nearly 10% of the total nodes, the ADOV averaged dropped packet value exceeds the OLSR value. The greater the number of attackers leads to greater lost links. The recovery time in AODV is slower than OLSR because OLSR protocol maintains its routing paths periodically while AODV recalculates its path when the source needs to send data. The recalculation process requires more computational process and time.

D. Average End-to-End delay

The fourth performance metric measured is the average end to end delay. The results against the number of attackers in both routing protocol is shown in Fig. 8.

It is noticed from Fig.8 that the values of average End-to-End delay produced by OLSR protocol is always less than the values of AODV. The routing path in OLSR is always available irrespective of the source needed to transmit data or not. AODV calculates the routing path only if the source needs to send data to its destination. The data remains waiting until the routing path calculation is completed, and then the data is forwarded to its destination. That led to the OLSR achieving a better average End-to-End delay than AODV. Since lost links in AODV need extra computational time to recalculate the routing path, the End-to-End delay in OLSR was less than AODV when the number of attackers becomes larger. In OLSR, the routing path is always ready, and there is no need to calculate it.

TABLE III. DETECTION RATE AND FALSE POSITIVE RATE

Routing Protocols	Detection rate	False Positive rate
OLSR	99.42 ±0.5%	1.1±0.01%
AODV	98.96±0.5%	1.21±0.012%

VI. DETECTION ACCURACY AND FALSE POSITIVE VALIDATION TESTS

To validate the DIPDAM scheme two more factors are measured for OLSR and AODV routing protocols: detection accuracy and false positive rate are calculated. Experimental results showed that DIPDAM in both OLSR and AODV achieved high performance with remarkably low false positives, and very high detection rate in any environment with high mobility, as shown in Table 3.

VII. DISCUSSION

From the performance metrics figures, it is obvious that the DIPDAM scheme can be considered as an effective scheme to detect and isolate any number of attackers from routing paths, irrespective of the routing protocol type.

AODV routing protocol achieved better performance metrics when the number of attackers is relatively small to the network size. On the other hand the OLSR seemed to be a more stable routing protocol in larger networks and achieved better performance than AODV, especially when the number of attackers was large.

It is clear that the AODV is more flexible for security solutions than the OLSR in small networks. Performance metrics of AODV protocol highly depends on the number of attackers, but OLSR protocol keeps the network performance the same, irrespective of the number of attackers.

DIPDAM performed efficiently with the validation tests performed. The scheme achieved high detection ratewith impressive low false positives on both the OLSR and the AODV protocols.

VIII. CONCLUSION

DIPDAM has been successfully implemented in OLSR and AODV. Experimental results show that DIPDAM in both

OLSR and AODV has low message overhead and low detection delay. This achieves higher performance with remarkably low false positives, and remarkably high detection rate in an environment with high mobility. Also, DIPAM proved to be a practical, scalable, and effective solution for securing both OLSR and AODV.

The simulation results showed that DIPDAM scheme was able to detect and isolate any number of attackers, while keeping a reasonably low overhead in terms of network traffic. The four performance metrics of the experiment demonstrate that the DIPDAM system can detect packet dropping attacks in both routing protocols (OLSR and AODV) with low message overhead, low detection delay, high rating under message loss and mobility conditions.

According to the simulation results, AODV protocol will perform better in networks with static traffic and relatively small numbers of attackers for the same network size of OLSR.AODV uses lower resources than OLSR, because the control message size used in AODV is kept small and requires smaller bandwidth for maintaining the routes. The AODV routing protocol maybe used in resource critical environments.

IX. Future Work

DIPDAM scheme must be tested in real MANETs with different conditions like variation on mobility, size, network traffic type, and node density.

The same scheme can be tried on different MANETs protocols from other categories, like multicast protocols. DIPDAM scheme can be upgraded to detect both types of attackers, data packet attackers and routing packets attackers.

References

[1] F.Tseng, L. Chou, and H. Chou. “A survey of Black Hole Attacks in wireless mobile ad-hoc networks”, Human-centric Computing and Information Sciences, 2011

[2] A.Abdalla, I. Saroit, A. Kotb, and A.Afsari. “An IDS for Detecting Misbehavior Nodes in Optimized Link State Routing Protocol”, International Journal of Advanced Computer Science, 1 (2), pp. 87-91, 2011

[3] A. Abdalla, I. Saroit, A. Kotb, and A. Afsari. “Misbehavior Nodes Detection and Isolation for MANETs OLSR Protocol”, World Conference on Information Technology. Procedia Computer Science,3, pp. 115–121, 2011

[4] A. Abdalla, A.Almazeed, I. Saroit, A. Kotb, and A.Afsari. “Detection and Isolation of Packet Dropping Attacker in MANETs”, International Journal of Advanced Computer Science and Application, 4 (4), pp. 29-34, 2013

[5] A. Fourati, and K. AlAghha.“An IDS First Line of defense for Ad Hoc Networks”, in Proceeding of IEEE WCNC, 2007

[6] Y. Hu, A. Perrig, and D. Johnson.“Ariadne: A secure On-Demand Routing Protocol for Ad hoc Networks”, Proc. of the MobiCom, Atlanta, Georgia, USA, 2002.

[7] C. Adjih, T. Clausen, P. Jacquet, A. Laouiti, P. Muhlethaler, and D. Raffo. “Securing the OLSR protocol”, Proc. of Med-Hoc-Net, Mahdia, Tunisia, June 25, 2003.

[8] D. Dhillon, T.Randhawa, M. Wang and L. Lamont. “Implementing a Fully Distributed Certificate Authority in an OLSR MANET”, IEEE WCNC2004, Atlanta, Georgia USA, 2004.

[9] D. Raffo, C. Adjih, T. Clausen, and P. Muhlethaler.“An Advanced Signature System for OLSR”, Proc. of the 2004 ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN 04), Washington, DC, USA, 2004.

[10] C. Adjih, D. Raffo, and P. Muhlethaler. “Attacks Against OLSR: Distributed Key Management for Security”, 2nd OLSR Interop/Workshop, Palaiseau, France, 2005.

[11] P. Glaus. “Locating a Black Hole without the Knowledge of Incoming Link, Algorithmic Aspects of Wireless Sensor Networks”, Lecture Notes in Computer Science, Vol. 5304. 128, 2009.

[12] N.Kaushik, and A.Dureja.“A comparative study of black hole attack in MANET” , International Journal of Electronics and Communication Engineering & Technology (IJECET) 4(2), 2013.

[13] J.Vilela and J. Barros. “A Feed Reputation Mechanism to Secure the Optimized Link State Routing Protocol”, The 3rd IEEE/CreateNet International Conference on Security and Privacy in Communication Networks, Nice, France, 2007.

[14] J.Vilela and J. Barros.“A Cooperative Security Scheme for Optimized Link State Routing in Mobile Ad-hoc Networks”, Proc. of the 15th IST Mobile and Wireless Communications Summit, Mykonos, Greece, 2006.

[15] P.Jaiswal, and R. Kumar.“Prevention of Black Hole Attack in MANET”, International Journal of Computer Networks and Wireless Communications (IJCNWC), 2(5), 2012.

[16] K. Lakshmi, S. Manjupriya, A. JeevaRathinam, K. Ram, and K. Thilagam. “Modified AODV Protocol against Black hole Attacks in MANET”, Proc, on International Journal of Engineering and Technology, 2(6), pp. 444-449, 2010.

[17] K.Taneja, and M. Rachna.“Security Issue on AODV Routing Protocol Suffering From Black holeAttack”. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE),1(7), 2012.

[18] Y. Rebahi, V. Mujica, C. Simons, and D. Sisalem. “SAFE: Securing packet Forwarding in ad-hoc networks”, In 5th Workshop on Applications and Services in Wireless Networks, 2005.

[19] S. Sen. “Evolutionary Computation Techniques for Intrusion Detection in Mobile Ad Hoc Networks”, doctoral diss., University of York Department of Computer Science, 2010.

[20] The Vint Project, “The Network Simulator –ns-2”. March 2005,www.isi.edu/nsnam/ns/index.html.

[21] F. J. Ro. (2005).“UM-OLSR Documentation”, University of Murcia. March 2005, http://masimum.dif.um.es/um-olsr/html

Control of Single Axis Magnetic Levitation System Using Fuzzy Logic Control

Tania Tariq Salim
Electrical &Electronics Engineering Department University of Gaziantep, 27310, Gaziantep,Turkey

Vedat Mehmet Karsli
Electrical &Electronics Engineering Department Universityof Gaziantep, 27310, Gaziantep,Turkey

***Abstract*—This paper presents a fuzzy logic controller design for the stabilization of magnetic levitation system (Maglev 's).Additionally, the investigation on Linear Quadratic Regulator Controller (LQRC) also mentioned here. This paper presents the difference between the performance of fuzzy logic control (FLC) and LQRC for the same linear model of magnetic levitation system .A magnetic levitation is a nonlinear unstable system and the fuzzy logic controller brings the magnetic levitation system to a stable region by keeping a magnetic ball suspended in the air. The modeling of the system is simulated using Matlab Simulink and connected to Hilink platform and the maglev model of ZeltomCompany. This paper presents a comparison for both LQRC and FLC to control a ball suspended on the air. The performance results of simulation shows that the fuzzy logic controller had better performance than the LQR control.**

***Keywords*—Magnetic Levitation; System MAGLEV; Fuzzy Control;Linear Quadratic Regulator Controller**

I. INTRODUCTION

Maglevsystems are getting huge interest from the designers over the world because of their different applications such as high-speed Maglev passenger trains, frictionless bearing, spacecraft etc. The controller design for any system is based on good understanding of the plant. This make the classical controller not satisfy the required control performance for a complex system like electromagnetic suspension system (EMS) because it is typically unstable, nonlinear and time-varying system. However there are a great number of plants already controlled by conventional PID controllers (or its derivatives) or optimal controlling

LQRC with acceptable results [1]. In recent years fuzzy controller is successfully applied to solve complex engineering problems without exact mathematical plant model [2]. Fuzzy logic control is very flexible controller with any system [3]. No controller design change will be required if any change acquire within the system only adding some new function to the controller design will be sufficient [4].This makes the fuzzy logic controller to be a good choice for controlling nonlinear systems such as electromagnetic levitation systems. In recent years many design approaches introduced by researcher to solve problems of maglev system. These researchers have adopted some different measures to solve the problems of overshoot, system oscillation and the stability. They obtained reasonable results using fuzzy neural network or adaptive fuzzy control to the controlled plant [4-6]. Several comparisons between fuzzy logic control and other types of controller done by many researcher over the past several years, a lots of them compared between the performance of FLC and PID, PD controller, some combined the modern with classical method using fuzzy-PD control [1,7]. Kashif Ishaque1, Yasir Saleem2, S.S Abdullah1, M. Amjad1, Munaf Rashid1 and Suhail Kazi1 they apply a new approach for fuzzy logic control using single input –single out put which can be easy implemented comparing to the multi input single out put[8].

In this work Linear Quadratic Regulator control method selected for comparison to Fuzzy Logic control. LQRC is part of optimal control strategy which has been widely developed and used in various applications [9]. The system must be described by state space model so this type of controller related directly to linear model of the system reverse FL controller.

II. MAGLEV DYNAMIC MODEL

The Maglev model consists of an electromagnet, a levitating magnet ball and a hall effect sensor for measuring the position of the levitated ball Fig1 shows the system model:

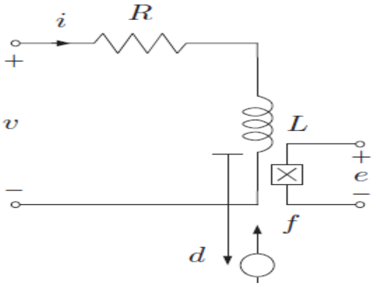


Fig. 1. Electromagnetic levitation system model

R: coil resistance

L: coil inductance

v: voltage across the coil

g: gravity constant

m: magnetic ball mass

d: distance measured from the coil bottom to the ball center

f : present the force on the levitating magnetic ball introduce by the electromagnet

e : is the voltage across the hall effect sensor

The electromagnetic force of the magnetic levitation system is described by the following nonlinear equation:

$$m * \frac{d^2 * x(t)}{dt^2} = m * g - f(x,i) (1)$$

The electromagnetic force on the levitated object is found using the concept of co-energy, the co- energy (w') is defined as:

$$w'(x,i) = \frac{1}{2} * i^2 * L(x) (2)$$

$$[f(x,i)]dx = dwe - dwf1 \ (3)$$

$$f(x,i) = \frac{dWe}{dx} - \frac{dWf1}{dx} (4)$$

Where dWe is change of electrical input and dWfl is change of stored energy, assuming dWe equals zero for linear system.

assumedWe=0

$$f(x,i) = -\frac{dwf}{dx} (5)$$

$$L(x) = L1 + \frac{L0 * x0}{x} (6)$$

where:

L(x) is the total inductance of the electromagnetic coil

X:is the position of the ball

L0: is additional inductance contributed by the presents of the object(magnet ball)

L1:is the inductance contributed without object presents (magnet ball)

X0:is equilibrium position

$$f(x,i) = -\frac{i^2}{2} * \frac{dL(x)}{dx} \qquad (7)$$

$$f(x,i) = -\frac{i^2}{2} * d\left(L1 + \frac{L0 * X0}{X}\right) (8)$$

$$f(x,i) = -\frac{i^2}{2} * \left(-\frac{L0 * X0}{x^2}\right) (9)$$

$$f(x,i) = \frac{L0 * X0}{2} * \frac{i^2}{x^2} (10)$$

$$f(x,i) = C * \frac{i^2}{x^2} (11)$$

where C=(L0*X0)/2

$$m\ddot{x} = m * g - C\left(\frac{i}{x}\right)^2 (12)$$

Taylor series expansion which was used to create a linear the equation:

$$f(x,i) = C * \left(\frac{i0}{x0}\right)^2 + \left[2 * C * \frac{i0^2}{x0^2}\right] * i(t) \left[2 * C * \frac{i0^2}{x0^3}\right] * x(t) (13)$$

*i*0 equals the current of the coil when the ball is at x0 (x0 equilibrium position). When the magnetic force balances the gravitational force on the ball, the acceleration of the ball is zero

$$f0 = C\left(\frac{i0}{x0}\right)^2 \qquad (14)$$

$$m * g = C\left(\frac{i0}{x0}\right)^2 (15)$$

Control force, f1 for keeping the ball balanced is given by the following equations:

$$f(i,x) = f - f0 (16)$$

$$f1 = \left(2C * \frac{i0}{x0^2}\right) i(t) - \left(2C * \frac{i0^2}{x0^3}\right) x(t) (17)$$

The voltage-current relationship for the coil is given by:

$$v(t) = R * i(t) + L1 * \frac{di(t)}{dt} (18)$$

The hall effect sensor equations used to measure the ball position is:

$$y = \beta * x (19)$$

$\beta$is the sensor gain.

The transfer function of the system is a ratio of the output to input in the Laplace domain. Laplace transformation of sensor and electrical and mechanical equations are shown below:

equation (18) is transformed to

$$v(s) = RI(s) + sL1I(s) (20)$$

$$I(s) = \frac{v(s)}{R + sL1} (21)$$

equation (13) is transformed to:

$$K1I(s) - KxX(s) (22)$$

$$\text{where } k1 = 2C * \frac{i0}{x0}$$

Equation number (19) becomes:

$$Vs(s) = \beta X(s) (23)$$

Finally, transfer function can be found using above equations. It is the ratio of sensor output voltage to input voltage

$$G(s) = \frac{Vs(s)}{V(s)} (24)$$

$$G(s) = \frac{K1}{s^3+p3*s^2-k2*s-p3*K2} (25)$$

where:

$$c = m*g*\left(\frac{x0}{i0}\right),\ KI = 2*c*\frac{i0}{x0}$$

$$Kx = 2*c*\frac{i0^2}{x0^3},\ K1 = -\frac{\beta KI}{mL1}$$

$$K2 = \frac{Kx}{m},\ p3 = \frac{R}{L1}$$

The linearized state space model yields:

$$\begin{bmatrix} \dot{x1} \\ \dot{x2} \\ \dot{x3} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ k2p3 & -k2 & p3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} * u (26)$$

$$y = \begin{bmatrix} K1 & 0 & 0 \end{bmatrix} * \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} (27)$$

III. Fuzzy Control

Fuzzy logic is an computing method, comparing to the traditional Boolean sets where variables take only two value true or false. Fuzzy logic have two limits, completely true limit take (1) value, and completely false limit take (0) value. The term "fuzzy logic" was introduced at 1965 by LotfiAZadeh. Fuzzy logic has been applied to many fields, from control theory to artificial intelligence, generally "error" and "change of error", for the process state and applying rules to decide a level of output.

There are a lot of models of FLC, but the most famous ones are the Mamdani model, Takagi- Sugeno-Kang (TSK) model andKosko's additive model (SAM) [10]. This paper uses Mamdani's model as given in Fig.2. The operational features of the FLC is explained in following:

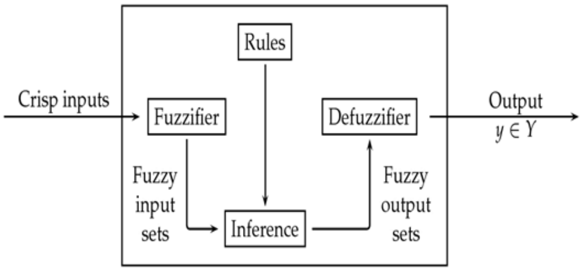


Fig. 2. Mamdani Model

Fuzzification: Fuzzification means converting a crisp value of process variable into a fuzzy set In Fuzzy set, numbers are converted into letters.

Rule Base: It consists of the IF-THEN rules, many approaches taken in determining the relation of the fuzzy rule:

- MamdaniìR(x,y) = min[ìA(x),ìB(y)].
- ZadehìR(x,y) = max{min[ìA(x),ìB(y)], 1 - ìA(x)}.
- LarsenìR(x,y) = ìA(x) . ìB(y).
- LukasiewicziR(x,y) = min{ 1, [1 - ìA(x) + ìB(y)]}

The rule can represent using these forms: IF error is zero AND change of error is zero THEN change of voltage is zero or by using Table 1:

TABLE I. SIMPLE RULE BASE

Error **ΔError**	**N**	**Z**	**P**
N	N	N	Z
Z	N	Z	P
P	Z	P	P

Defuzzification: Defuzzification operation is the reverse of the fuzzification operation which means the conversion of the fuzzy output values into crisp values [10]. There are many type of defuzzification:

- Mean of Maximum method (MoM)
- Center of Area (CoA)
- Center of Maximum (CoM)

IV. Hilink Board

The Hilink platform board is used as interface between physical plants and MatLab /Simulink for achievement of hardware in the loop of real time control systems, and is shown in Fig.3. This platform enables MATLAB/Simulink real time windows target to communicate with the control board in real time [11].

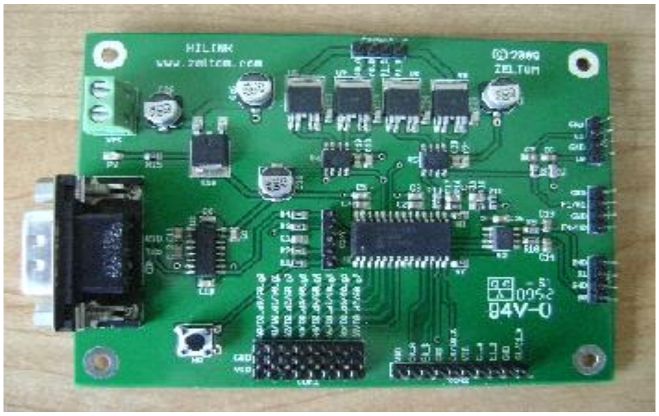


Fig. 3. Hilink platform

V. Fuzzy Logic Controller Design

Fuzzy logic controller is applied to single-axis levitation system. The best choice to get robust, flexible, faster and real time speed control was to use Mamdani model for fuzzy controller. In that two input error and change of error (e, ce) and one output change of voltage (cv), seven membership (triangular, Z membership) functions were used for fuzzification step with selected range [-1,1] for the inputs and output. The Fig.4 shows membership function of the fuzzy controller using GUI fuzzy toolbox of MATLAB software. Rule base of the system is given in Table 2.

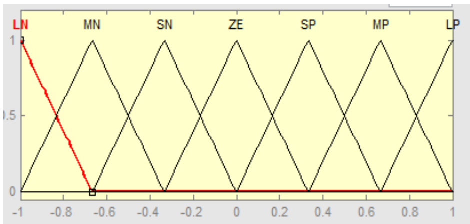


Fig. 4. Memberships of Inputs and Output Fuzzy Controller

TABLE II. Simple Rule Base

e ce	LN	MN	SN	ZE	SP	MP	LP
LN	LN	LN	LN	LN	MN	SN	Z
MN	LN	LN	LN	MN	SN	Z	SP
SN	LN	LN	MN	SN	Z	SP	MP
ZE	LN	MN	SN	Z	SP	MP	LP
SP	MN	SN	Z	SP	MP	LP	LP
MP	SN	Z	SP	MP	LP	LP	LP
LP	Z	SP	MP	LP	LP	LP	LP

VI. Simulation And Discussion

Fig.5 shows the output of the system drown using transfer function of the system without applying any controller. When no controller applied to the system the ball will fall down or attract to the coil.

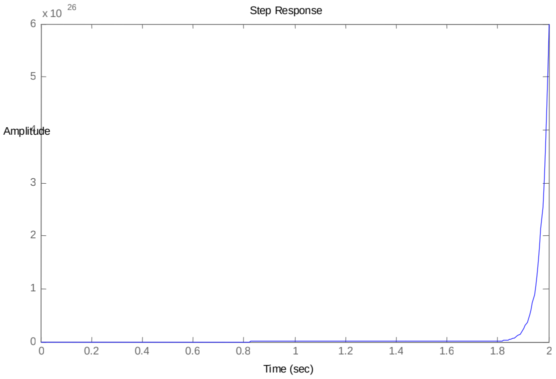


Fig. 5. Step response of the system without controller.

After applying the designed fuzzy controller it will be connected to a magnetic levitation MatLab / Simulink model. The model will built inside the Hilink /MatLab toolbox connected to maglev model then operates the system in real time target using T=1/2048 and S the stop time equals to inf.

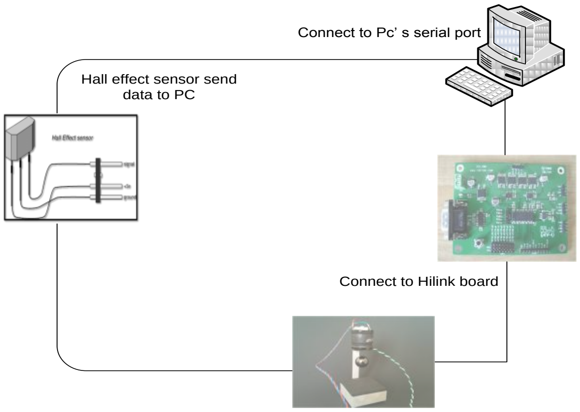


Fig. 6. Block diagram of physical system

Fig.6.shows the system that we used for testing FLC which consist from the Hilinkpaltform used as interface for connecting Matlab design with electromagnet. The position data measured using hall effect sensor then position will send to PC where the slider gain we use in our Matlabdesign is adjust slightly to bring the system to stable region .

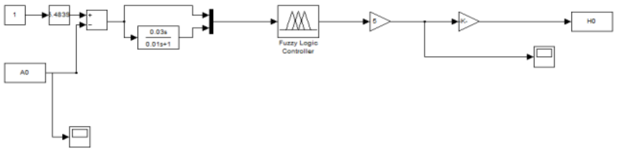


Fig. 7. .Matlab simulation of fuzzy Controller

Fig.7 shows the Matlab design for fuzzy logic control where the blocks A0 and H0 is taken from the Hilinksimulink library which present the analogue input and plus output respectively

Fig.8 shows the sensor output voltage which present A0 the system input .

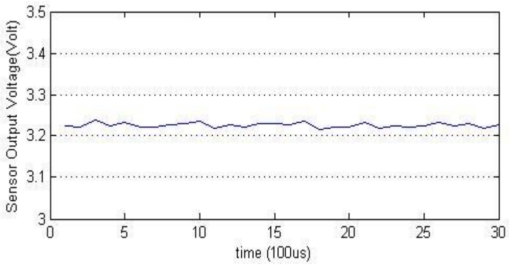


Fig. 8. Sensor output voltage

Fig.9 shows the output of the system after Appling FLC and Fig.10 shows the step response of the system after applying FLC

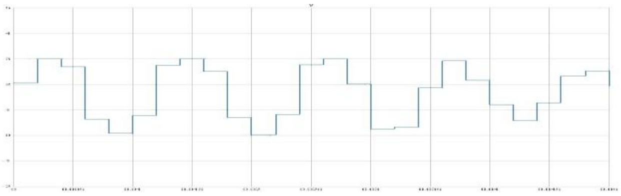


Fig. 9. Output of the system after connecting FLC

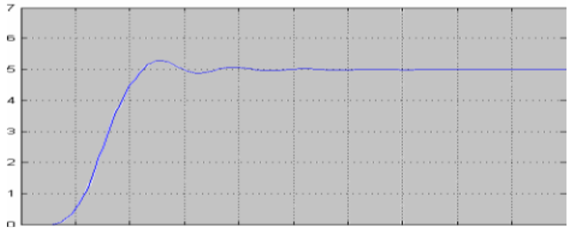


Fig. 10. Step response of the system using FLC

The LQRC interacts with the linearized model. The performance results of the system obtained from MATLAB are depicted in Figures 11, 12 and 13.

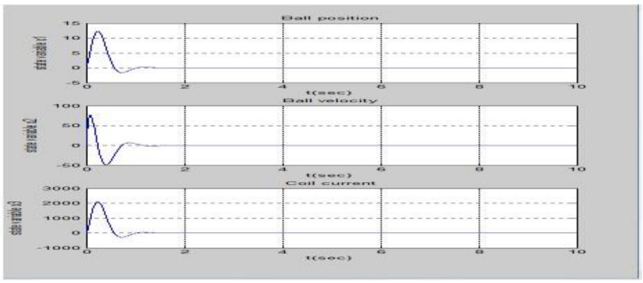


Fig. 11. x1 position, x2 velocity and x3 coil current versus time using LQRC

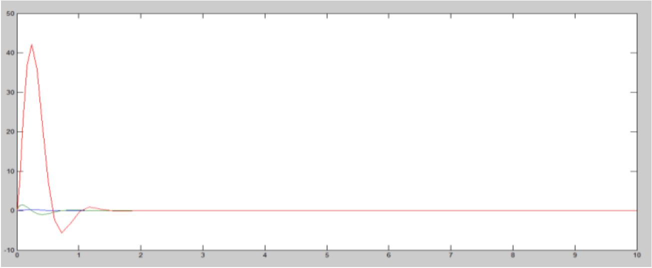


Fig. 12. x1, x2 and x3 versus time using LQRC

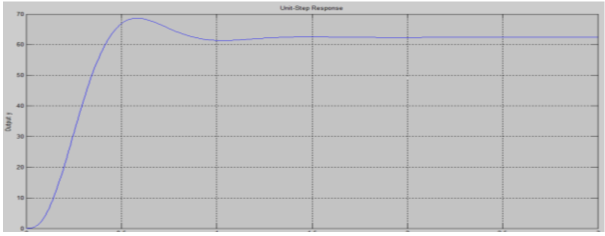


Fig. 13. Step response for LQRC control

These results shows that all of the variables x1, x2 and x3 will go to zero when time became infinity. So we can consider that the system is controllable, after two seconds the system will be stable from initial condition operation point which is about two centimeter of the ball position. But with the presence of surrounded disturbance makes the stabilizing system based on LQRC is not robust. LQRC not achieved the performance of Robust control strategies which need to reject the disturbance and work properly where the system dynamic model has uncertainty. From the step response for the FLC and LQRC we can see that the smallest amount of error is coming from FLC.

VII. Conclusion

Magnetic levitation system model mathematically presented, the Maglev system is experimentally analyzed using HILINK platform. Two type of controller were applied, FLC and LQRC. This paper presents a comparison between simulation results of both applied controllers.The system is controlled with FLC and LQRC but only with FLC control the system achieves robust control, fuzzy controller hasslower responsecomparing to LQRC. However, fuzzy controller shows the better performance because of its lowest overshoot among LQRC also steady state error not acquires in case of FLC. If the working conditions exceed the initial linear area, FLC presents robustness against model uncertainties and permits very precise positioning of the levitated object. While LQRC technique is not stable with the presence of a disturbance and cannot reject the perturbation. So it can be concluded that FLC has superiority over LQRC.

Comparing to pervious researches that mention in this paper LQRC is used instead of classical control like PID,both controllers control the magnetic levitation system but as we present in this paper they depends on the system parametersR,L, etc which may be change due to temperature or any type of disturbance,which is not happened with FLC.This paper applies control method to actual plant (ZELTOM/HILINK) to study the response of real system. In the future it can be implemented to connect FLC and LQRC output and apply to the system to solve the problem of the two controllers and get better results, also apply FLC to actual plant designed by our self.

References

[1] Dukaadrian –Vasile,GrifHoratiyStefan,OlteanStelianEmilian, A fuzzy-PD control design method for a class of unstable system, Interdisciplinary in engineering,scientific international conference TG. MUREȘ – ROMÂNIA, 15 -16 November 2007

[2] Hexiang Liu, Minqiang Hu, Haitao Yu and Li Yu, Study on Fuzzy Control in Electromagnetic Suspension SystemBased on the Prediction Model, International Conference on Intelligent Control and Information Processing August 13-15, 2010 - Dalian, China.

[3] Hosam Abu Elreesh, Hosam Abu Elreesh, FPGA Fuzzy Controller Design for Magnetic Ball Levitation, *I.J. Intelligent Systems and Applications*, 2012, 10, 72-81 Published Online September 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2012.10.08.

[4] Rong-Jong Wai, *Senior Member, IEEE*, and Jeng-Dao Lee,Adaptive Fuzzy-Neural-Network Control for Maglev Transportation System,IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 19, NO. 1, JANUARY 2008.

[5] Rong-Jong Wai, *Senior Member, IEEE*, and Jeng-Dao Lee,Robust Levitation Control for Linear Maglev Rail System Using Fuzzy Neural Network,IEE transaction on control system technology vol,17,vo.1 January 2009 .

[6] Hexiang Liu, Minqiang Hu, Haitao Yu and Li Yu,Study on Fuzzy Control in Electromagnetic Suspension System Based on the Prediction Model, International Conference on Intelligent Control and Information ProcessingAugust 13-15, 2010 - Dalian, China.

[7] A.K. Ahmad, Z. Saad, M.K Osman, I.S Isa, S.Sadimin,S. S. Abdullah,Control of Magnetic Levitation System Using Fuzzy Logic Control Second International Conference on Computational Intelligence, Modeling and Simulation

[8] Kashif Ishaque1, Yasir Saleem2, S.S Abdullah1, M. Amjad1, Munaf Rashid1 and Suhail Kazi1Single Input Fuzzy Logic Controller for Magnetic Levitation System,978-1-4577-0005-7/11/$26.00 ©2011 IEEE

[9] *SamehBdran, Prof.MaShuyuan, SamoSaifullah, Dr.HuangJie, C*omparison of PID, pole placement and LQR controllers for speed ratio control of continuously variable transmission (CVT),Mechatronic Center, Mechanical Engineering Beijing Institute of Technology (BIT) Beijing, China

[10] Hosam Abu Elreesh,Design of GA-Fuzzy Controller For Magnetic Levitation Using FPGA,A Thesis Submitted To The Faculty Of Engineering. In Partial Fulfillment of the Requirements For The Degree of Master of Science in Electrical Engineering,June 2011.

[11] HILINK: Real-time Hardware-in-the-loop Control Platform for Matlab/Simulink/datasheet.

Developing Parallel Application on Multi-core Mobile Phone

DhuhaBasheer Abdullah
Computer Science Dept.
Computer Sciences and Mathematics College
Mosul University
Mosul, Iraq

Mohammed M. Al-Hafidh
Computer Science Dept.
Computer Sciences and Mathematics College
Mosul University
Mosul, Iraq

Abstract—**One cannot imagine daily life today without mobile devices such as mobile phones or PDAs. They tend to become your mobile computer offering all features one might need on the way. As a result devices are less expensive and include a huge amount of high end technological components. Thus they also become attractive for scientific research. Today multi-core mobile phones are taking all the attention. Relying on the principles of tasks and data parallelism, we propose in this paper a real-time mobile lane departure warning system (M-LDWS) based on a carefully designed parallel programming framework on a quad-core mobile phone, and show how to increase the utilization of processors to achieve improvement on the system's runtime.**

Keywords—Mobile phone; Parallel programming; Multi-core processor; Lane detection

I. Introduction

Mobility is a key term in the world of today. No matter where you are or what you are doing you are surrounded by a world of mobile devices. No longer are we confined to work at a desktop terminal, we can now work and communicate from virtually any location. This new mode of interaction has been made possible by the advances in the world of miniaturization[1]. We can now work and communicate by using a myriad of devices such as: Laptops, Ultra Mobile PC's, PDA's and mobile phones. Even though we may be surrounded by these devices, the question must be raised: are we using them to their fullest potential? One possible solution to this lies within the area of parallel processing, without realizing the true potential of embedded multi-core architecture, we are not making full use of this now technology[2].

Multi-core architectures for personal computers is an important field in our everyday computing , many frameworks and APIs focusing on parallel programming have been proposed for multi-core processors,that achieve speedup and maximum processor utilization.

However, very little researches have been proposed in the area of multi-core architecture for mobile phones primarilybecause this is a relatively new concept and not many end-products have embraced such architecture. The recent release of multi-core mobile phones optimized for both performance and power consumption, such as Samsung Galaxy SII and SIII[3], has revolutionized mobile computing and opened up the door to new research paradigms, especially for real-time processing.

Now we can consider complex algorithms for potential implementation that previously regarded as impractical for deployment on mobile phones platforms. For instance, performing certain computation on big data matrices on mobile phone was very difficult in the past, due in part to its memory constraints but primarily to the processing power of the mobile phone. It is desirable to be able to use complex algorithms whenever possible because they generally yield more accurate results. Fortunately, the recent release of multi-core mobile phones has empowered us to do exactly that, as true parallelism can now be achieved [4].

II. Contribution

In this paper, we consider animplementation oflane departure warning systems, relying on the principles of task and data parallelism, we propose in this paper a parallel programming approach on quad-core mobile devices to detect road lanes and warn the driver if the car is departing the lane, that has the following properities:

1) Implementing parallel algorithm on multi-core mobile phone.

2) Show how to increase the utilization of processors to achieve improvement on the system's runtime.

III. Related Work

On the parallel programming front, making a task parallelizable and run on multiple cores can be a grueling process. Challenging issues include thread synchronization, data race, and starvation.

Many attempts have been made by researchers and programmers alike to design a high-level framework that provides an abstraction layer for programmers to use. Such framework allows the programmers to fully focus on application development without unnecessary worry about parallel programming. The ParLab at Berkeley, UPCRC at Illinois, and the Pervasive Parallel Laboratory at Stanford propose a two-layer framework, consisting of the productivity layer where domain experts, assumed to have limited experience with parallel programming, can focus on application development, and the efficiency layer where computer scientists with strong background in parallel programming can focus on improving the efficiency of the application [5].

Similar to the aforementioned framework, programming models such as algorithmic skeletons have also been proposed, aiming to benefit from multi-core architectures while decoupling the hassle of thread management from common programming. Skandium and Calcium [6] provide high-level parallel programming libraries based on the thread pool and ExecutorService frameworks in JAVA. Users only need to provide a threshold for threads and a set of initial parameters.

Daniel C. Doolan, and Laurence T. Yang in year 2006. They considered the problem of matrix multiplication to show and demonstrates that mobile devices are capable of parallel computation using Mobile Message Passing Interface (MMPI). MMPI allows parallel programming of mobile devices over a Bluetooth network [7].

PanyaChanawangsa, and Chang Wen Chen in year 2012. They demonstrate how proper utilization of a dual-core mobile processer can achieve tremendous speedup in mobile application [4].

Massimo Bertozzi, and Alberto Broggi in year 1998.They describes the Generic Obstacle and Lane Detection system (GOLD), a stereo vision-based hardware and software architecture to be used on moving vehicles to increment road safety [8].

Mars Lan, MahsanRofouei, Stefano Soatto and Majid Sarrafzadeh in year 2009. They built a SmartLDWS, that employs a novel lane detection algorithm that is both robust and scalable to overcome poor camera quality and limited processing power faced by most smartphones [1].

IV. Parallel Processing

Parallel Processing refers to the concept of speeding-up the execution of a program by dividing it into multiple fragments that can execute simultaneously, each on its own processor. A program being executed across ***N*** processors might execute n times faster than it would using a single processor [9].

A. *The benefit of using Parallel Processing*

In the earliest computers, only one program ran at a time. A computation-intensive program that took one hour to run and a tape-copying program that took one hour to run would take a total of two hours to complete their task. An early form of parallel processing allowed the execution of both programs simultaneously. The computer would start an input/output operation, and while it was waiting for the operation to complete, it would execute the processor-intensive program. The total execution time for the two jobs would be a little over an hour [10][11].

V. Dependencies

Understanding data dependencies is fundamental in implementing parallel algorithms. No program can run more quickly than the longest chain of dependent calculations (known as the critical path), since calculations that depend upon prior calculations in the chain must be executed in order. However, most algorithms do not consist of just a long chain of dependent calculations. There are usually opportunities to execute independent calculations in parallel.

Let Pi and Pj be two program fragments. Bernstein's conditions describe when the two are independent and can be executed in parallel. For Pi, let Ii be all of the input variables and O_i the output variables, and likewise for P_j. P_i and P_j are independent if they satisfy:

- Ij ∩ Oi = Φ
- Ii ∩ Oj = Φ
- Oi ∩ Oj = Φ

Violation of the first condition introduces a flow dependency, corresponding to the first statement producing a result used by the second statement. The second condition represents an anti-dependency, when the second statement (P_j) would overwrite a variable needed by the first expression (P_i). The third and final condition represents an output dependency: When two statements write to the same location, the final result must come from the logically last executed statement[12][13].

VI. Speedup

The speedup of code explains how much performance gain is achieved by running our program in parallel on multiple processors.A simple definition is that it is the length of time it takes a program to run on a single processor, divided by the time it takes to run on a multiple processors.Speedup generally ranges between 0 and **p**, where **p** is the number of processors [14]. Speedup is defined by the following formula:

$$\text{Total Speedup} = T_s / T_p \qquad (1)$$

T_s : is the runtime without parallelism.

T_p : is the runtime with parallelism.

VII. Multithreading And Processor Utilization

Since a mobile phone is considered a general-purpose device, application-level parallelism is the best we can achieve. Without explicitly using multiple threads, speedup from a multi-core architecture will not be obvious. In this paper, we propose a general guideline for breaking down a global task into multiple subtasks and later demonstrate how to apply this idea on a mobile lane detection system.

The first step towards parallelizing a task is to determine the optimal number of threads to use. Limiting thread contention is crucial for application speedup. Spawning too many threads than necessary not only disrupts other applications, but may also result in a longer execution time of the application due to the overhead associated with context-switching. A processor core can handle only one thread at a time. For efficiency purposes, a simple rule is to spawn as many threads as the number of cores available, thereby delegating one thread to each core and eliminating the need for time-slicing. A simulation was conducted by spawning different numbers of threads to execute certain tasks. A dramatic improvement in execution time can be seen when we increase the number of worker threads from 1 to4. However, since there are only four available cores, increasing the number of threads do not enhance but aggravates the performance, resulting in a slightly longer execution time[15][16].

VIII. Proposed Mobile Lane Detection System

Transportation safety is an issue of ever increasing concern. In the U.S. alone, more than 30,000 casualties suffered from traffic crashes in 2009, according to the National Highway Traffic Safety Administration. A considerable number of car accidents occurred as a consequence of the drivers failing to keep the vehicles within the designated lane. In the wake of such tragedies, attempts to incorporate an Intelligent Vehicle Assistant System into automobiles have been made by adopting various computer vision approaches.

While LDWS has been installed on many trucks and othercommercial vehicles in both Europe and North America and has shown to significantly reduce preventable accidents, it still often remains as an option even for the luxury passenger vehicles. Cost is often cited as one of the main reasons impeding a wide-spread adoption of LDWS. On the other hand, the professional skill required to install such a system, to calibrate the camera, and to integrate it into vehicle's electronics should not be overlooked, either[1].

In order to bring LDWS to the mainstream market, we propose Mobile-LDWS,to make the system available for everyone. The system acquires images of the road and intelligentlylabels the lane marks. The heading of the vehicle as well as its position with respect to the road boundaries can thus be determined.If the system believes the vehicle is about to depart from the current lane, it triggers an alarm to alert the driver.Other applications of lane detection include autonomous driving for cruise control as well as robot navigation.

A. Development Platform

The system was developed on Android 4.1.2 (Jelly Bean). Released in September 2012. The phone's most outstanding feature is its processor – a superscalar quad-core 1.4 GHz Arm Cortex-A9 with 2 GB RAM [3]. Optimized for high performance and low power consumption, the Galaxy SIII is indeed an ideal platform for this system.

B. Data Acquisition

The proposed lane detection system uses the phone camera as the means of sensing. It captures raw image frames at the rate of 30 frames per second (fps) and feeds them to the preprocessing module. The phone is attached onto the car windshield by an off-the shelf GPS mount. It should be placed in such a way that the camera is able to see the road clearly while precaution should be taken during this step so that the device does not obstruct the driver's view (Figure 1).

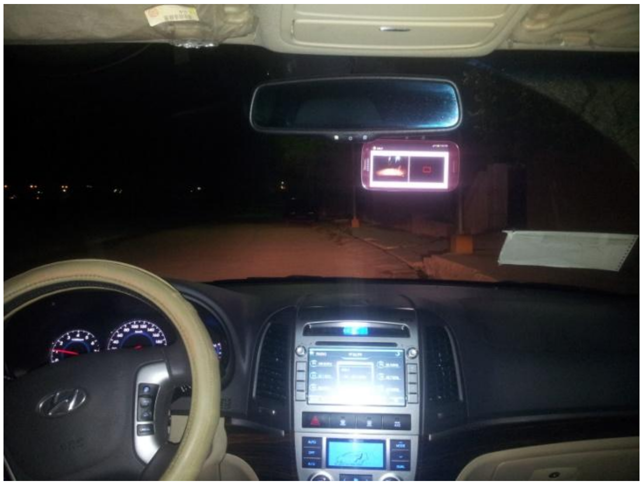


Fig. 1. The M-LDWS on a Samsung GalaxySIII in action on an off-the-shelf GPS mount.

Android's camera API allows programmers to process the preview frames directly. This feature enables us to overcome the storage constraint, as we can process those frames and display the detected lane boundaries on the phone screen without having to record any of them. The frame data come in the form of byte arrays whose default format is YCbCr, to get the three base colors (read , green , blue), we convert the frame format from YCbCr to ARGB8888. Finally, each pixel values are stored in an image matrix. Now we can implement the basic matrix and linear algebra operations.

C. Lane Detection

The lane detection algorithm is based on the assumptions that the lanes are defined by a clearly painted white line.

To detect a line we first must find out the most important features of it, which are color and shape. We first used the color feature to detect a line. Since the color of the road lane is white, we perform an operation to extract white color and change it to green, based on the following:

$$g(x,y) = \begin{cases} 255, & \text{if } r(x,y) > val,\ g(x,y) > val,\ b(x,y) > val \\ 0, & \text{otherwise} \end{cases}$$

Where :
r(x,y) is the red value of the pixel in the position (x,y).
g(x,y) is the green value of the pixel in the position (x,y).
b(x,y) is the blue value of the pixel in the position (x,y).
val is a predefined value.

In the next step, the shape feature will be used to detect the lines, and perform a vertical edge detection filter to extract the lines (Figure 2).

D. Lane Departure Warning

The next step is to find out if the vehicle is in the lane or it is about to depart it. If the system believes the vehicle is about to depart from the current lane, it triggers an alarm to alert the driver. This can be done by scanning a specific region in the image for any lines that have been detected.

If the result of the detection process is true, that means that the vehicle is departing the current lane so the system alert the driver.

E. Proposed Parallel Framework

The next step is to determine if data parallelism is possible and appropriate. For many image processing tasks, data comes in the form of image matrices, or on the lower level 2-dimensional arrays. In many cases, they can be split up into smaller independent chunks and processed concurrently, reducing the execution time while producing the same output as when processed sequentially. For the image edge detection, we can split up the image matrix into smaller sub-matrices (matrix slicing) and slide an edge detection convolution mask over each sub-matrix concurrently. However, if the size of data to be processed is not significantly large, employing data parallelism will not yield much speedup as a result of thread overhead. In the case of this application, since the image are large, data parallelism is well worth a try. Generally speaking, given k processor cores, we should divide the input data of size n into n/k smaller chunks anddistribute them across k cores with each core running a single thread. Figure (3).

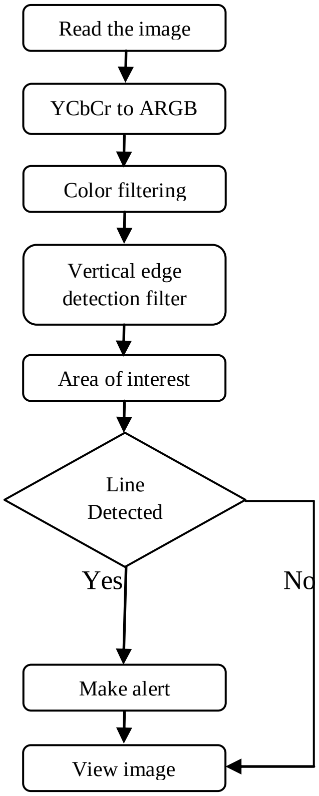


Fig. 2. The proposed serial lane detection system

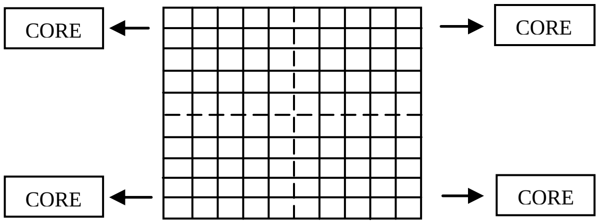


Fig. 3. Matrix slicing

The parallel computation requires several steps that are not required in the sequential version of the application (Figure4). One of the first main differences is that we need to determine the number of threads suitable for the phone hardware. Since Samsung Galaxy SIII has four cores, four threads are ideal for execution the task in hand. So that the image must be split in to four same size blocks called sub-image, then send each sub-image to one of the four cores we have. Once this operation has been completed each core can compute its own section of the image. The final stage is to gather all the results back in to one image. The result of the computation can then be available for the next steps.

To gain full control of task management, we also make use of theFutureTask class , allowing to track the progress of the submitted tasks and block until all of them have been completed. The four tasks are eventually submitted to an ExecutorService, which takes care of thread pool creation and assigns a submitted task to an available thread. More importantly, by using the ExecutorService, memory consistency is guaranteed, thus eliminating the trouble of thread synchronization.

IX. EXPERIMENTAL RESULTS

The system tested on image size (320×240), The results are shown in (Table (1)). Time results represents the average runtime of hundred frames.

TABLE I. EXPERIMENTAL **RESULTS**

Method	Run time
Serial	11043ms
Parallel	6506ms

From the time results in (Table (I)), we can calculate the speedup as below:

Total speedup = 11043/6506=1.69

This means that the reduction in the overall processing time is 41.08%.Figure (5) shows lane detection results.

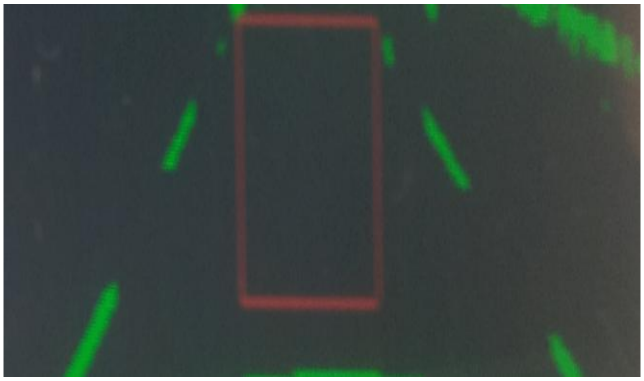


Fig. 5 Lane detection results

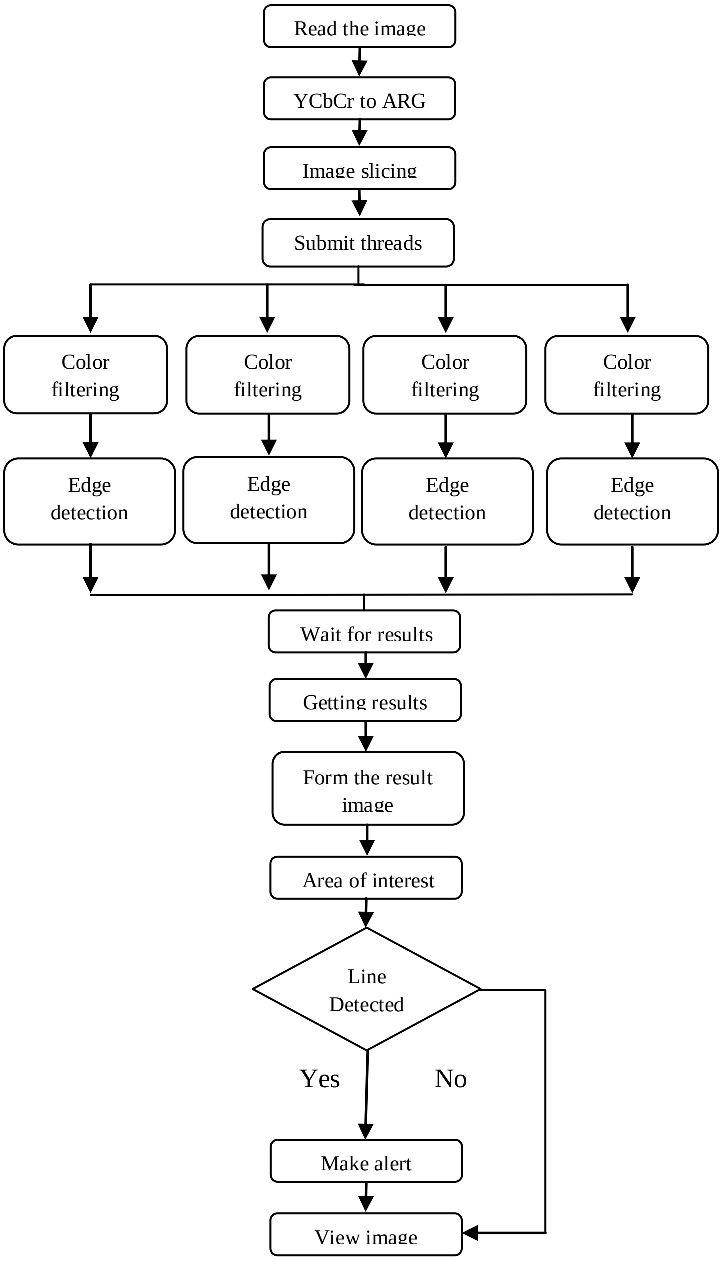


Fig. 4 Modified lane detection system with the proposed parallel framework

X. Conclusions

In this paper, we have demonstrated how to achieve speedup in a mobile lane detection system written entirely in JAVA by using the proposed parallel programming approach, based on the idea of task and data parallelism. Running on a quad-core Samsung Galaxy SIII. The proposed system shows significant reduction in the overall processing time and good speedup.

References

[1] M. Lan, M. Rofouei, S. Soatto, M. Sarrafzadeh," SmartLDWS: A Robust and Scalable Lane Departure Warning System for the Smartphones", Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, October 3-7, 2009.

[2] D. Doolan, S. Tabirca, L. Yang," MMPI a Message Passing Interface for the Mobile Environment", Proceedings of MoMM2008, Linz, Austria, 2008.

[3] Samsung I9305 Galaxy S III Full Specifications, http://www.gsmarena.com/samsung_i9305_galaxy_s_iii-5001.php

[4] P. Chanawangsa, C. Chen, "A New Smartphone Lane Detection System: Realizing TruePotential of Multi-core Mobile Devices", MoVid'12, 2012, pp.19-24.

[5] B. Catanzaro, et al., "Ubiquitous Parallel Computing form Berkeley, Illinois, and Stanford", IEEE Computer Society, 2010, pp. 41-55.

[6] T. Panagiotis, "Evaluating Skandium's Divide-and-Conquer Skeleton", Master Thesis, School of Information, University of Edinburgh, 2010.

[7] T. Yang, D. Doolan, "Mobile Parallel computing", Proceedings of The Fifth International Symposium on Parallel and Distributed Computing, IEEE International,2006.

[8] M. Bertozzi, A. Broggi, "GOLD: A Parallel Real-Time Stereo Vision System for Generic Obstacle and Lane Detection", IEEE Transaction on image processing, VOL.7,NO. 1, JANUARY 1998.

[9] J. Nancy, J. Richard, A. James, "PARALLEL PROCESSING: THE NEXT GENERATION OF COMPUTERS", National Energy Technology Laboratory,2011.

[10] M. Sasikumar, D. Shikhare, P. Prakash, Introduction To Parallel Processing, Prentice-Hall of India Private Limited, 2006.

[11] C. Evangelinos, C. Hill," Cloud Computing for parallel Scientific HPC Applications: Feasibility of running Coupled Atmosphere-Ocean Climate Models on Amazons EC2.", ratio, vol. 2, no. 2.40, pp. 2–34, 2008.

[12] J. Hennessy, D. Patterson,"Computer Architecture: A Quantitative Approach", Morgan Kaufmann Publishers, SF, CA, 1996.

[13] H. Dietz, "Linux Parallel Processing HOWTO", v980105, 5 January 1998.

[14] D. Marshall, Parallel Programming with Microsoft Visual Studio, Microsoft Corporation by: O'Reilly Media, 2011.

[15] D. Abdullah and M. Al-Hafidh, " The True Powers of Multi-core Smartphone", IJCSI, , Vol. 10, Issue 4, No 2, July 2013.

[16] H. Guihot, Pro Android Apps Performance Optimization, Apress, 2012.

Cross Correlation versus Mutual Information for Image Mosaicing

Sherin Ghannam and A. Lynn Abbott
Bradley Department of Electrical and Computer Engineering
Virginia Tech
Blacksburg, Virginia, USA

***Abstract*—This paper reviews the concept of image mosaicking and presents a comparison between two of the most common image mosaicing techniques. The first technique is based on normalized cross correlation (NCC) for registering overlapping 2D images of a 3D scene. The second is based on mutual information (MI). The experimental results demonstrate that the two techniques have a similar performance in most cases but there are some interesting differences. The choice of a distinctive template is critical when working with NCC. On the other hand, when using MI, the registration procedure was able to provide acceptable performance even without distinctive templates. But generally the performance when using MI with large rotation angles was not accurate as with NCC.**

Keywords—mosaicing; normalized cross correlation; mutual information

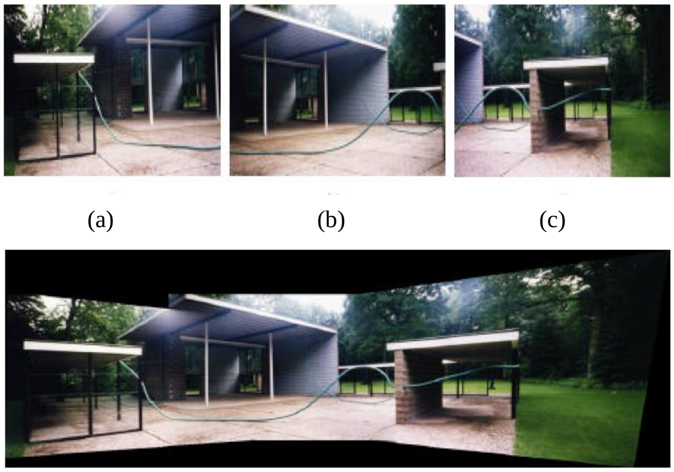
I. INTRODUCTION

Mosaicing refers to the process of combining multiple photographic images with overlapping fields to produce a single image of the whole scene. Mosaicing, also known as panographic photography or image stitching, has an extensive research literature [1-3] and several commercial applications [4-6].The automatic construction of large and high-resolution image mosaics is an active area of research in the fields of photogrammetry, computer vision, image processing, medical image, robot vision and computer graphics [7,8]. The most traditional application is the construction of large aerial and satellite photographs from collections of images.

The report is organized as follows. An overview of image mosaicing is provided in section II. Details about image registration are presented in section III. Section IV demonstrates the matching process. The methodology is presented in section V. A brief discussion about optimization is presented in section VI. Quality assessment is illustrated in section VII. Section VIII introduces the experimental results. Finally, section IX concludes the work.

II. IMAGEMOSAICING

Mosaicing, combines overlapping images to produce one composite image for a scene [1], as shown in Fig.1. The first part of an image mosaicing operation consists of identifying correspondences between some features present in both images, in order to determine the geometric transformation necessary to align the two images. This alignment operation is called *image registration*. After alignment, a composite image is created by merging or averaging pixel values of the overlapping portions and retaining pixels where no overlap occurs.



(a) (b) (c)

(d)

Fig. 1. Illustration of mosaicing (from [9]). (a-c) Input images. (d) Resulting mosaic.

III. IMAGE REGISTRATION

Image registration is the process of overlaying two or more images of the same scene taken at different times, from different viewpoints, and/or by different sensors. The registration process requires computational methods for determining point-by-point correspondences between two images of a scene. Registration may be used to fuse complementary information in the images or to estimate the geometric and/or intensity difference between the images [10].

From corresponding positions in two images, a transformation function can be determined to get correspondences between the remaining points in the images. The aim of registration is find the transformation parameters that maximize a similarity metric or minimize dissimilarity metric between the images to be registered. The optimization problem can be formulated as follows:

$$\hat{T} = \arg\ \max_{T\in\tau} S(V,U,T) \qquad (1)$$

where$U$ and V are the first and second images to be registered, T is the transformation, τ is the search space, S is

This work was supported by Egyptian Ministry of Higher Education,

the similarity measure and $\hat{T}$ is the optimal solution. These concepts are illustrated in Fig. 2.

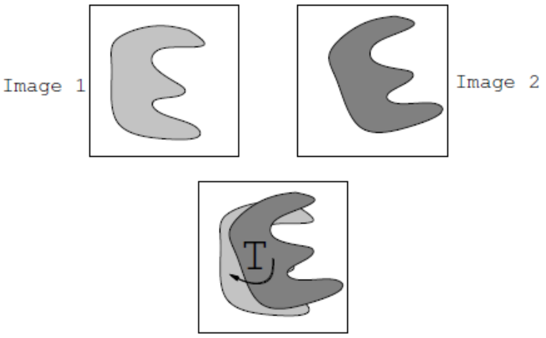


Fig. 2. Registration process (from [11]).

The success of the registration often requires that the search space$\tau$is relatively small with respect to the input data. Large τ increases the probability of getting trapped in a local minimum.

Transformations can be rigid, depending only on translation and rotation. In that case a transformation can be represented as

$$T(X) = \boldsymbol{R}\, X + \boldsymbol{t} \qquad (2)$$

where$\boldsymbol{R}$ is a 2×2 rotation matrix with one degree of freedom, and $\boldsymbol{t}$ is a2×1 translation vector. More generally, $\boldsymbol{R}$can incorporate additional degrees of freedom, and the resulting affine transformation represents a composition of rotation, dilation, and shear. For either case, the transformation can be represented in homogeneous form using a single matrix $\boldsymbol{M}$ as

$$T(X) = \boldsymbol{M}\, X \qquad (3)$$

where the single matrix ***M***is 3×3 and *X* is 3×1.

The majority of registration methods consist of the following four steps: feature detection, feature matching, transform model estimation, and image resampling and transformation [3].

Feature detection refers to the detection of salient and distinctive locations in the images, such as intensity edges, corners, line intersections, etc. These features are called control points (CPs) in the literature. In the matching step, the correspondences between the features detected in the input image and those detected in the reference image are established. A detailed discussion for this crucial stage is presented in the next section.

The next step is to estimate transform model parameters, as needed in (3), using the detected correspondences. The final step is to perform that transformation, using an appropriate resampling technique (such as nearest neighbor, linear or cubic interpolation) to represent the transformed image.

IV. (DIS)SIMILARITY METRICS FOR MATCHING

Physically corresponding features can be quite dissimilar in appearance due to imaging conditions.
To identify correspondences, two major categories of matching methods are often used: area based methods and feature based methods.

A. Area (Intensity) Based Methods

Area based measures rely on computations between “windows” of pixel values in the two images. Two such methods are normalized cross correlation and mutual information. As described here, these methods provide measures of image similarity, because larger values result for corresponding points. Area based examples of dissimilarity include sum of squared difference, and sum of absolute difference [12].

An advantage of these techniques is that they can be applied to image data directly, and do not require higher-level structural analysis. But they have the disadvantage of sensitivity to intensity changes, introduced for instance by noise, varying illumination, and/or by using different sensor types.

An approach to enhance dealing with intensity changes is to use the zero-mean normalized cross correlation (NCC); also called cross covariance. It is defined as

$$NCC(i,j) = \frac{\sum_{x,y}(V(x,y)-\bar{V})(U(x-i,y-j)-\bar{U})}{\sqrt{\sum_{x,y}(V(x,y)-\bar{V})^2}\sqrt{\sum_{x,y}(U(x-i,y-j)-\bar{U})^2}} \qquad (4)$$

Where *x* and *y* are the pixel coordinates while *i* and *j* refer to the shift at which the *NCC* coefficient is calculated. The resulting matrix *NCC* contains correlation coefficients with values between -1.0 and 1.0.Note that Urefers to the input image after being transformed by (3).

Mutual information (MI) is another popular matching metric used for image registration [13,14]. It is based on information-theoretic concepts, andcan be considered a measure of the statistical dependency between the data sets. This metric requires the computation of joint histograms as shown in Fig.3. In the figure, n_{ij} is the number of pixels with color *i*in *I* and with color *j* in *J*. The valuesn_i and n_j are the marginal values, i.e., histograms of *I* and*J*.

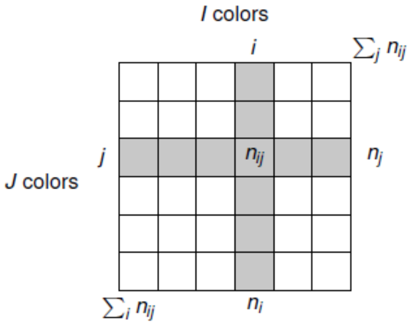


Fig. 3. Joint histogramfor images *I* and *J*.

To define the MI between two images, we regard them as random variables X and Y and their intensity values at certain coordinates in the images as the joint outcome of a random experiment. MI is defined in terms of entropy as follows:

$$\begin{aligned} MI(X,Y) &= H_e(Y) - H_e(Y|X) \\ &= H_e(X) + H_e(Y) - H_e(X,Y) \end{aligned} \qquad (5)$$

where $H_e(X) = -\sum_X p_d(X)\log(p_d(X))$ represents the entropy of random variable X and $p_d(X)$ is the probability density, as estimated by a histogram. Then $H_e(X,Y) = -\sum_X\sum_Y p_d(X,Y)\log(p_d(X,Y))$ represents the joint entropy between the two random variables X and Y.

In information theory, entropy is considered to be a measure of the uncertainty in a variable. For illustration, example probability distributions and their associated entropy values are shown in Fig. 4. Flatter distributions represent higher levels of uncertainty in the output of an experiment.

Typically, the joint probability distribution of two images is estimated as a normalized joint histogram of the intensity values. The marginal distributions are obtained by summing over the rows or over the columns of the joint histogram. The normalized joint histogram is calculated by counting the number of pixels having an intensity value *i* in the first image and an intensity value *j* in the second image at the same location and denoting it by n_{ij}. At each location in the first image, the pixel intensity value *i* will be examined and its corresponding value *j* in the same position in the second image. This will increase the counter $n_{ij}$by one. This process is repeated for each pixel. To obtain the normalized joint histogram, all values of n_{ij}should by divided by the sum of these values.

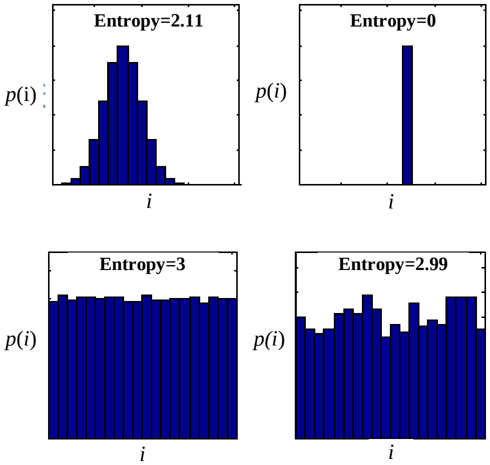


Fig. 4. Examples of probability distributions with associated entropy values.

Registration is achieved by adjustment of the relative position and orientation until the MI between the images is maximized. Some studies have shown the use of MI can give good performance even for multimodal situations [13].In such cases, intensity values are not linearly related so the cross correlation is not likely to succeed.

B. Feature Based Methods

These methods are preferred when the local structural information is more significant than the information carried by the image intensitiesal one. Features should be distinctive, distributed well over the images, and efficiently detectable in both images. These features can be based on regions, lines, and points.

Feature detection from regions relies on the ability to extract a useful subset of an image, possibly by applying a thresholding operation to a high-contrast image. Examples of region features are centroid, area, and elongation.

Line features can result from edge detection followed by line fitting. Common edge detection methods, such as the Canny technique or the Laplacian of Gaussian, result in a set of points in the image. A line-fitting algorithm such as the Hough transform can result in a set of lines, and their properties can be used as features for matching.

Point features often rely on the detection of intensity edges or corners in an image. Fig. 5 shows corner correspondences that have been detected in two images [9].

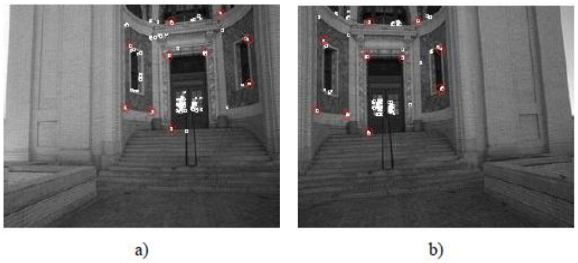


Fig. 5. Corner correspondences between two images (from [9]).

V. Methodology

This section illustrates the problem of creating an image mosaic from two overlapping 2D images of the same 3D scene. The NCC and MI similarity measures are used for registration, and the results are compared.The steps are summarized in Fig. 6.

A. Image Preprocessing

If image involves more than one color band, e.g. RGB, typically only one band is taken into consideration while performing the matching process. But after finding the optimal transformation parameters, all color bands will be processed during the transformation (alignment) stage to produce a color mosaic.

B. Matching

Given the two images, the task is to find correspondences between them. This is done by performing registration using two area-based similarity measures, NCC and MI. The position at which either NCC or MI is maximized will be stored.

To accommodate rotation, this process will be repeated at several rotation angles of the original image. This search interval can be reduced using an optimization technique which will be illustrated in the next section.

Applying this matching process over the whole image consumes a lot of time especially for large images. This process can be accelerated using small template(s) taken from each of the images. The matching process will be carried out using these templates. The template-matching process that finds the correspondence produces more reliable matches if the selected templates are locally unique [15]. A template that is relatively homogeneous may easily lead to false matches.

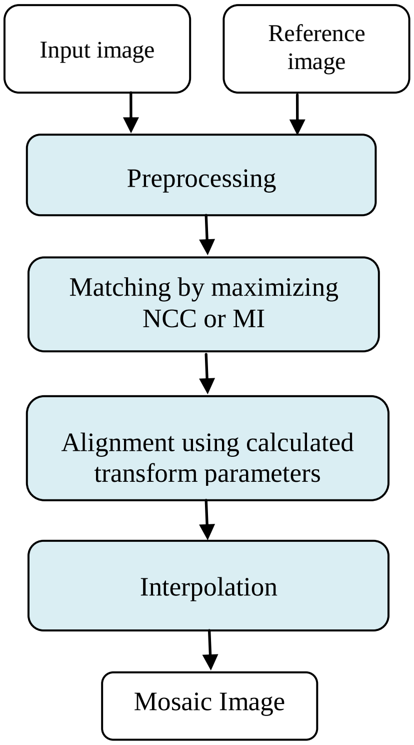

Input image
Reference image
Preprocessing
Matching by maximizing NCC or MI
Alignment using calculated transform parameters
Interpolation
Mosaic Image

Fig. 6. Flow chart of image mosaicing system.

C. Transformation

After using image templates to determine the geometric transformation Tthat is needed, the entire input image is transformed using (3).This process is sometimes called alignment.

D. Interpolation

Typically, a transformation T will require the computation of new pixel values using several pixels from the original image. This step can be performed by averaging pixel values locally, although other techniques such as spline-fitting have also been employed.

VI. Optimization

Optimization techniques often rely on the maximization or minimization of an objective (cost) function[16].Many optimization techniques can suffer from finding solutions that correspond to a local optimum of the objective function, instead of the solution that is the best overall.

For optimizating functions of n variables, many algorithms work by doing a sequence of 1D optimizations. For the case of 1D minimization, as illustrated in Fig. 7, it is possible to subdivide a given range [a,c]iteratively in an attempt to find the optimum solution.

Pseudocodefor the process is givenin Fig. 8.Line search techniques proposed in the literature differ in the way of calculating the value of the intermediate point b. For example, in the golden search method, b is either $\left(\frac{1}{2}-\frac{\sqrt{5}}{2}\right)$ or $\left(\frac{\sqrt{5}}{2}-\frac{1}{2}\right)$times the distance between $a$and$b$.

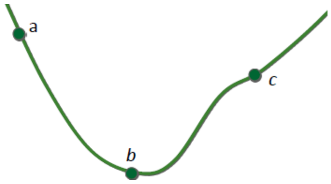


Fig. 7. Subdividing process for 1D optimization.

```
% Line search algorithm stops when reaching a prespecified tolerance (TOL) or maximum number of iterations K
$while((c-a) > TOL \mid iterations < K)$
    $if\ f(b) < f(a) \& f(b) < f(c)\ then$
        $if\ f(a) > f(c)\ then$
            $a = \frac{a+b}{2}$
        $else$
            $c = \frac{b+c}{2}$
        $endif$
    $else$
        $if\ f(a) < f(b)\ then$
            $dx = b - a$
            $c = b$
            $b = a$
            $a = b - dx$
        $else$
            $dx = c - b$
            $a = b$
            $b = c$
            $c = b + dx$
        $endif$
    $endif$
$endwhile$
```

Fig. 8. Pseudocode for basic line search.

In the problem of image registration, the only method that guarantees a globally optimum solution is an exhaustive In the problem of image registration, the only method that guarantees search over the entire image. Although computationally demanding, such a global search is often used if only translations are to be estimated. In case of transformations with more degrees of freedom or in case of more complex similarity measures, more sophisticated optimization algorithms are required [12].

VII. Quality Assessment

Different metrics can be used to measure the quality of the mosaicing result [17]. Mean square error (MSE) and peak signal-to-noise ratio (PSNR) are commonly used. They are applied on the overlapping region *O*:

$$MSE = \frac{\sum_{(x,y)\in O}\left(I_2(x,y) - I_1(x,y)\right)^2}{N} \qquad (6)$$

$$PSNR = 10\, log_{10} \frac{(255)^2}{\sum_{(x,y)\in O}\left(I_2(x,y) - I_1(x,y)\right)^2} \qquad (7)$$

A reconstruction error (*E*) measuring the mean of the absolute intensity differences between two successive images on the overlapping area (*O*) has also been used [18]. It is defined as

$$E = \frac{1}{|O|} \sum_{(x,y)\in O} |I_2(x,y) - I_1(x,y)| \qquad (8)$$

Exposing one image to changes in luminance or noise greatly affects these parameters. This means that low PSNR or high MSE or high *E* does not always mean poor registration.

CVLab [19] has suggested an evaluation methodology for the comparison of image mosaicing algorithms. The idea is to compare mosaics to their ground truth versions. This work was inspired by the work of in the performance evaluation of stereo reconstruction algorithms [20].

VIII. Experimental Results

Mosaicing techniques using NCC-based and MI-based similarity measures were implemented using Java, and were applied to two different image pairs to investigate their strengths and shortcomings. The first image pair, shown in Fig. 9(a-b), is from a football field. The images contain some obvious distinctive parts against a background that is largely homogenous. The second image pair is from remote sensing. For these experiments, only rigid transformations, translation and rotation, were considered.

The first two images overlap by nearly 8%. Templates for matching were obtained by applying a thresholding operation, and they are shown in Fig. 10. Because the two templates are fairly small in size, searching for the optimal translations and is relatively fast. But repeating this process for many possible rotation angles is computationally intensive, especially for the MI-based technique. In order to reduce the computation time, a line search (as described previously) was applied to the rotation-angle search space.

The second image pair is shown in Fig. 11(a-b).The two images overlap by nearly 25%.Templates for matching were obtained in this case, as shown in Fig. 12, through a cross correlation using binary templates that were extracted from the two images.

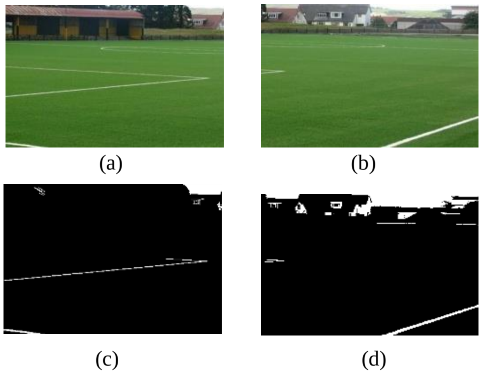


Fig. 9. Thresholding operation on first image pair. (a-b) First and second images. (c-d) Binary images.

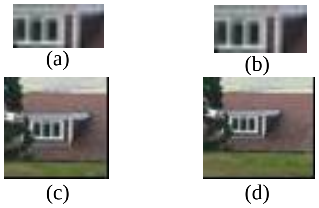


Fig. 10. Template selection. (a-b) Selected objects from binary images. (c-d) Corresponding templates.

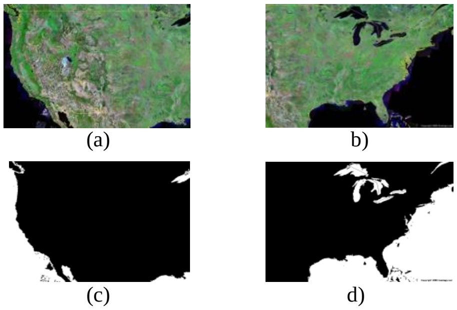


Fig. 11. Thresholding operation on second image pair. (a-b) First and second images. (c-d) Binary images.

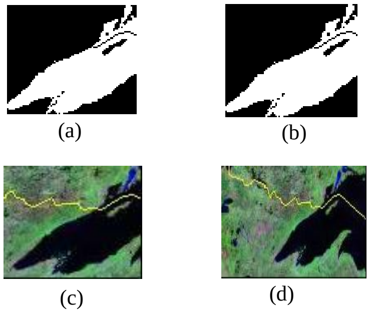


Fig. 12. Template selection. (a-b) Selected objects from the two thresholdedimages. (c-d) Corresponding templates.

Mosaicing results for different rotation angles are shown in Table I. The two input images are displayed in the first column, the two chosen templates are in the second column and the third column contains the resulting mosaics. Values of NCC, MI, and PSNR for the overlapping areas are displayed under each result. A number of experiments were carried out to test the performance of the mosaicing process under non-modeled variations such as intensity manipulation, shearing, and Gaussian noise addition. The resulting mosaics after applying these operations for zero rotation angles are shown in Table II.

To clearly compare the performance of the two techniques, graphs showing the computed rotation angles from each technique at each case versus the actual rotation angles are plotted in Figs.14 to 21.

From Fig. 13and Fig. 17, it can be seen that registration using MI may have a considerable amount of error at large rotation angles, while the performance of the NCC-based technique was better. If there is a substantial difference in luminance between the two images, these errors increase. For example, the MI-based technique successfully deduced the correct angle when the actual rotation angle was -30 in the second test set before increasing the brightness of one image. After changing the luminance, the MI-based test yielded an angle of 71°, as shown in Fig. 18. On the other hand, NCC succeeded in finding angles very close to the correct ones.

Adding Gaussian noise (with a variance of 12) also affected the resulting mosaics and produced some errors as shown in Fig.19. The errors in the case of NCC-based were around 5° while for the case of MI, errors were fairly large at some large rotation angles. The same scenario happened after applying shearing with factors of 0.15 and 0.35 as shown in Fig.20. Shearing caused almost all computed angles to be incorrect by about 5°, plus some dramatic errors in the case of MI-based at large rotation angles.

Althought it seems that the NCC-based technique is superior to the MI-based technique, especially at large rotation angles, there are some situations that the NCC-based technique could not handle and MI-based performance was much better.

One of these situations corresponds to the case that the chosen templates do not have distinctive details (i.e. nearly homogenous texture). Table III shows some results of the mosaicing techniques based on templates that are not distinctive. Even at small rotation angles such as 10°, the NCC-based technique could not extract accurate registration parameters, as shown in Fig. 21. However, the MI-based technique succeeded.

TABLE I. SOME RESULTS FOR DIFFERENT ROTATION ANGLES.

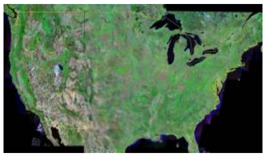
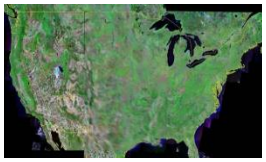
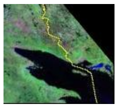
Test Image 1: Horizontal and vertical translation and zero rotation		
		 NCC-based result NCC=0.997, MI=1.84, PSNR=29.45
		 MI-based result NCC=0.997, MI=1.84, PSNR=29.45
Test Image 1: Horizontal and vertical translation and 10° rotation		
		 **NCC-based result** NCC=0.996, MI=1.85, PSNR=29.06
		 **MI-based result** NCC=0.996, MI=1.85, PSNR=29.06
Test Image 1: Horizontal and vertical translation and 50° rotation		
		 NCC-based result NCC=0.996, MI=1.85, PSNR=29.1
		 **MI-based result** NCC=0.996, MI=1.185, PSNR=29.1

Table I, continued

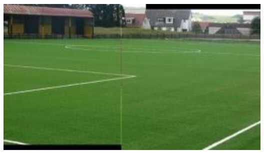
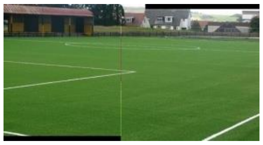
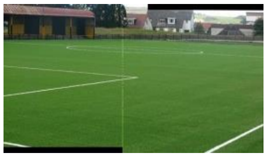
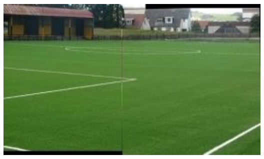
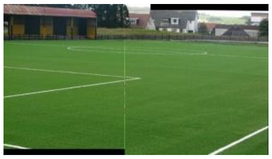
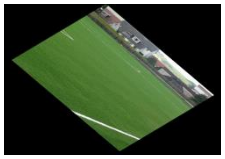
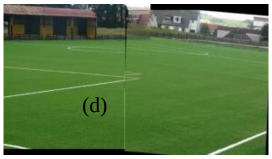
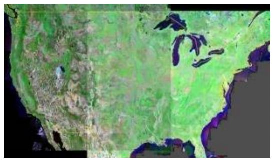
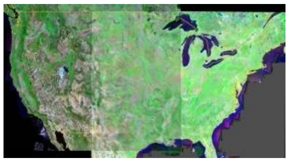
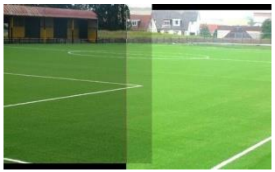
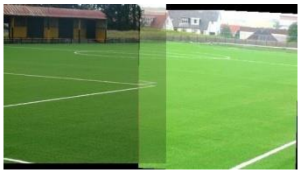
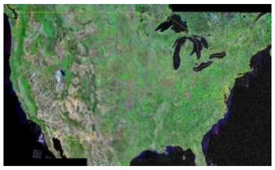
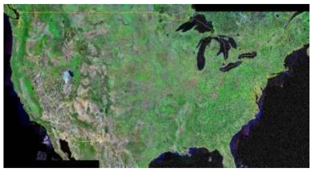
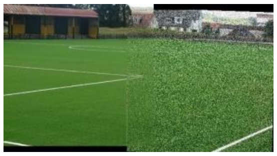
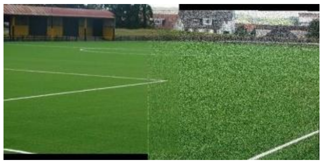
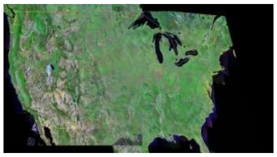
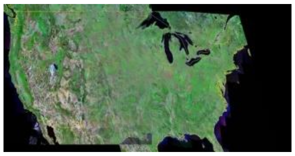
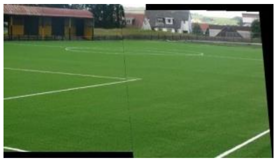
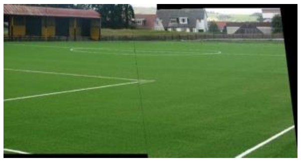
Test Images 2: Horizontal and vertical translation and zero rotation		
		 **NCC-based result** NCC=0.999, MI=2.66, PSNR=35.46
		 **MI-based result** NCC=0.999, MI=2.66, PSNR=35.46
Test Images 2: Horizontal and vertical translation and 10° rotation		
		 **NCC-based result** NCC=0.999, MI=2.69, PSNR=35.22
		 **MI-based result** NCC=0.993, MI=2.35, PSNR=27.69
Test Images 2: Horizontal and vertical translation and 50° rotation		
		 **NCC-based result** NCC=0.996, MI=2.33, PSNR=30.13
		 **MI-based result** NCC=0.967, MI=1.52, PSNR=20.91

TABLE II. RESULTS OF SOME EXPERIMENTS.

Intensity Adjustment	
 **NCC-based result**	 **MI-based result**
 **NCC-based result**	 **MI-based result**
Noise Addition	
 **NCC-based result**	 **MI-based result**
 **NCC-based result**	 **MI-based result**
Shearing	
 **NCC-based result**	 **MI-based result**
 **NCC-based result**	 **MI-based result**

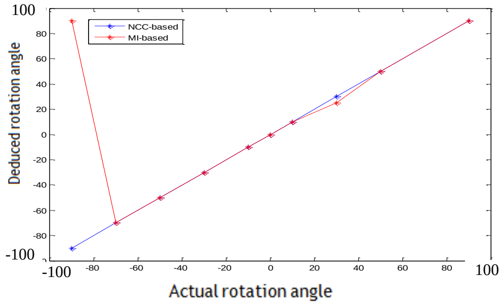


Fig. 13. Computed rotation angles versus actual rotation angles for test image 1.

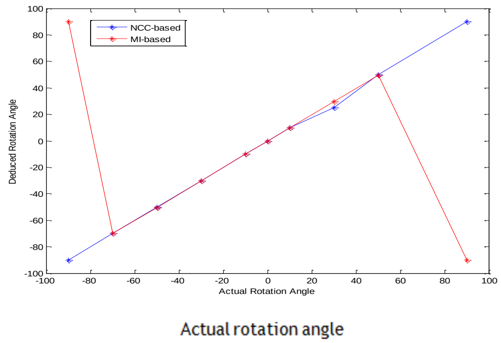


Fig. 14. Computed rotation angles versus actual rotation angles for test image 1 after intensity adjustment.

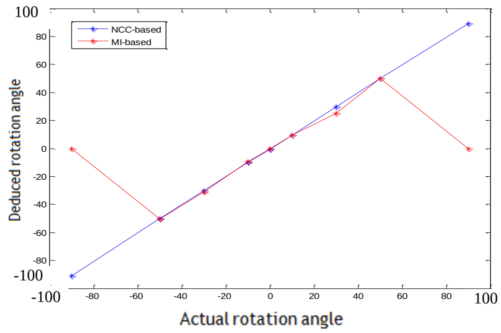


Fig. 15. Computed rotation angles versus actual rotation angles test image 1 after noise addition.

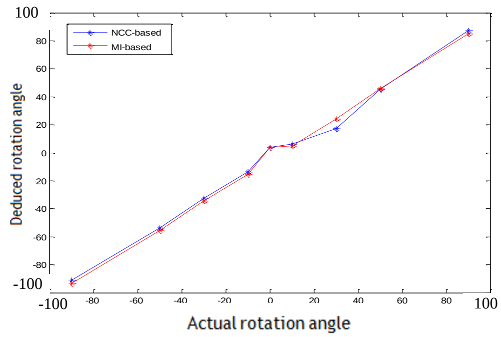


Fig. 16. Computed rotation angles versus actual rotation angles for test image 1 after shearing.

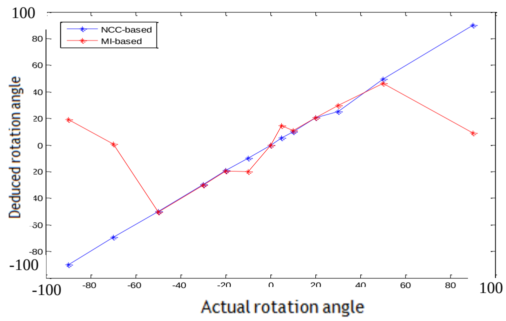


Fig. 17. Computed rotation angles versus actual rotation angles for test image 2.

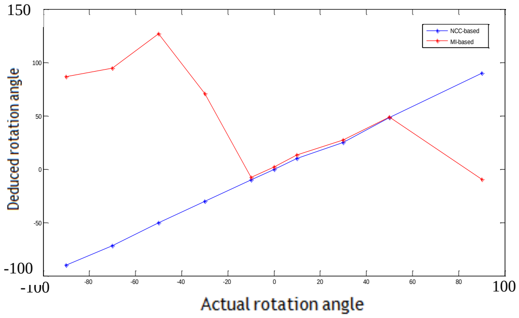


Fig. 18. Computed rotation angles versus actual rotation angles for test image 2 after intensity adjustment.

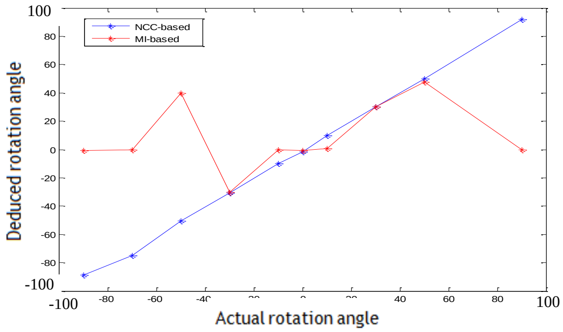


Fig. 19. Computed rotation angles versus actual rotation angles for test image 2 after noise addition.

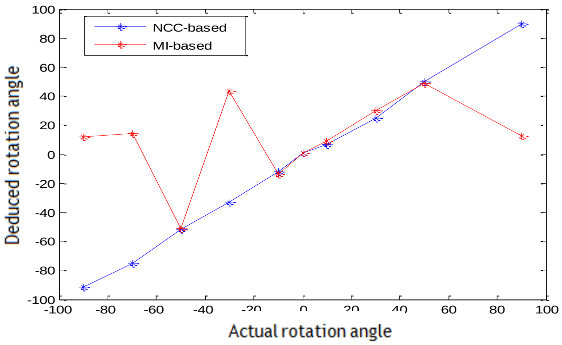
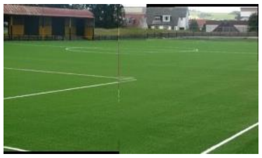
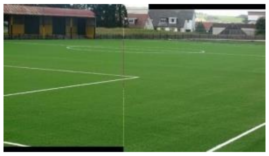
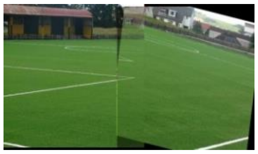
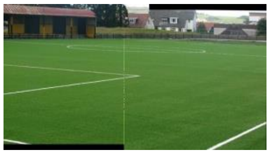


Fig. 20. Computed rotation angles versus actual rotation angles for test image 2 after shearing.

TABLE III. RESULTS BASED ON INDISTINCTIVE TEMPLATES.

Horizontal and vertical translation and zero rotation		
		 NCC-based result
		 MI-based result
Horizontal and vertical translation and 10° rotation		
 NCC-based result	 MI-based result	

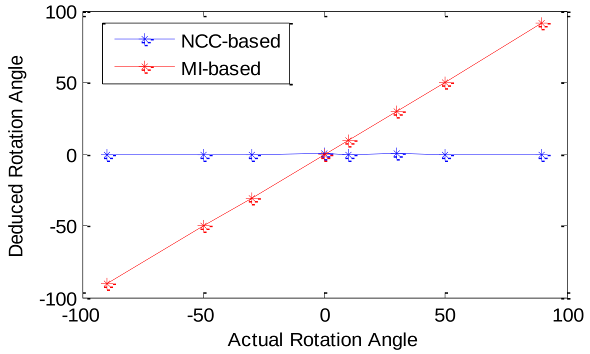


Fig. 21. Computed rotation angles versus actual rotation angles based on indistinctive templates.

IX. CONCLUSION

This paper has provided a comparative evaluation of the performance of template-based mosaicing techniques using two common similarity measures:normalized cross correlation (NCC) and mutual information (MI). Their performance has been tested at different rotation angles and under some un-modeled distortions and intensity variations.

The results indicate that NCC-based and MI-based mosaicing techniques have very close performance in many cases, but NCC-based performance is usually better for large rotation angles.

To provide faster results, small registration parameters were obtained using small image templates. When working with NCC-based mosaicing, steps should be taken to ensure that these templates are distinctive. Otherwise, the system may fail to provide reliable rotation angles. In the results shown here, the MI measure was less sensitive to the choice of image templates.

REFERENCES

[1] R. Szeliski, "Image alignment and stitching: A tutorial", Technical Report MSR-TR-2004-92, Microsoft Research, 2004.

[2] D. Milgram, "Computer methods for creating photomosaics",*IEEE Transactions on Computers*, C-24 (11):1113–1119, 1975.

[3] M. Brown and D. Lowe, "Recognising panoramas", Proc.9th Intl. Conference on Computer Vision, vol. 2, pp. 1218–1225, 2003.

[4] S. E. Chen, "Quick Time VR—An image-based approach to virtual environment navigation", Proc. SIGGRAPH'95, pp. 29–38, 1995.

[5] Realviz, http://www.realviz.com

[6] http://www.microsoft.com/products/imaging

[7] R. B.Inampudi, "Image mosaicing", Proc. IEEE Int.Geoscience and Remote Sensing Symposium(*IGARSS'98*), vol. 5, pp.2363 – 2365, 1998.

[8] H. Shum, and H. and R. Szeliski, "Construction and refinement of panoramic mosaics with global and local alignment", Proc. IEEE Intl. Conf. on Computer Vision, pp. 953-958, 1998.

[9] A. Rocha, R. Ferreira and A. Campilho, "Image mosaicing using corner detection", Proc. Ibero-American Symposium on Pattern Recognition, 2000.

[10] L. G. Brown, "Survey of imageregistration techniques", *ACM Computing Surveys*, vol. 24, no. 4, pp. 325-376, 1992.

[11] http://www.creatis.insa-lyon.fr/~srit/tete/2012_master_eeap_si_m5.pdf

[12] B.Zitováand J. Flusser, "Image registration methods: a survey", *Image and Vision Computing Journal*, vol. 21, no. 11, pp. 977-1000, 2003.

[13] P. Viola, "Alignment by maximization of mutual information", PhD thesis, MIT, 1995.

[14] A. Collignon, "Multi-modality medical image registration by maximization of mutual information", Ph.D. Thesis, Catholic University of Leuven, Leuven, Belgium, 1998.

[15] A. Ardeshir Goshtasby, *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*, John Wiley & Sons, 2005.

[16] D. A. Wismer and R. Chattergy, *Introduction to Nonlinear Optimization*, North Holland, 1978.

[17] A. Bevilacqua, A. Gherardi, and F. Piccinini, "Quantitative quality assessment of microscopic image mosaicing", *Journal of World Academy of Science, Engineering and Technology (WASET)*, vol. 47, no. 48, 283-286, 2010.

[18] E. Zagrouba, W. Barhoumi, and S. Amri, "An efficient image-mosaicing method based on multifeature matching," *Machine Vision and Applications*, vol. 20, pp. 139–162, 2009.

[19] http://www.vision.deis.unibo.it/MosPerf/MosPerf-Home.aspx

[20] http://vision.middlebury.edu/stereo/

Service Pyramid Concept of Knowledge Intensive Business Services in the Small and Medium Sized Enterprises Sector

Sándor Nagy
Institute of Business Sciences
University of Miskolc
Miskolc, Hungary

Abstract—**The activity of service companies is more complex today than before the use of Information and Communication Technology (ICT) became widespread in business. Fewerservice companies can be found on the market based solely on physical customer interaction; however, they have beenreplaced with standardized and automated services. This structural change can be observedparticularly among Knowledge Intensive Business Services (KIBS) companies where the use of ICT is not simply a market requirement, but a competition tool, as well. Without the application of any ICT instrument, a KIBS company could not meet market expectations, and furthermore, business developments are mainly ICT-based. On the other hand, companies need to develop their businesses more and more, as KIBS companies which are not able to differentiate themselves as trustworthy on the business services market will disappear. 'Just another business service company' cannot really sell today. This is applicable to a greater extent to Small and Medium Sized Enterprises (SME), where ICT has opened market opportunities that have never existed for them before. The aim of this paper is to highlight the new features of KIBS business activities inthe SME sector and the related research questions.**

Keywords—knowledge intensive business services; information and communication technology; small and medium sized enterprises; productisation; service product; service pyramid

I. Introduction

The network-like allocation of value creation in the open global economy contributed to the replacement of labour-intensive activities with knowledge intensive activities in the developed countries. This can easily be observed in the national statistical databases, but research programs and studies also deal with the contemporary economic questions of knowledge-based products and services. Derived from the most important features of services, this sector is highly receptive for technological innovations, which makes it a rapidly changing business area where prompt market reactions mean strategic advantage. During the last two decades, new companies emerged and grew to a global level based on Information and Communication Technology (ICT) innovations and finally they made available several high value business services for Small and Medium Sized Enterprises (SME), as well. In this paper a new service pyramid concept will be presented which can provide solutions for Knowledge Intensive Business Service (KIBS) companies to synthesize new features of services and the advantages of ICT spread in the SME sector in order to develop a successful market strategy.

II. New Features of Services

Earlier theoretical works determined services according towhat they are not (they are not agriculture and not industry) and their most important feature was that the production and consumption inseparably coincide with each other, while in case of the physical goods, a lot of time may pass between production, sales and consumption. The first modern approaches toservices were published after the 1950s, whenthe increasing role of complex services becamea main feature of the capitalist economies, which is the next stage in the economic development of industrial countries. Beside this *tertiarisation* in the developed economies, research isdiscovering new characteristics of services, indicatingthat serviceswill increasingly behave as physical products: in contrast with the earlier scientific view, more and more services can be kept in stock and can be mass-produced and mass-consumed. These factors point to a redefining of the economic role of services. Of course, it could be supposed that some services simply become physical products.[1]

Moreover, it can decreasingly be said that services only adoptto innovation. This finding come from analyzing Research and Development (R&D) expenditures, but this was not obviously the accurate conclusion. As the interpretation and classification of R&D expenditures of service companies was not quite precise –people both in surveys and in accounting basically were not sure what can be regarded as research activity in relation to service– research that operateswith thesenumbers may mislead with its conclusions that service companies do not innovate. In many companies, innovation costs are classified as R&D when they are linked with physical products or when outputs of developments are physically embodied. Empirical results have shown that service innovation is a hardly recognizable activity for companies [4]. On the other hand, in new publications, R&D expenditures of the service sector are similar to those ofthe process industry in developed countries. In the service sector, innovations are hard to reconcile with the traditional definitions for the following reasons:

[1] It is interesting that in contrast, physical products incorporate more services than before.

- they do not obviously result in new technologies or new technical solutions,
- they are mostly not tangible, which make them hard to measure,
- service innovation sources often come from customer relationships,
- they are combinations of in-house and exteriorfactors
- they are more difficult to protect with patents.

These problems require a wider interpretation of innovation in the service domain, where not only technological advances but also changing customer needs drive developments, and it is sourced in market cooperation.

Current service developments are clearly aimed to eliminate, or at least reduce, the inseparable nature of service delivery and consumption. With this, enterprises can significantly expand (or even wholly eliminate) the time constraints arising from inseparability. For example, if banking services can only be accessed duringopening hours, then this causes a large number of lost market interactions, while the ATMs (cash dispensers) or online banking services offer virtually non-stop availability. (The reduction of operation costs is an additional benefit.)

Research has shown that the use of information technology improves productivity and advancesglobalization in the service sector. In recent decades the global trade of services grew more intensely than the trade of goods, making the former scientific approach disputable. Researchers traditionally used to believe that services cannot be traded. This approach was first criticized already in the 1980s. Today, the global trade of services hardly needs to be proven; rather,it causes challenges to governments (due to e.g. taxation issues). Thus, services have become as globalized as physical products; this is mainly driven by ICT developments, knowledge transfer and innovation. In the last two decades, the spread of ICT catalyzed this process by offering new business operation models to service enterprises.

However, due to service development, companies have to face new challenges such as diversified needs of customers. Although service providers are interested in standardization and automation, ICT applications enable and allow them to offer solutions that satisfy the customers' individual needs,also. Similarly to the processing industry, mass production is combined with mass customization in the services market as well.

The idea behind mass customization is to be able to deliver services tailored to individual needs, while keeping the effectiveness and cost benefits of mass production. Considering the traditional characteristics of services, it can be observed that mass customization may ease the disadvantages of services: it reduces inseparability, helps systemize heterogeneity, makes tangible the intangible, improves expected quality and, where necessary, promotes long-term customer relationships. (As a relatively new phenomenon, customers are actively involved not just in the production but in the distribution process, as well.)

In order to fit in the market expectations and to utilize the spread of ICT in business use, service companies could innovate with new and commercially valuable goods which are more tangible. In the literature,related terms can be found as *commodification* and *productisation* referring to the activity where service companies provide more product-like solutions through systemization of their components. The first need for systemization emerged in the field of marketing, but now it is widespread. Newer research isfocusing mainly on formal service development processes with diverse stages (from idea generation to commercialization), which approach comes from the domain of manufacturing. Another study method is to model the service, which has a wider view and has the benefit of combining unique services with standard services based on the output perceptions of the customer. Modeling is built on the experiences of service processes as prerequisites with three equally important components:

- the *service concept* and structure based on customer needs,
- the *service process* as the prototype, and
- the*service system* which involves the resources to deliver [6].

As stated by the supporters of the modeling method, it allows developers to focus on customer needs. Problems emerge from gathering and selecting relevant customer information as an input to the model.

Despite the advantages of productisation,there are obstacles on the market. Companies are concerned about converting their knowledge into tangible products as they could be copied and stolen without any difficulty. In non-academic publications, consultants highlight the fact that complex knowledge products are not attractive to imitate and they are encouraging companies to use productisation bravely.

III. Information and Communication Technology in Knowledge Intensive Business Services

The above changes can also be observed in knowledge-intensive activities. The term KIBS in general refers to knowledge-intensive inputs provided for the business processes of business organizations, including both private and public organizations. Consequently, KIBS are not goods for private consumption. The term was first introduced by *Miles* in the academic literature, who proposed to distinguish between two groups in the KIBS sector: traditional professional services (P-KIBS) and advanced technology-based services (T-KIBS). The P-KIBS group includes traditional professional services that intensely apply the advanced technology (business and management services, legal and accounting activities, market research, etc.), while T-KIBS services are more closely related to the ICT sector, meaning that they are mainly focused on providing advanced technology-based services (IT services; engineering tasks; R&D consulting, etc.) [2].

The KIBS sector presented strong growth, and European studies have revealed several driving factors of this[1]:

- companies buy more solutions from outside their organizations (outsourcing and offshoring),

- increasing demand for technological knowledge (especially ICT related ones),
- increasing demand for specialized knowledge of social, administrative, and regulatory issues (compliance),
- internationalization and globalization issues,
- a growing role of services and intangible elements in production and in products,
- agrowing number of knowledge workers in labour markets.

The economic role of the KIBS sector is a central question in all academic publications as it emerges in the knowledge and innovation processes nationally and globally where KIBS companies are strong partners for other organizations. In this relation effective business services are focused on interaction between providers and customers in order to emphasize the connection in delivering processes. Keeping the traditional view of services where customers actively participate in the service production processes, co-production means

- interactivity in knowledge transfer and in learning,
- a tailor-made approach, where
- services necessarily start from scratch [6].

Co-production is essentially different from routine purchases, because during this process a client accesses to the expertise of the provider, rather than just gettinga commodity. This ledresearchers to debate standardization and automation opportunities in the KIBS market, but in this research concept it is not necessarily the only conclusion. Rather, it is better to perceive multiple ways to deliver a KIBS service while both customized and standardized solutionscan be observed. When it is proper to use a tailored approach or standardization isestablished as a more appropriate research question.This perception could much better explain how KIBS companies can utilize the penetration of new technologies.

ICT plays an important role in the life of KIBS companies, which can be explained by its digital nature, since ICT enables these companies to record, process, store and distribute the information which play vital role in producing the output of KIBS. Thus, the ICT enhances the classic service provisioning function by supporting operational processes, reducing costs and increasing the speed of operation. The integration of ICT developments into business processes has brought significant progress in collecting, storing and distributing business knowledge. As knowledge is the most important property for KIBS companies,howICT could help them use their knowledgeis a central question, whether that knowledgeis in the employee’s head or in documents. This process is called knowledge management, where knowledge could be differentiated as

- explicit, which is formal and independent from individuals (andcan be copyrighted), or
- tacit which is informal and dependent on individuals (andcannot be copyrighted).

With the spread of the ICT, a huge volume of explicit knowledge has become available for a wide range of companies – more than ever before. Considering the role of ICT in knowledge management, the companies’ aim is to convert tacit knowledge into explicit knowledge as much as they can. This often depends on the organizational culture. By formalizing the knowledge, KIBS companies advance the knowledge transfer which could be realized

- in customer interactions,
- through in-house relations,
- in sectorial interactions.

In addition to improving operations, the spread of ICT greatly promotes the opening and developing of new markets, because it can reduce the high transaction costs of knowledge-intensive activities. Some new KIBS services have appeared in the markets which have changed the opportunities of SME sector: certain services, which were previously overpriced, have become affordable for SMEs, and new services have also appeared. For example, it was not possibleearlier for a small company to plan and execute a targeted international advertising campaign. These days it is quite easy to do research in marketing databases and to set up fine-tuned campaigns as numerous companies offer real-time advertising services online for a relatively low price. Furthermore, instead of human agents, SME companies work with scheduled software workflows such as robots.As businesses apply ICT tools to support more and more business processes, the range of affordable KIBS services and mass customized services expands, since digitalization offers more opportunities for innovation. In this way, small organizations can have access to highly value-added inputs, which greatly improves their innovation capabilities, and, in turn, market competitiveness – even at global level.

Thus, the spread of ICT has brought geographical changes as well, since it eases the distribution of KIBS services, which may balance spatial inequalities and enable clients to access KIBS services irrespective of their geographical location. Resulting from this, KIBS service providers which have wide market coverage can operate not only in big cities, but can serve their clients from relatively peripheral areas. This phenomenon can be observed globally. The spread of ICT also supports the cooperation between KIBS companies physically located far away from each other, enabling the creation of new types of inter-organization collaborations and networks. The development and the possible innovation respectively of a given KIBS activity or of a KIBS company obviously are not so evident, because the necessary information and skills are still focused onthe relevant central economic areas, althoughto a lesser extent.[2]

[2] As the information required for KIBS related developments is digitalized and uploaded to the internet, the physical location of a KIBS company becomes less important. Due to the general cost reduction endeavors, the number of personal meetings between stakeholders in the KIBS sector has largely dropped; they may meet personally only a few times a year. For instance, broadband internet access enables webinars (seminars broadcasted via the web), an increasing portion of the literature is available online, and developers can also have online access to databases. Metropolitan headquarters are oftenmaintained for prestige reasons only, in many cases just as a virtual office.

The phenomena outlined above opens up entirely new prospects in smaller economies, as well, because SMEs can have a more balanced opportunity in market competition, which pushes companies to innovate, giving a new boost to their efficiency. Now ICT and the internet are not about simply having a website for KIBS companies; it is more about having an effective *digital presence*.

IV. Changes in Knowledge Intensive Business Services Concepts

Revealing structural and organizational changes and their triggersin the KIBS sector could lead to a better understanding of knowledge-based economies and globalization, where scientific concepts are focused on the fact thatthat knowledge has an increasing role in creating the economic value. It could also help policymakers work out regulations not only for national development, but the international trade, as well. During the last few decades the content of the knowledge-intensive work has changed, as organizations require more and more complex and prompt solutions from business service companies. This indicates parallel changes in service developments, where companies try to find cost effective answers for customer needs.

Regarding the KIBS sector, not only fast volume growthcan be observed, but animprovement in quality, as well, which comes mainly from

- development of their clients,
- a growing proportion of outsourced complex business works,
- development of knowledge-based societies,
- increasing regulations,
- internationalization and globalization,
- technological innovations.

The general sector-level impacts of ICT on services, from a scientific point of view, can be studied sooner in case of KIBS services, because the knowledge-intensity puts these companies onthe leading edge in adaptation to changes. KIBS enterprises are open to accepting ICT-related innovations and, equally importantly, they are involved in significant innovation activities. Given that KIBS companies which operate on the SME market are heavily pushed to improve their efficiency, it may be said that the spread of ICT is a great opportunity for them to both gain new markets and enhance their productivity.

ICT is assisting KIBS companies to codify certain parts of their knowledge and also helps them gather information about the relevant market cost-effectively with online research tools for sales and innovation purposes, serving several functions:

- multiplying occasions to get sales leads,
- strengthening and automating customer relationship management,
- providing a source forgathering information about customer demand and market competition to innovate,
- eases recognizing and adopting external developments,
- using online tools to develop more efficient workflows, and
- Achieving spatial independence.

The KIBS-related changes listed in the previous sectionscan easily be seen in the SME sector when observing their impacts on the service concepts of KIBS companies. This approach could be considered a proper research methodology, because in this research it allows one to reveal casual relationships. These trends have been integrated in the service provision concepts and are fundamental parts of the strategic planning to such a degree that failure to consider these would be a competitive disadvantage. Nowadays in the SME sector there is an increasingly lower demand for business services that are based on classic organizational and operational principles. The reason is simply because companies on the side of clients are changing in their behavior when operatingwith a KIBS company. This starts with the search phase, when they are looking for solutions on the market. Previously, decision makers initially asked around in their narrow environment, or there were printed and periodically updated catalogues to find a business provider; now nearly every search starts in a web browser.

Before introducing the new service concept, it is useful to compare the relevant features of the previous traditional (classic) KIBS services and the currently identifiable versions.

Based on the earlier theoretical works of services in Table 1,seven main features have been highlighted in order to draw attention tothe key characteristics of KIBS. Each feature could help KIBS companies as important innovators in service innovations identify developing areas. The comparison begins with splitting production and consumption, which introduces other new features because by their separation, the phenomenon of intangible goods arises, where other advantages could be derived: by digitalizing a service, it could be disconnected from humans on the provider side, whereby the possibilities of the digital worldcan be utilized. Here, the geographical and time limits disappear, which opens new opportunities for innovation. From the operational and policymaking points of view, the decreasing transaction cost could be a significant change, as it is an important factor both in micro and macro levels in economies.

TABLE I. Features of Classic and New Knowledge Intensive Business Services

feature	classic KIBS	new KIBS
production and consumption	inseparable	separable
transaction costs	high	low
uniqueness	customized, unique	customized, unique and mass customization
geographical location	centers	anywhere
access to SME sector	limited	limitless
innovation	mainly user, catalyzing other innovations	user, catalyst, and important innovator
trade	cannot be traded	could be traded

In order to access the SME sector as much as possible and to achieve spatial independence, KIBS companies could focus on mass customization as well,in addition to the fully customized versions. Thus, the new service concept derived from the characteristics described above is also shown in Fig. 1.

Fig. 1 demonstrates that due to the changes identified in Table 1, the new concept of KIBS services can be recognized as a pyramid, where benefits related to the spread of ICT are apparent.

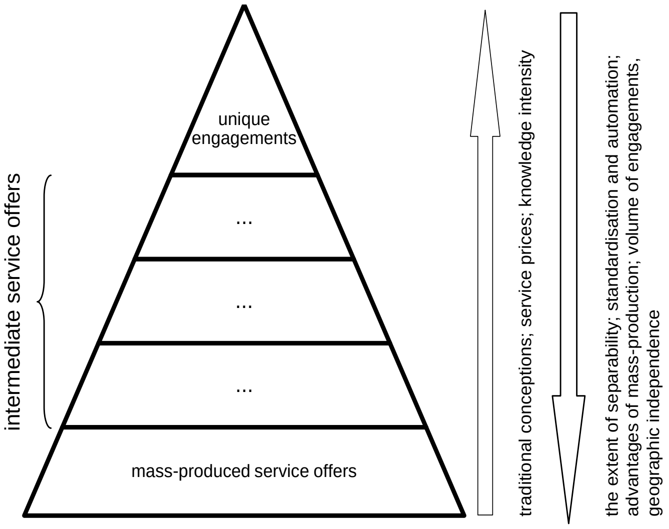

unique engagements
...
...
...
mass-produced service offers
intermediate service offers
traditional conceptions; service prices; knowledge intensity
the extent of separability; standardisation and automation; advantages of mass-production; volume of engagements, geographic independence

Fig. 1. Pyramid service conceptions of Knowledge Intensive Business Services.

The interpretation of the offering pyramid of KIBS services should start from its both ends: the bottom of the pyramid represents mass supply, while the top refers to absolutely customized services. Between the two extremes, there is an intermediate supply, which becomes more tailored as one movestowards the top of the pyramid. Consequently, the benefits of ICT (in terms of mass service delivery, standardization, automation and new types of services) impact the services at the bottom of the pyramid. As one movesdownwards from the top of the pyramid, services act more like a product, while those located in the top of the pyramid can be interpreted in the classic way.

In this research the '*service concept*' is seen as a high-level and more general term, compared to the term of '*business model*'. One maysay the business model can give operative tools that help implement the strategic service concept. KIBS service concepts can become a successful market strategy when a company is able to associate it with a logical, consistent, original, hard-to-copy business model.

Just asversioning is a well working business strategy for designing a product line of information goods, the *service pyramid concept* is a similarly proper approach for KIBS companies. Service pyramids let the service provider recommend different offers at a different level of complexity. KIBS companies are originally focused on providing complex and high quality version services. In a business manager's mind, a typical KIBS company is visualized as a well-fashioned, highly professional person, who is present himself in the manager's office and works hard with him to solve difficultbusiness problems. It is generally equal with the offer of any newly opened KIBS company. With service pyramids, a KIBS company usually creates the high-quality version first, and then as a top-down method they could subtract value from it to get lower-level services, which could be inserted in their product line as mass-market service products [5].

Additionally,it can be observed that in this complex system of service pyramids, companies can use modularity to develop new services. In the literature, modularity is presented as a design strategy that can stimulate market success and innovation mainly in ICT relation.

Modularity means a stable and clear architecture (both vertically and horizontally) with well-defined operation functions that reduce costs and decrease the uncertainty of ad-hoc work. Overall, modularity can be introducedinto services based on systemization, which helps moderate the disadvantages of heterogeneity. Service modularity is not only a source of effective operation but of a successful competition strategy, as well, where there is huge space for differentiation, especially in this intermediate time of new service strategies.

Note that in this perspective of KIBS services; they are similar to information products, as customers must experience themto value thembefore they decide to buy, so theycan be named '*experience goods*' as well. For SME companies, advertising their KIBS services in a certain market does not mean getting clients directly from this promotionalactivity. As KIBS services are highly complex intangible goods, their qualities are not easy to explain to the customer simply throughadvertisements [3]. Actually, the successful advertising strategy for KIBS companies is not only about concluding a final contract, as it could be considered nearly impossible. Who will engage a tax advisor to harmonize a company's tax structure just because of their billboard? Advertising is rather about thinking in a big and subtle network of marketing tools and channels. High professionals traditionally have to build their reputations on the market, which helps them get through the first sales obstacles originating from information asymmetry and risks of fears of the customers. Building reputation is about widely communicating business results, goals and merits etc., which means sharing information with customers and building a closer relationship to eliminate risk from the first business relationship. And for communication in business to business (B2B) relation, the ICT is the main (but, of course, not the only) channel to send out messages today.

V. Conclusions

It is put forwardthat ICT is a great help for KIBS companies in obtaining new customers, retaining old ones, selling more, and improvingand innovating services. Changes on the markets make it harder for companies to use the relevant information while the importance of tacit knowledge is rising. This all stimulates the combination of external and internal knowledge in SME organizations which providenew possibilities for KIBS companies with a high competition level. Moreover, since only a few start-ups survive, KIBS companies need to find new, profitable ways to innovate services utilizing the standardization and automation advantage of the trend that services tend to be rather intangible products.

The scope of this paper does not allow a detailed description of the KIBS service pyramid concept presented inFig. 1, but this was not the goal of this research. The objective here is to highlight a research direction that is worth following in order to have a better understanding of the subject matter. The short history of research into knowledge-intensive business services has some very exciting questions left unanswered, which will be answered in the course of further research. For example, it is important to collect some empirical evidence about organization sizes, because the new KIBS service pyramid concept described here may effectively impact the SME sector in the near future.

As a next research steps it is neededto set up relevant KIBS clusters internationally (in developed and emerging markets) to see how these companies are utilizing the service pyramid concept and productisation in their market strategies (between leaders and followers), then benchmark their business performance with the market and with each other. This could give empirical evidence for the effective use of the service pyramid concept, both ata micro and macroeconomic level. Itis also expected that additional research areas will emerge, especially regarding the process of service productisation.

References

[1] EMCC, European Monitoring Centre on Change, “Sector futures - The knowledge-intensive business services sector,” European Foundation for the Improvement of Living and Working Conditions, Dublin, 2005.

[2] I. Miles, “Knowledge Intensive Business Services: Prospects and policies,” Foresight 7(6), pp. 39-63, 2005.

[3] C.C.J.M. Millar, and C.J. Choi, “The innovative future of service industries: (anti-) globalization and commensuration,” The Service Industries Journal, Vol. 31, (1) pp. 21–38, 2011.

[4] F. J. Mosoniné, M. Tolnai, and A. Orisek, Research and development and innovation in the service sector. Kutatás-fejlesztés és innováció a szolgáltatási szektorban. Nemzeti Kutatási és Technológiai Hivatal, Budapest, 2004.

[5] C. Shapiro, and H. R. Varian, Information rules: a strategic guide to the network economy. Harvard Business School Press, Boston, 1998.

[6] K. Valminen, and M. Toivonen, “Seeking efficiency through productisation: a case study of small KIBS participating in a productisation project,” The Service Industries Journal, Vol. 32, No. 2, pp. 273-289, 2012.

Software Development Effort Estimation by Means of Genetic Programming

Arturo Chavoya, Cuauhtemoc Lopez-Martin, M.E. Meda-Campaña
Department of Information Systems
University of Guadalajara
Guadalajara, Mexico

***Abstract*—In this study, a genetic programming technique was used with the goal of estimating the effort required in the development of individual projects. Results obtained were compared with those generated by a statistical regression and by a neural network that have already been used to estimate the development effort of individual software projects. A sample of 132 projects developed by 40 programmers was used for generating the three models and another sample of 77 projects developed by 24 programmers was used for validating the three models. Results in the accuracy of the model obtained from genetic programming suggest that it could be used to estimate software development effort of individual projects.**

Keywords—genetic programming; feedforward neural network; software effort estimation; statistical regression

I. INTRODUCTION

The estimation of how long it takes to develop specific software projects is an ongoing concern for project managers [1]. The software development effort estimation can begin with individual projects within academic environments [2], as is the case in this study. There are several techniques for estimating development effort, which could be classified into: 1) expert judgment that aims at deriving estimates based on the experience of experts on similar projects [3][4]; 2) those based on models such as a statistical regression model [5][6]; and 3) those based on techniques from computational intelligence [7], such as fuzzy logic [8][9], neural networks [10] and genetic programming [11].

Considering that no single estimation technique is best for all situations, and that a careful comparison of the results from several approaches is most likely to produce realistic estimates [12], this study compares estimates generated with a genetic programming model against the results obtained with a neural network and with the most commonly used model: statistical regression[4].

Data samples for this study were integrated by 132 and 77 projects for generating (verifying) and validating the models, respectively, and were developed by 40 and 24 programmers, respectively. All of the projects were created following practices of the Personal Software Process (PSP) [13].

The three models were generated from data of small projects individually developed using practices of PSP because this approach has proven its usefulness when applied to individual projects [2].

The hypothesis of this research is the following: Prediction accuracy of a model based on genetic programming is statistically better or equal than a statistical regression model or a model obtained with a feedforward neural network, when these three models are generated from two kinds of lines of code and are applied to the prediction of software development effort of individual projects that have been developed with personal practices.One reason for choosing genetic programming in this work was that this technique is capable of modeling non-linear behaviors, which are common when correlating independent variables with the development effort of software projects [14].

The rest of the paper starts with a section describing the genetic programming algorithm used to generate the corresponding model, followed by a section with the related work. The next section presents the methods used for evaluating the three models, followed bya section on the generation of the models. Respective sections on the verification and validation of the models are presented next. The paper ends with a section of conclusions.

II. GENETIC PROGRAMMING

Genetic programming (GP) is a field of evolutionary computation that works by evolving a population of data structures that correspond to some form of computer programs [15]. These programs typically represent trees varying in shape and size where the internal nodes correspond to functions and the leaves represent terminals such as constants and variable names. The trees can be implemented as the list-based structures known as S-expressions, with sublists representing subtrees.

Fig. 1 presents the flowchart followed by a typical implementation of the GP algorithm [15]. The GP algorithm starts with a population of M randomly generated programs consisting of functions and terminals appropriate to the problem domain. If the termination criterion has not been reached, each program is then evaluated according to some fitness function that measures the ability of the program to solve a particular problem.The fitness function typically evaluates a problem against a number of different fitness cases and the final fitness value for the program is the sum or the average of the values of the individual fitness cases. GP normally works with a standardized fitness function in which lower non-negative values correspond to better values, usually with zero as the best value.

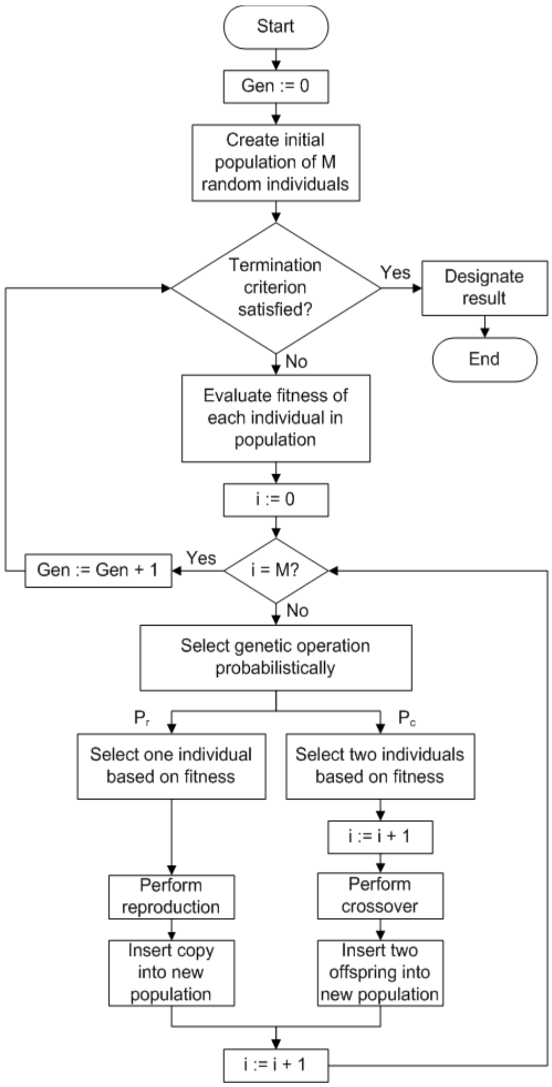

Start
Gen := 0
Create initial population of M random individuals
Termination criterion satisfied?
Yes
Designate result
End
No
Evaluate fitness of each individual in population
i := 0
Yes
Gen := Gen + 1
i = M?
No
Select genetic operation probabilistically
Pr
Pc
Select one individual based on fitness
Select two individuals based on fitness
i := i + 1
Perform reproduction
Perform crossover
Insert copy into new population
Insert two offspring into new population
i := i + 1

Fig. 1. Flowchart followed by a typical implementation of the GP algorithm.Symbols are as follows: Gen = Generation counter; i = Individual counter; M = Population size; Pr= Probability of reproduction; Pc = Probability of crossover.

After all programs in the population have been evaluated, a selection is made among the individuals in the population to produce the next generation. This selection is usually made proportionate to fitness so that programs with better fitness values have a higher probability of being selected. The Darwinian selection of the fittest individuals in the population is the biological basis on which the various evolutionary computation paradigms are inspired. A number of operations can be applied to selected individuals to provide for variability in the new generation. The reproduction operation consists of selecting a fixed percentage of individuals to pass unchanged to the next generation according to a certain probability of reproduction (Pr). In the crossover operation, two individuals are selected according to a probability of crossover (Pc) to function as parents to produce two offspring programs. In each of the parents a node in the corresponding trees is selected randomly to constitute a crossover point.

The subtrees that have the selected nodes as roots are then exchanged generating two new individuals that are usually different from their parents.

Fig. 2 shows an example of two parental trees before crossover, with the corresponding S-expression below each tree; arrows point at the root nodes of the subtrees chosen to be exchanged, with the corresponding subexpressions shown in boldface.

Fig. 3 presents the generated offspring trees resulting from the exchange of the subtrees in Fig. 2 whose root nodes are pointed at by the arrows. The exchange of subtrees corresponds to the exchange of the sublists shown in boldface below each tree.

A fixed portion of the next generation is produced using the crossover operation, having the possibility of forcing that a fixed percentage of the selected nodes correspond to functions, whereas the rest correspond to either functions or terminals. Unlike genetic algorithms, the mutation operation is normally not necessary in GP, as the crossover operation can provide for point mutation when two nodes corresponding to terminals in the parents are selected to be exchanged.

The process of evaluating, selecting and modifying individuals to produce a new generation is continued until a termination criterion is satisfied. The GP run usually terminates when either a predefined number of generations has been reached or a desired individual has been found.

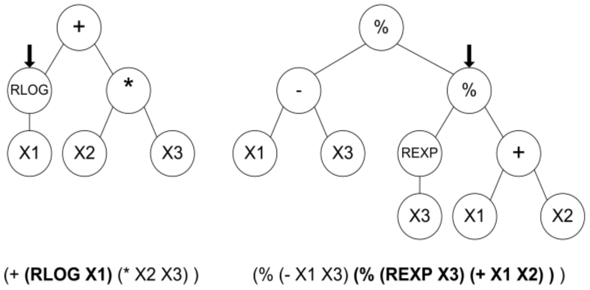


Fig. 2. Example of two parental trees before crossover and the corresponding S-expressions.

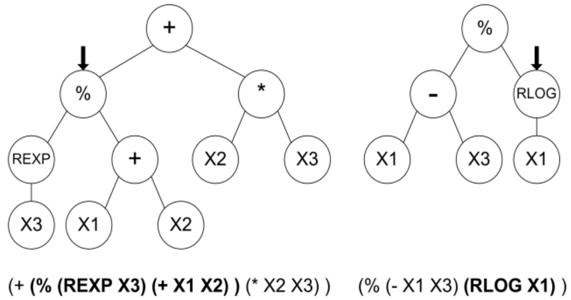


Fig. 3. Offspring trees after crossover and the corresponding S-expressions.

III. RELATED WORK

Results from the application of neural networks and statistical regression have shown that the estimation accuracy of both techniques are competitive with models generated from data of large projects [16][17][18], and of small projects [19].

The accuracy of the genetic programming model used in the present work is compared against the accuracies obtained from the neural network and the multiple linear regression models described in [19]. These two models were generated using data from small-scale projects. The kind of neural network used was a feedforward multi-layer perceptron with a backpropagation learning algorithm (the most commonly used in the effort estimation field [20]). The feed-forward neural network used the Levenberg-Marquardt algorithm due to its reported efficiency [21].

Genetic programming has already been applied to large projects; however, we did not find any study related to its application for predicting the software development effort of small projects developed in laboratory learning environments [22]. Some of the methods reported in previous publications resemble the approach taken in the present work, in which a mathematical model that best fits the data is searched.

The main difference of the present work with most previous reports lies in the genetic programming parameters they used and the data on which they applied the genetic programming algorithm. In [10] a GP algorithm was implemented having a population size of 1000 individuals reproducing for 500 generations during only 10 runs. They used a dataset of 81 software projects that a Canadian software company developed in the late 1980s. They suggested that the genetic programming approach needed further study to fully exploit its advantages. On the other hand, in [23] GP was used with the goal of comparing the use of public datasets against company-specific ones. The techniques they used (GP, artificial neural networks and multiple linear regression) were slightly more accurate with the company-specific database than with publicly available datasets. They used the same GP parameters as in [10]. They concluded that companies should base effort estimates on in-house data rather than on public domain data. In [24] GP was compared against artificial neural networks and multiple linear regression using a number of publicly available datasets. Using less individuals in the GP population (from 25 to 50) than normally employed in the typical implementation of the algorithm (several hundred), they found that although GP was better at effort prediction than neural networks and multiple linear regression with some datasets, in general, none of the techniques they tested rendered a good effort estimation model. These authors concluded that the datasets used to build a prediction model had a great influence in the ability of the model to provide adequate effort estimation. In [25] a different approach was used with GP; instead of finding the mathematical model that best fitted the data, they developed a grammar-based technique they called Grammar Guided Genetic Programming (GGGP) and compared it against simple linear regression. They used the data of 423 software development projects from a public repository and randomly divided them into a training set of 211 projects and a test set of 212 projects. The results obtained using the GGGP technique were not very encouraging, as the effort prediction they found was not very accurate. In [26] GP was also applied for predicting the effort of large projects, and their results showed that GP was better than case-based reasoning and comparable with statistical regression. Finally, GP was applied in [27] using the same methodology as in the present work, but the model found had a slightly higher validation MMER than the model presented here.

IV. METHODS

In this study, the independent variables for all three models were New and Changed (N&C) as well as Reused code, and all of them were considered as physical lines of code (LOC). N&C is composed of added and modified code. The added code is the LOC written during the current programming process, whereas the modified code is the LOC changed in the base project when modifying a previously developed project. The base project is the total LOC of the previous projects, whereas the reused code is the LOC of previously developed projects that are used without any modification [13]. Source lines of code represent one of the two most common measures for estimating software size [28]. Finally, the dependent variable Effort was measured in minutes.

The accuracy criterion for evaluating models in this work was the Magnitude of Error Relative to the estimate for observation *i*, or MER*i*, defined as follows:

$$MER_i = \frac{\left|\text{Actual Effort}_\text{i} \text{ - } \text{Estimated Effort}_\text{i}\right|}{\text{Estimated Effort}_\text{i}}. \quad (1)$$

The MER value is calculated for each observation *i* whose effort is estimated. The aggregation of MER over multiple observations can be achieved through the mean (MMER).

Another criterion that has been used in the past for evaluating prediction models is the Magnitude of Relative Error (MRE), which is calculated for the *i*-th observation as follows:

$$MRE_i = \frac{\left|\text{Actual Effort}_\text{i} \text{ - } \text{Estimated Effort}_\text{i}\right|}{\text{Actual Effort}_\text{i}}. \quad (2)$$

The mean of MRE over multiple observations is denoted as MMRE.

The accuracy of an estimation technique is inversely proportional to the MMER or the MMRE. It has been reported that an MMRE ≤ 0.25 is considered acceptable [29]; however, no studies or argumentations supporting this threshold value have been presented [30].Results of MMER in [31] showed better results when compared to other studies; this fact is the reason for choosing MMER as evaluation criterion in the present work.

Experiments for this study were done within a controlled environment having the following characteristics:

- All of the developers were working for a software development company. However, none of them had previously taken a course related to personal practices for developing software at the individual level.
- All developers were studying a graduate program related to computer science.

- Each developer wrote seven project assignments. Only the last four of the assignments of each developer were selected for this study. The first three projects were not considered because they had differences in their process phases and in their logs, whereas the last four projects were based on the same logs and had the following phases: plan, design, design review, code, code review, compile, testing, and postmortem.
- Each developer selected his/her own imperative programming language whose coding standard had the following characteristics: each compiler directive, variable declaration, constant definition, delimiter, assign sentence, as well as flow control statement was written in a line of code.
- Developers had already received at least a formal course on the object oriented programming language that they selected to be used through the assignments, and they had good programming experience in that language. The sample for this study only involved developers whose projects were coded in C++ or JAVA.
- Because this study was an experiment with the aim of reducing bias, we did not inform the developers about our experimental goal.
- Developers filled out a spreadsheet for each project and submitted it electronically for examination. This spreadsheet contained a template called “Project Plan Summary”, which included the completed data by project. This summary had actual data related to size, effort (time spent in the development of the project) and defects. This document had to be completed after each project was finished.
- Each PSP course was given to no more than fifteen developers.
- Since a coding standard establishes a consistent set of coding practices that is used as a criterion for judging the quality of the produced code [13], the same coding and counting standards were used in all projects. The projects developed during this study followed these guidelines. All projects complied with the counting standard shown in Table I.
- Developers were constantly supervised and advised about the process.
- The code written in each project was designed by the developers to be reused in subsequent projects.
- The kind of the developed projects had a similar complexity as those suggested in [13], and all of them required a basic knowledge of statistics and programming topics learned in the first semesters of an undergraduate program. From a set of 18 individual projects, a subset of seven projects was randomly assigned to each of the programmers. Description of these 18 projects is presented in [19].
- Data used in this study are from those programmers whose data for all seven exercises were correct, complete, and consistent.

TABLE I. COUNTING STANDARD.

Count type	**Type**
Physical/logical	Physical
Statement type	**Included**
Executable	Yes
Non-executable	
Declarations	Yes (one by text line)
Compiler directives	Yes (one by text line)
Comments and Blank lines	No
Delimiters:	
{ and }	Yes

V. GENERATION OF MODELS

Data from 132 individual projects developed by 40 programmers from the year 2005 to the year 2008 were used in the three models: GP, neural network and multiple linear regression. The projects that contained reused code were selected for the sample.

A. Multiple Linear Regression

The following multiple linear regression equation was generated [19]:

$$E\phi\phi o\rho\tau=45.06+(1.20*N\&X) \text{ □ } (0.30*P\epsilon\upsilon\sigma\epsilon\delta). \qquad (3)$$

The intercept value of 45.06is the value of the line where the independent variables are equal to zero. On the other hand, the signs of the two parameters comply with the following assumptions related to software development:

- The larger the value of new and changed code (N&C), the greater the development effort.
- The larger the value of reused code, the lesser the development effort.

An acceptable value for the coefficient of determination is $r^2 \geq 0.5$ [13],with this equation having an r^2 equal to 0.58. The ANOVA for this equation had a statistically significant relationship between the variables at the 99% confidence level and the two independent variables were statistically significant at the 99% confidence level.

B. Neural Network

There is a variety of tasks that neural network can be trained to perform. The most common tasks are: pattern association, pattern recognition, function approximation, automatic control, filtering and beam-forming. In the present work, a feedforward neural network with one hidden layer was applied for function approximation. This network had already been trained to approximate a function [19]. The effort was considered as a function of two variables: N&C (number of new and changed lines of code) and Reused (number of reused lines of code).

It has been shown that a feedforward network with one layer of hidden neurons is sufficient to approximate any function with a finite number of discontinuities on any given interval [21]. This is the reason for using a fully-connected feedforward neural network with one hidden layer of neurons in this work. The fully-connected part of the description means that each neuron in a layer receives a signal from each of the neurons in the preceding layer. There were two neurons in the input layer of the network: one received the number of N&C lines of code and the other received the number of reused lines of code. The output layer consisted of only one neuron indicating an estimated effort. The number of neurons in the hidden layer was empirically optimized. A range from two to 40 neurons was explored and the best results were obtained with ten neurons in the hidden layer. The optimized Levenberg-Marquardt algorithm was used to train the network.

The network passed through two phases: training and application. The first group of 132 software projects was used to train the network. This group of projects was randomly separated into three subgroups: training, validation and testing. The training group contained 70% of the projects. The input-output pairs of data for these projects were used by the network to adjust its parameters. The next 20% of data was used to validate the results and identify the point at which the training should stop. The remaining 10% of data was randomly chosen to be used as testing data, to make sure that the network performed well with the data that was not presented during the parameter adjustment.

C. Genetic Programming

A LISP implementation of the GP algorithm was used for generating a model to predict software development effort. The following standard parameters were used on all runs [15]: the initial population consisted of 500 S-expressions randomly generated using the ramped half-and-half generative method. In this method, an equal number of trees are created with a depth that ranges from 2 to the maximum allowed depth (6 in this work) for new individuals. For each depth, half of the programs corresponded to full trees, and the other half consisted of growing trees of variable shape. The maximum depth for individuals after the application of the crossover operation was 17. The reproduction rate was 0.1, whereas the crossover rate was 0.7 for function nodes and 0.2 for any node. Finally, each GP run was allowed to evolve for 50 generations and the individual with the best fitness value was selected from the final generation.

The set of terminals was defined by the two independent variables X1 and X2 corresponding to the New & Changed and Reused lines of code, respectively. Additionally, terminals also consisted of floating-point constants randomly generated from the range [-5, 5).

The set of functions consisted of the arithmetic operators for addition (+), subtraction (–) and multiplication (*), along with the following protected functions shown in prefix notation. To avoid division by zero, the protected division % was defined as follows:

$$(\%\quad x\quad y)=\begin{cases}1 & y=0\\ x/y & y\neq 0\end{cases}. \tag{4}$$

To account for non-positive variable values, the protected logarithmic function RLOG was defined as

$$\left(RLOG\quad x\right)=\begin{cases}0 & x=0\\ \ln|x| & x\neq 0\end{cases}. \tag{5}$$

Finally, the protected exponential function REXP was defined as

$$\left(REXP\quad x\right)=\begin{cases}0 & |x|\geq 20\\ e^{x} & |x|<20\end{cases}, \tag{6}$$

where the boundary value 20 was arbitrarily chosen to avoid over- and underflows during evaluation.

Since the standardized fitness function f is required to consist of non-negative values, with zero as the best match, this function was defined as

$$f=\sum\left|\text{Actual Effort}_{\text{i}}\ \text{- Estimated Effort}_{\text{i}}\right|. \tag{7}$$

The MMER value was not considered an appropriate fitness measure, as the denominator in the MER formula can yield negative values if the estimated effort in the LISP model is negative itself.

Fifty experiments each consisting of 1,000 GP runs were made. From each experiment, the run with the highest fitness value (lowest *f* value) was selected and finally an individual program from all runs was selected according to how well it predicted software development effort on both the verification and validation data sets. The selected program from the 50,000 runs is presented next in LISP notation, where X1 is New and Changed code, and X2 is Reused code.

```
(- (- (+ (- X1 (* (- X2 X2) 3.7990248)) (REXP 3.7627742))
    (% (+ (* 2.2606792 X1) (- X2 -4.461488)) (+ (+ X1 X2) X1)))
  (% (+ (% X2 (+ X1 2.2606792)) (- 4.497994 X1))
    (+ (% (% X2 (% -1.1173002 X2)) (+ X1 X1)) (+ X1 2.2606792))))
```

After evaluation of constant subexpressions and simplification of additions involving subexpressions evaluating to zero, the next equivalent program was obtained.

```
(- (- (+  X1 43.06774)
    (% (+ (* 2.2606792 X1) (- X2 -4.461488)) (+ (+ X1 X2) X1)))
```

(% (+ (% X2 (+ X1 2.2606792)) (- 4.497994 X1))

(+ (% (% X2 (% -1.1173002 X2)) (+ X1 X1)) (+ X1 2.2606792))))

VI. Verification of Models

The GP, the multiple linear regression equation, and the neural network models were applied to the original dataset of 132 projects for estimating effort; then their accuracy by project (MER), as well as by model (MMER), were calculated giving the following results for MMER:

- Genetic Programming = 0.25
- Multiple Linear Regression = 0.26
- Neural Network = 0.25

The following three assumptions of residuals for MER ANOVA were analyzed:

- Independent samples: in this study, groups of developers are made up of separate programmers and each of them developed their own projects, rendering the data independent of each other.
- Equal standard deviations: in a plot of this kind the residuals should fall roughly in a horizontal band centered and symmetrical about the horizontal axis (as shown in Fig. 4).
- Normal populations: a normal probability plot of the residuals should be roughly linear (as shown in Fig. 5).

Once these three residual assumptions had been met, the ANOVA for MER of the projects was calculated, which showed that there was not a statistically significant difference among the prediction accuracy for the three models (*p*-value of Table II is greater than 0.05).

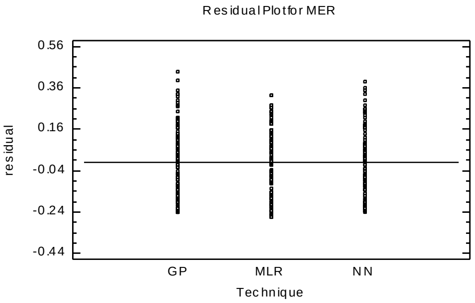


Fig. 4. Equal standard deviation plot of MER ANOVA – verification stage.

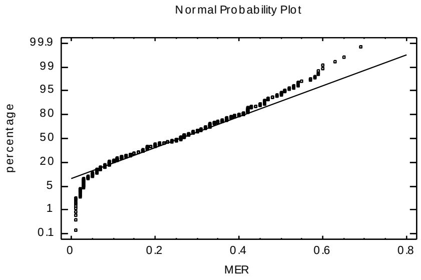


Fig. 5. Normality plot of MER ANOVA– verification stage.

TABLE II. ANOVA Table for MER by Model (verification)

Source	Sum of squares	Degrees of freedom	Me an square	F-ratio	p-value
Between groups	0.0317	2	0.0158	0.60	0.5488
Within groups	10.3644	392	0.0264		
Total	10.3962	394			

VII. Validation Of Models

Another group of developers consisting of 24 programmers developed 77 projects through the year 2009. These projects were developed using the same standards, logs, and following the same processes as the 132 programs used for generating the models presented in Section V. Once the three models for predicting effort were applied to these data, the MER by project as well as the MMER by model were calculated yielding the following MMER values:

- Genetic Programming = 0.23
- Multiple Linear Regression = 0.24
- Neural Network = 0.22

An ANOVA for the MMER models (Table III) showed that there was not a statistically significant difference among the accuracy of prediction for the three models (*p*-value is greater than 0.05) at 95% of confidence. Fig. 6 and Fig. 7 show how ANOVA residuals assumptions as described in the previous section are met.

TABLE III. ANOVA Table for MER by Model (validation).

Source	Sum of squares	Degrees of freedom	Me an square	F-ratio	p-value
Between groups	0.045	2	0.0226	1.00	0.3703
Within groups	5.0962	225	0.0226		
Total	5.1414	227			

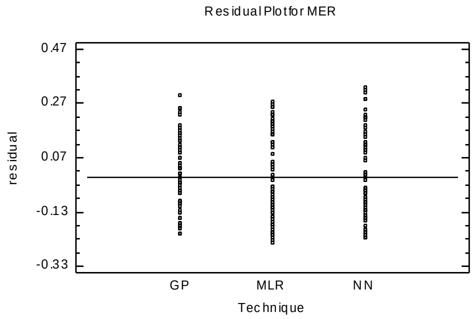


Fig. 6. Equal standard deviation plot of MER ANOVA – validation stage.

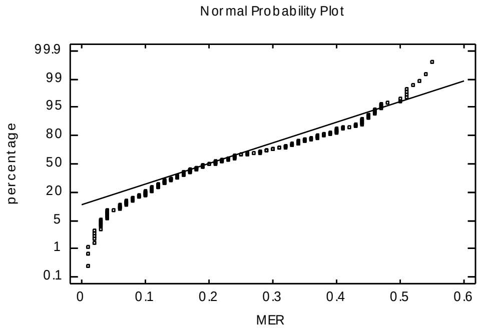


Fig. 7. Normality plot of MER ANOVA – validation stage.

VIII. Conclusion

Taking into account that no single estimation technique is best for all situations, this study compared a GP model with the results obtained from a neural network, as well as from statistical regression.

Data samples integrated by 132 and 77 individual projects for verifying and validating the three models were developed by 40 and 24 programmers, respectively. All the projects were developed following the same practices from the personal software process. The independent variables used in the models were New & Changed code as well as Reused code, whereas the dependent variable was the effort measured in minutes.

The accepted hypothesis in this study was the following: prediction accuracy of a genetic programming model is statistically equal to those obtained from a feedforward neural network, and from a statistical regression model when these three models are generated from two kinds of lines of code and they are applied for predicting software development effort of individual projects that have been developed with personal practices.

Even though we found that the three estimation techniques we tested had a similar power of prediction, GP can have an advantage over other techniques in those cases where specific non-linear functions are suspected to be part of the prediction function, as GP allows the use of any function desired, and the final solution can be a composition of the selected functions.

Future research involves the application of genetic programming for estimating the effort of individual projects involving more independent variables and larger datasets.

Acknowledgment

The authors of this paper would like to thank CUCEA of Guadalajara University, Jalisco, México, ConsejoNacional de Ciencia y Tecnología (Conacyt), as well as Programa de MejoramientodelProfesorado (PROMEP).

References

[1] Jørgensen, M., T. Halkjelsvik, T.: The effects of request formats on judgment-based effort estimation. The Journal of Systems and Software, 83, 29–36 (2010)

[2] Rombach, D., Münch, J., Ocampo, A., Humphrey, W.S., Burton, D.: Teaching disciplined software development. Journal of Systems and Software, 81, 747- 763 (2008)

[3] López-Martín, C., Abran, A.: Applying expert judgment to improve an individual's ability to predict software development effort. International Journal of Software Engineering and Knowledge Engineering 22(4): 467-484 (2012)

[4] Jørgensen, M.: Forecasting of software development work effort: Evidence on expert judgment and formal models. Journal of Forecasting, 23(3), 449-462 (2007)

[5] Boehm, B., Abts, C., Brown, A.W., Chulani, S., Clark, B.K., Horowitz, E., Madachy, R., Reifer, D., Steece, B.: COCOMO II. Prentice Hall (2000)

[6] Kok P., Kitchenhan, B.A., Kirakowski, J.: The MERMAID approach to software cost estimation. In: Proceedings ESPRIT (1990)

[7] Pedrycz, W.: Computational intelligence as an emerging paradigm of software engineering. In: ACM 14th International Conference on Software Engineering and Knowledge Engineering, pp 7-14 (2002)

[8] Lopez-Martin, C., Yañez-Marquez, C., Gutierrez-Tornes, A.: Predictive accuracy comparison of fuzzy models for software development effort of small programs. Journal of Systems and Software, 81(6), 949-960 (2008)

[9] Lopez-Martín, C: A fuzzy logic model for predicting the development effort of short scale programs based upon two independent variables. Journal of Applied Soft Computing, 11(1), 724-732 (2011)

[10] Wen, J., Li, S., Lin, Z., Hu, Y., Huang, C.: Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 54, 41–59 (2012)

[11] Burguess, C.J., Lefley, M.: Can genetic programming improve software effort estimation? A comparative evaluation. Journal of Information and Software Technology, 43, 863-873 (2001)

[12] Boehm, B., Abts, C., Chulani, S.: Software development cost estimation approaches: A survey. Journal of Annals of Software Engineering, 10(1-4), 177-205 (2000)

[13] Humphrey, W.S.: A discipline for software engineering. Addison Wesley (1995)

[14] Hsu C.J., Huang C.Y.: Comparison of weighted gray relational analysis for software effort estimation. Software Quality Journal, 19(1) 165-200 (2011)

[15] Koza, J.R.: Genetic programming: On the programming of computers by means of natural selection. The MIT Press (1992)

[16] Lopez-Martin C., Isaza C., Chavoya A.: Software development effort prediction of industrial projects applying a general regression neural network. Journal of Empirical Software Engineering, 17(6) 738-756 (2012)

[17] De Barcelos Tronto, I.F., Simoes da Silva, J.D., and Sant'Anna, N.: An investigation of artificial neural networks based prediction systems in software project management. Journal of Systems and Software, 81(3), 356-367 (2008)

[18] Heiat, A.: Comparison of artificial neural network and regression models for estimating software development effort. Journal of Information and Software Technology, 44(15), 911-922 (2002)

[19] Lopez-Martin, C.: Applying a general regression neural network for predicting development effort of short-scale programs. Journal of Neural Computing and Applications, 20(3), 389-401 (2011)

[20] Park, S.: An empirical validation of a neural network model for software effort estimation. Journal of Expert Systems with Applications, 35, 929-937 (2008)

[21] Haykin, S.: Neural Networks: A comprehensive foundation. Prentice Hall (1998)

[22] Afzal, W. and Torkara, R.: On the application of genetic programming for software engineering predictive modeling: A systematic review. Journal of Expert Systems with Applications, 38(9), 11984-11997 (2011)

[23] Lefley, M., Shepperd, M.J.: Using genetic programming to improve software effort estimation based on general data sets. LNCS, 2724, 2477-2487 (2003)

[24] Dolado, J.J., Fernández, L.: Genetic programming, neural networks and linear regression in software project estimation. In: International Conference on Software Process Improvement, Research, Education and Training, pp 157-171 (1998)

[25] Shan, Y., McKay, R.I., Lokan, C.J., Essam, D.L.: Software project effort estimation using genetic programming. In: Proceedings of the IEEE 2002 International Conference on Communications, Circuits and Systems, 2, pp. 1108-1112 (2002)

[26] Ferrucci, F., Gravino, C., Oliveto, R., Sarro, F.: Genetic programming for effort estimation: An analysis of the impact of different fitness functions. In: The 2nd International Symposium on Search Based Software Engineering, pp. 89-98 (2010)

[27] Chavoya, A., Lopez-Martin, C., Meda-Campaña, M.E.: Applying genetic programming for estimating software development effort of short-scale projects. In: IEEE 2011 Eighth International Conference on Information Technology: New Generations (ITNG 2011), pp. 174-179 (2011)

[28] Sheetz, S.D., Henderson, D., Wallace, L.: Understanding developer and manager perceptions of function points and source lines of code. Journal of Systems and Software, 82, 1540–1549 (2009)

[29] Conte S.D., Dunsmore H.E, Shen V.Y.: Software engineering metrics and models. Benjamin/Cummings Pub Co. (1986)

[30] Jørgensen, M.: A critique of how we measure and interpret the accuracy of software development effort estimation. In: The First International Workshop on Software Productivity Analysis and Cost Estimation (SPACE'07). Information Processing Society of Japan, pp. 15-22 (2007)

[31] Foss, T., Stensrud, E., Kitchenham B., Myrtveit I.: A simulation study of the model evaluation criterion MMRE. IEEE Transactions on Software Engineering, 29(11), 985-995 (2003)

An analysis of Internet Banking in Portugal:

the antecedents of mobile banking adoption

João Pedro Couto
Business and Economics Department
University of Azores
Ponta Delgada, Portugal

Teresa Tiago
Business and Economics Department
University of Azores
Ponta Delgada, Portugal

Flávio Tiago
Business and Economics Department
University of Azores
Ponta Delgada, Portugal

Abstract—**In recent years, mobile operations have gained wide popularity among mainstream users, and banks tend to follow this trend. But are bank customers ready to move forward? Mobile banking appears to be a natural extension of Internet banking, . Thus, to predict consumer decisions to adopt mobile banking, it's useful to understand the pattern of adoption of Internet banking (IB). This investigation seeks contribute to an expansion of the knowledge regarding this matter by researching Portuguese consumers' patterns and behaviors concerning the acceptance and use intention of IB as a foundation for establishing growth strategies of mobile banking. For data collection, we used an online "snowball" process. The statistical treatment used included a factor analysis in order to allow examination of the interrelationships between the original variables. The analysis was made possible by developing a set of factors that expresses the common traits between them. The results revealed that the majority of respondents did not identify problems with the use of and interaction with the IB service. The study generated some interesting findings. First, the data generally supports the conceptual framework presented. However, some points need to be made: (i) trust and convenience, from all the elements referenced in the literature as relevant from the client' perspective, continue to be a very important elements; (ii) the results did not support the paradigm that the characteristics of individuals affect their behavior as consumers, (iii) individual technological characteristics affect consumer adoption of IB service; (iv) consumer perceptions about the IB service affect their use, as reveal by the existence of three types of customers that show different practices and perceptions of IB; and (v) intention to use IB is dependent upon attitudes and subjective norms regarding the use of IB.**[1]

Keywords—internet banking; mobile banking; technology adoption models; Portugal

I. Introduction

Traditionally, banking was a simple branch-based operation and therefore the intensive street presence was considered a success critical factor. Over the past three decades, the technological evolution path impact the way services processes and the bank sector was no exception in the use of multiple technologies and applications used on daily base activities. The ICT adoption brings along a more and complete services offer, with high levels of security involve and cost reduction. For these reasons, more and more banks were investing in this form of customer interaction. Internet banking has created a paradigm shift, enriching banks offers and taking advantage of customers' predisposition to engage into a virtual relationship with their banks.

However, with the internet banking market increasing competition driven by consumer expectations and technology developments, IB left is central role and become a launching platform for the latest IT-driven bank-offer: mobile banking.

The literature review reveals that many studies were (and currently are being) conducted on the adoption of mobile banking by customers primarily on their perceptions about cost reduction, ease of use and convenience, reliability, and lastly but not the least, security and privacy. It also refers that these costumers had past experience with internet banking solutions. Therefore, this paper identifies the different segments of IB customers, paying special attention to trust and convenience as key concepts of consumer behavior, since they are common influences identified in IB and mobile-banking.

In this study, therefore, building upon prior research regarding consumer adoption and use of internet banking we try to identify the antecedents of mobile banking adoption, considering that the intention to use mobile banking is impacted by past IB experiences.

For practical reasons, this study focuses on the Portuguese consumers' patterns and behaviors' concerning the acceptance and use intention of internet banking. The data was gathered online and in order to test the hypothesis, a set of multivariate statistical analysis were performed.

The rest of this paper is organized as follows: Section 2 discusses the related literature reviewed for this research study; the subsequent sub-sections outline the conceptual model and experimental hypothesis on which the model is based; presents methodology and the discuss the empirical findings; Section 5 describes concludes the paper's results. The implications for industry as well as for research and limitations and scope for future research have been discussed in last Sections 6 and 7 respectively.

II. Research Background

It seems very cliché to start a work on internet defining the different paths of evolution of the home banking. However, when looking to this evolution and the adoption patterns follow by clients and its implications in banks strategies, it seems quite useful to quickly review this processes.

In the last two decades, the topic of home banking has gained its own space both in academics and business circles [1-

Acknowledgment: Funding for this work is granted by FCT – CEEAplA, Research Center for Applied Economics and data gathered by Ricardo Borges, to whom we thank for the sharing.

3]. However, home banking is a concept with a history greater than it appears to be, since it started with sales over phone process, followed by ATM and "dial-up" computers access and nowadays it's an umbrella combining of all of these with internet banking and mobile banking.

The explosion of internet usage and the huge funding in ICT initiatives, allowed the design of internet banking services offers, overcoming the spatial and time constrain of banking services, since it provides a 24/7 and global coverage [4].

The concept internet banking refers to the use of the internet as a remote delivery channel for banking services [5], allowing clients to access their bank and bank account and perform almost all the different types of transactions available through internet [4].

Since the nineties, the number of households with Internet access has increased dramatically, offering new markets possibilities for internet-based services such as internet banking [6]. In general, Europe has been and still is the leader in internet banking technology and usage. Particularly, in Portugal most banks offer internet banking solutions (See, BESnet was created in 1998 by BancoEspirito Santo).

There are two main reasons, listed in most research works, to traditional financial institutions engage internet banking activities. The first is related to lowering operational costs. The second regards, improving consumer banking services, increasing retaining consumers' rate and expanding consumer' share. Back in 1999, [4] suggested that IB segment was the most profitable business unit. More recent evidences confirm his conclusions and added that besides been more profitable this business model retained loyal and committed consumers when compared with traditional banking (ABA, 2004; Fox, 2005). For all these reasons, banks have recognized the importance to differentiate themselves from other financial institutions through new distribution channels. If in the last decade, banks were investing in internet banking.

Nowadays, mobile banking has been rapidly gaining popularity as a potential medium for electronic commerce. However, the diffusion of this channel is still in an early stage.

Both systems operate over the internet and enables customers 24 hour 7 day access to their account, and allows customers to conduct more complicated transactions, such as pay bills, applying for housing loan applications, online shopping, account consultation, and stock portfolio management.

The work of Dhungel, Acharya, &Upadhyay-Dhungel [7] remembers that banks invest in electronic channels to take advantage of its unique features as universal applicability, more speed to conduct transactions and less financial costs. But points to the existence of different stages of adoption of digital channels.

These authors also defend that when a new innovation becomes commercial feasible, the adoption of this new technology by potential users leads to its diffusion. They stressed on the importance of identifying a list of customers' profiles that will embrace new technologies in the earliest stages of its implementation.

Although, the development of the electronic banking supply, the number of mobile banking users is still very weak in comparison with the other e-banking services, such as internet banking and ATMS. Therefore, there is a need to understand the factors that influence intention to use mobile banking. Following a similar approach of the Dhungel, Acharya, &Upadhyay-Dhungel [7] work, it's relevant to determinate the key customer-specific factors that will predict actual digital behavior.

Few studies have focused on mobile value from the distinctive feature of a mobile technology perspective as an internet continuum technology, with specific customers' expectations [6]. Novel perspectives point to the replication of patterns from e-commerce to mobile solutions [8, 9].

In recent years, a multiplicity of theoretical perspectives have been applied to provide an understanding of the determinants of Internet banking adoption and use and more recently to understand mobile banking [1, 7, 10-13].

With the growing acceptance of internet banking, a constant analysis of customer behavior is needed, considering the factors affecting its adoption. In this field, often behavioral models are used, such as theory of reasoned action (TRA) [14] or the technology acceptance model (TAM) [15]. More recent studies employing a TAM-base theoretical lens have identified additional constructs that may be influential in internet banking service adoption: (i) online consumer behavior and online service adoption (channel knowledge, convenience, experience, perceived, accessibility and perceived utility; time savings; site waiting time; security, privacy and trust; cost; service quality); (ii) service switching cost (procedural, financial and relational); (iii) adoption of internet banking (convenience, service quality, perceived relative advantage, compatibility, trialability, complexity, demographics, gender, consumer attitudes and beliefs, security, privacy, trust, risk, needs already satisfied, familiarity, habit, convenience, adaptability, computer and technology confidence, knowledge and high levels of internet use at work) [14, 16-22].

With increasing technology adoption, bank performance is progressing steadily, therefore customer service is the area with major improvements made, which is a major advantage for users [19].

Regardless the importance of the subject, the investigation regarding European bank services is still scarce. When looking to the development of banking industry is clear that Portuguese banks are in the upstream of IT based solutions use. The questions that remain unanswered are how customers feel regarding internet banking and what influences their acceptance and adoption processes, including the adoption of mobile solutions.

III. Framework and Hypothesis

Given the broad range of contributing theories and factors identified in the literature regarding banking in internet and mobile era, internet banking consumer behavior needs to study with a different approach to these factors. As Venkatesh et al. [21] recalled, the theory of reasoned action (TRA) from Fishbein and Ajzen [14] is still one of the most prominent theories regarding human behavior. According to this approach, behavior intention relies in two constructs: (i) the attitude toward behavior, which is "the positive or negative feelings about an individual's adoption of a particular behavior" (Fishbein&Ajzen, 1975, p. 216); (ii) and the subjective norm that is consider as “the perception of a person that most people who are important to him think he should or should not behave in a certain way "(Fishbein&Ajzen, 1975, p. 302). Since attitude resides in the mind, precedes and produces behavior and thus can be used to predict behavior (Yang and Yoo, 2004), it will reflect the personal characteristics of the individuals, such as gender, age or profession.

The principles of the theory of reasoned action were applied by Davis et al. [15] in the acceptance of IT by the individual, showing similar trends to other domains where TRA was applied. Driven from TRA, these authors present an instrument to predict the likelihood of a new technology being adopted within a group or individuals: technological acceptance model (TAM). Regardless the criticisms and the limitations found in TAMs, the model has been used in numerous studies seeking to gauge the determinants of behavior adoption and use of new technologies.

The positive results obtained from the use of this model in internet banking context (Daniel and Storey, 1998; Eriksson, Kerem, & Nilsson, 2005; Hernandez &Mazzon, 2007; Laukkanen, Sinkkonen, &Laukkanen, 2008; Mols, 1999; Prompattanapakdee, 2009; Qureshi & Khan, 2008; TeroPikkarainen, Pikkarainen, Karjaluoto, &Pahnila, 2004) and its non-application to Portuguese banking environment were the driving reasons for considering this model as the basis of their work. However, it was felt necessary to make some modifications, considering the inputs of the IBAM (internet banking adoption model) proposed by Alsajjan and Dennis [23] in order to enrich the model and a better match to the Portuguese reality. Therefore, special attention was given to trust and convenience as key concepts of online consumer behavior. For instance, from a psychology or relationship marketing perspective trust is the key element of relationships and in internet context has been found as important factor in the adoption decision process, especially in the internet banking context [3].

In light of the above discussion, the following hypotheses were constructed for testing in this study:

H1: Attitude towards the use of internet banking is dependent upon the behavioral beliefs relative to trust and convenience.

H2: Attitude towards the use internet banking is dependent upon individual demographic characteristics.

H3: Attitude towards the use internet banking is dependent upon individual technological characteristics.

H4: Intention to use internet banking is dependent upon attitudes towards the use of internet banking and subjective norms about the use of internet banking.

The hypotheses described reflect the study aims to explore the self-reported behaviors of online customers and their intention to use internet banking services. From the literature review that considers internet banking as the previous stage of mobile banking, it’s expectable that the profiles found indicate future users of mobile banking [9].

IV. Methodology and Results

After the extensive literature survey, the research methodology has been centered on the already identified existing core variables. Therefore, and in order to validate the assumptions set, we applied a methodology consisting of four phases, namely, sample definition (1), questionnaire development (2), data collection (3), and statistical analysis (4). Based on the theorize model developed in the course of a detailed review of the related literature on user acceptance of technology and new technology diffusion, a questionnaire was compose as a measurement scale for the research, including two sections: (i) the first addresses issues of socio-demographic features of respondents, and (ii) the second adds a set of closed questions that capture differences in the perceptions of respondents to the use and adoption IB.

This study focuses on the Portuguese consumers for practical reasons and also because Portugal is one of the countries with more intensive electronic banking systems. Adding to it, the last two decades reflect a growing adoption of ICT in Portugal. According to the Eurostat, Portugal has an internet penetration rate larger than 61% and the Portuguese are the Europeans who most use the Internet to access bank services. However, the mobile banking penetration is still undersized, turning Portugal a good field for exploring the antecedents of mobile banking as a natural path after internet banking.

For data collection we used an online “snowball” process. The statistical treatment used, included a factor analysis in order to allow to study the interrelationships between the original variables, by developing a set of factors which expresses the common traits between them. Based on these factors we grouped the respondents according to their behavior relatively to the IB activities, carrying out a cluster analysis. To test the hypotheses we then applied tests of multiple means differences of to determine the existence of distinct patterns between the groups obtained in cluster analysis.

The final sample obtained comprises a set of 277, where 193 were males and 84 females, with 44% of the individuals between the ages of 25 to 34 years. The results revealed that the majority of respondents did not identify problems with the use and interaction with the IB service. Since the fear that consumers feel about the safety of online transaction is one of the biggest inhibitors of IB use [24], we tried to check the perception that the respondents had on the security provided by IB service of their bank, with the majority of respondents demonstrating confidence in the service using IB.

In order to test the hypothesis, a set of multivariate statistical analysis were performed. First, a factor analysis was applied, to reduce the number of variables on customer’s perceptions of the service.

The results permitted the extraction of two factors, representing 63.41% of the total variance explained. The suitability of the technique is supported by the statistical significance of Bartlett's test and Kaiser-Meyer-Olkintest (KMO). Considering the variables associated with each component, the first factor was designated by “Trust” and reflects the way users perceive confidence and confidentiality and data security, the second factor, was designated by “Convenience”, and is mainly associated with ability to perform all kind of operations, faster and at any given time.

TABLE I. FACTOR ANALYSIS

Rotated Component Matrix(a)	1	2
Offers na trustable service	0,870	0,245
Ensures clients privacy	0,865	0,264
Is a safe storage system	0,833	0,312
Has a positive reputation	0,813	0,211
Offers conveniente location	0,155	0,744
Offers conveniente hours	0,395	0,709
Offers the necessary operations	0,368	0,707
Offers more rapid transactions	0,233	0,676
Offers convenient information	0,182	0,625
Reduces the costs of transactions	0,120	0,624

With the factors found, a cluster analysis was performed, using the K-means method. The solution showed three clusters, as can be seen in the following table. Given the factors found with higher incidence in these three clusters they were designated: Intermediate users (clusters 1), Full Users (cluster 2) and Basic users (cluster 3) (Fig. 1)

TABLE II. CLUSTER ANALYSIS

Dimension/ Cluster	**Intermediate Users (n= 105)**	**Full Users (n= 130)**	**Basic Users (n= 42)**
Trust	,240	,323	-1,599
Convenience	-,961	,760	,050

Aiming to determine the existence of differences in how individuals perceive the activity of Internet Banking, based on their socio-demographic characteristics, a chi-square test was used crossing the cluster membership and these variables, namely sex, age, location, education background and profession. The results suggest the absence of significant differences between individuals with different demographic characteristics.

The existence of differences arising from the level of technological expertise of the individual was also analyzed, in line with that suggested by Eriksson et al. [25]. The results suggest the existence of significant differences in how respondents use the IB, depending on the experience level of Internet use, and IB use and the intensity of usage.

To understand how the three clusters obtained, are different from each other, in relation to other key issues, a variance analysis was performed as well as a set of tests of multiple comparisons of means. Regarding the attitude towards IB services we found that the respondents included in group 2 (full users), show a more favorable attitude on the various dimensions analyzed of attractiveness, sensibleness, beneficial and valuable service usage.

TABLE III. ANALYSIS OF VARIANCE: ATTITUDE

	Sum of squares	df	Mean Square	F	Sig.	Means Difference
Attractive to customers	35,333	2	17,667	39,602	0	2>1,3
A sensible option	38,15	2	19,075	40,541	0	2>1,3
Beneficial to customers	37,208	2	18,604	41,96	0	2>1,3
A good idea	34,748	2	17,374	46,966	0	2>1,3
An option with added value	26,636	2	13,318	12,895	0	2>1,3

We also asked respondents about the advantages, type of transactions performed online and the process of utilization. The results showed significant differences among the three groups, and the full users cluster showed a higher rating of this items when compared to other two groups.

V. DISCUSSION AND CONCLUSIONS

With recent advances in information and communication technologies, internet-base commerce is having an increasingly profound impact on our daily lives, offering appealing and advantageous new services. As Cheng, Lam and Yeung [26] suggested banks are taking advantage of these opportunities and Internet banking is widely seen as the key and most popular delivery channel for banking services in the cyber age. The growing importance of IB is highly associated to cost reduction, but especially to the benefits in customer relationship.

From the customers viewpoint, the decision to use the IB may be motivated by convenience and ease of use, but to these may be added financial gains, since many institutions practice lower prices for their IB services, as well as some of the features inherent in the use of the Internet itself: easily and quickly access the information, available 24/7, and ubiquity. This suggests that there are notable trends in the use of mobile banking as an extension of internet banking, since the systems is supported by internet and the benefits are quite similar [9].

However, despite the wide range of attributes, not all customers of banking institutions adhere to these type of service and those who do not always employ the same reasons or use it with the same intensity.

Therefore, the aim of the work is to gain awareness of the various reasons explaining why Portuguese consumers are or not becoming internet banking users and to determine their profiles of use, since the most intensive users can be potential mobile clients. Based on a random sample of internet users, demographic, attitudinal, and behavioral characteristics of Internet banking (IB) users and non-users were examined.

The results confirm the first hypothesis which states that perceived usefulness and ease of use influence the intention and actual use of IB solutions.

The cluster analysis carried out revealed the existence of three types of clients that demonstrate distinct practices and perceptions of IB: basic users (1), intermediate users (2) and full users (3).

Basic users are characterized by a lack of confidence in the IB service to ensure their safety, while intermediate users, although do not appreciating the ease of use IB, are individuals who feel that the service is safe, and full users I. Users highlight the full indulgence that IB provides them, also have confidence in the security presented in IB, while the basic users

Based on the results obtained in the multivariate analyzes, we can infer significant differences in the way individuals perceive the use of IB and vary your level of experience of IB. We can observe that there are marked differences between the three clusters, and full users stand out for having an positive attitude to IB and intended use superior.

The second hypothesis was based on the concept that the characteristics of individuals affect their behavior as consumers, namely their demographic characteristics.

The results contradict this hypothesis and conclusions presented by SadiqSohail and Shanmugham [27]. Thus, we could not find empirical substantiation, in our sample, to validate the influence of demographic characteristics of individuals, such as gender, age, education, in the adopting the IB services.

The third hypothesis had as a reference the dimensions that have been extensively studied with regard to IB consumer behavior, and from which there is no reference for the Portuguese context, that is the technology acceptance by consumers. This measure the technological characteristics of consumers effect on the adoption of Internet Banking services.

In this case the data support the hypothesis, and the results show that the way individuals relate to the technological conditions its performance in terms of Internet Banking.

VI. Contributions and Implications

The results generated some interesting findings. First, data support in general the conceptual framework presented. However, some mentions need to be made: (i) trust and convenience, from all the elements referenced in the literature as relevant from the client' perspective, continue to be a very important elements [2, 3, 28-30]; (ii) the results did not support the paradigm that the characteristics of individuals affect their behavior as consumers, which contradicts the conclusions of Sadiq and Shanmugham [27]; (iii) individual technological characteristics affect consumer adoption of internet banking service; (iv) consumer perceptions over the Internet Banking service affect their use, as reveal by the existence of three types of customers that show different practices and perceptions of IB; and (v) intention to use internet banking is dependent upon attitudes towards the use of internet banking and subjective norms about the use of internet banking. These last results are consistent with the literature, in particular the models of adoption of new technologies proposed by Eriksson and Such (2008) and by Alsajjan and Dennis [23]. These last findings also suggest that the intention to engage in a higher-technology relationship with the bank is dependent on past experiences in the digital context.

The main theoretical contributions of this study highlight is the importance of the evidence that the components associated with the TAM model - the perception of usefulness and ease of use are relevant in IB adoption.

For the sample analyzed, it was possible to demonstrate that perceived usefulness and ease of use is a determinant condition in the intention and actual use of IB solutions. It was also possible to verify the existence of the influence on the intensity of use of the IB solutions.

This research has, however, not supported the conclusions of past studies about the influence of the personal characteristics of individuals in the adopting the IB. We could not find validation for the concept that the intrinsic characteristics of consumers, such as age, sex, educational level and income affect the intensity and how they adhere to Internet Banking.

Once the demographic characteristics, contrary to what happened in most other studies, emerge as not influencing the compliance behavior, it seems urgent to deepen the knowledge of IB customers in order to make possible the development of services and communication strategies for more customized Internet banking services.

It was also demonstrated, for the sample analyzed, that of all the elements referenced in the literature as relevant in the perspective of the client, the trust remains a highly valued element as well as ease of use. Therefore the banks need to seek solutions that are relevant in a user's perspective to minimizing uncertainty and maximizing the use.

The bank branches and their own interactions with clients continue to be a key element, but banks should not neglect the service provided to customers in the digital environments.

Considering the aspects mentioned above, we can conclude that this work may contribute to the identification of new target segments and that the results obtained may be taken into account in the development of future communication campaigns and digital offers.

Given the nature of its operations and services, the banking sector emerged as an area prone to dissemination and adoption of technological innovations. Since the nineties that has witnessed the proliferation of home banking activities and more recently the Mobile and Internet Banking.

The use of the Internet as a channel of distribution and communication can be considered another step in the assimilation of technologies that began with the appearance of ATM machines. But if for banks using the IB has advantages for clients can also be a wide range of benefits.

From the point of view of the customer, the decision to use IB may be driven by convenience and ease of use. To these may be added the financial gains , since many institutions practice tables lower prices for their services IB , as well as some of the features inherent in the use of the Internet itself : easily and quickly access the information , available 24/7 and, ubiquity .

It appears, however, that despite the wide range of attributes, not all customers of banks adhere to this type of service and that they do not always employ the same reasons or with the same intensity.

VII. Limitations and Future Research

Some useful preliminary insights are produced, however, leaving a considerable number of issues for future research, providing scholars with an opportunity to conduct further research in this field and practitioners with an opportunity to enhance adoption rates based on consumer behavior knowledge.A limitation of this study is the sample size, while it has sought to achieve a larger sample, financial and time constraints of a research prevented the full achievement of initial objective. Considering this situation, the sample was limited to a lower set of participants that somehow share a similar educational and cultural background. Future work should seek to extend the sample, as well as minimizing the effects of educational and cultural proximity.

It would also be interesting to assess the existence of attitudes and patterns of behavior, in the face of technologies, at the different regions of the country, which would allow the evaluation of the impact of the development of technological infrastructure on a regional basis and the effect on in the local customer's behavior.

Furthermore, the scope of the study could be broadened to include the comparison of buying behavior in physical versus virtual banking services, or even consider the inclusion of mobile banking.

From the standpoint of credit institutions, it also appears as an important aspects the assessment of the effect of substitution of channels, interactive communication and even loyalty and overall satisfaction of stakeholders. A line of future research could include the assessment of the impact of this type of service in the overall performance of banks.

Additionally, despite the Portuguese consumer's exhibit preferences similar to those observed in other developed countries, there are still elements that require a deeper future analysis.

References

[1] H. Hoehle, E. Scornavacca, and S. Huff, "Three decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis". Decision Support Systems. 54(1): p. 122-132. 2012.

[2] B. Suh and I. Han, "Effect of trust on customer acceptance of Internet banking". Electronic Commerce Research and Applications. 1: p. 247-263. 2002.

[3] B. Suh and I. Han, "The Impact of Customer Trust and Perception of Security Control on the Acceptance of Electronic Commerce". International Journal of Electronic Commerce. 7(3): p. 135-161. 2003.

[4] N.P. Mols, "The Internet and the banks' strategic distribution channel decisions". International Journal of Bank Marketing. 17(6). 1999.

[5] S. Liao, et al., "The adoption of virtual banking: an empirical study". International Journal of Information Management. 19(1): p. 63-74. 1999.

[6] R.W. Bons, et al., "Banking in the Internet and mobile era". Electronic Markets. 22(4): p. 197-202. 2012.

[7] A. Dhungel, B. Acharya, and K. Upadhyay-Dhungel, "Perception of Bank Customers about Automated Teller Machine (ATM) Service Quality". Banking Journal. 2(2): p. 23-38. 2012.

[8] C. Srinuan, P. Srinuan, and E. Bohlin, "An analysis of mobile Internet access in Thailand: Implications for bridging the digital divide". Telematics and informatics. 29(3): p. 254-262. 2012.

[9] G. Kim, B. Shin, and H.G. Lee, "Understanding dynamics between initial trust and usage intentions of mobile banking". Information Systems Journal. 19(3): p. 283-311. 2009.

[10] M. Kazemi and A. Kariznoee, "Prioritizing factors affecting bank customers using kano model and analytical hierarchy process". International Journal of Accounting and Financial Management-IJAFM. 6. 2013.

[11] T. Teo and J. Noyes, "Exploring attitudes towards computer use among pre-service teachers from Singapore and the UK: A multi-group invariance test of the technology acceptance model (TAM)". Multicultural Education & Technology Journal. 4(2): p. 126-135. 2010.

[12] S. Prompattanapakdee, "The Adoption AND Use OF Personal Internet Banking Services IN Thailand". The Electronic Journal on Information Systems in Developing Countries. 37(6): p. 1-31. 2009.

[13] C.M. Matei, C. Silvestru, and D.Ş. Silvestru, "Internet Banking Integration within the Banking System". RevistaInformaticaEconomică nr. 46(2). 2009.

[14] M. Fishbein and I. Ajzen, Belief, attitude, intention and behaviour: An introduction to theory and research: Addison-Wesley. 1975.

[15] F.D. Davis, R.P. Bagozzi, and P.R. Warshaw, "User acceptance of computer technology: a comparison of two theoretical models". Management Science. 35(8): p. 982-1003. 1989.

[16] P. Gerrard, J.B. Cunningham, and J.F. Devlin, "Why consumers are not using internet banking: a qualitative study". Journal of Services Marketing. 20(3): p. 160-168. 2006.

[17] J.B. Pinho, O poder das marcas. vol. 53: Summus Editorial. 1996.

[18] J.H. Wu and S.C. Wang, "What drives mobile commerce?:: An empirical evaluation of the revised technology acceptance model". Information & Management. 42(5): p. 719-729. 2005.

[19] T.M. Qureshi and M.B. Khan, "Customer Acceptance of Online Banking in Developing Economies". Journal of Internet Banking and Commerce. 13(1). 2008.

[20] TeroPikkarainen, et al., "Consumer acceptance of online banking: an extension of the technology acceptance model". Internet Research. 14(3): p. 224-235. 2004.

[21] V. Venkatesh, et al., "User acceptance of information technology: Toward a unified view". MIS Quarterly. 27(3): p. 425-478. 2003.

[22] H. Yang and Y. Yoo, "It's all about attitude: revisiting the technology acceptance model". Decision Support Systems. 38(1): p. 19-31. 2004.

[23] E.M. Serra and J.A.V. González, A marca: avaliação e gestão estratégica: Verbo. 1998.

[24] W. Chung and J. Paynter. "Privacy issues on the Internet". in 35th Hawaii International Conference on System Sciences. Hawaii: Computer Science. 2002.

[25] K. Eriksson, K. Kerem, and D. Nilsson, "Customer acceptance of internet banking in Estonia". 23(2): p. 200-216. 2005.

[26] T. Cheng, D.Y.C. Lam, and A.C.L. Yeung, "Adoption of internet banking: an empirical study in Hong Kong". Decision Support Systems. 42(3): p. 1558-1572. 2006.

[27] M. SadiqSohail and B. Shanmugham, "E-banking and customer preferences in Malaysia: an empirical investigation". Information Sciences. 150(3-4): p. 207-217. 2003.

[28] S. Grabner-Krauter and R. Faullant, "Consumer acceptance of internet banking: the influence of internet trust". International Journal of Bank Marketing 26(7): p. 483-504. 2008.

[29] M. Kivijärvi, T. Laukkanen, and P. Cruz, "Consumer trust in electronic service consumption: a cross-cultural comparison between Finland and Portugal". Journal of Euromarketing. 16(3): p. 51-65. 2007.

[30] G. Milne and M. Boza, "Trust and concern in consumers' perceptions of marketing, information management practices". Journal of Interactive Marketing. 13(1): p. 5-24. 1999.

APPENDIX

Table a1: Users Age

Users Age	Frequency	Percent	Valid Percent	Cumulative Percent
18 a 24	68	24,5	24,5	24,5
25 a 34	123	44,4	44,4	69,0
35 a 44	55	19,9	19,9	88,8
45 a 54	21	7,6	7,6	96,4
55 a 64	10	3,6	3,6	100
Total	277	100	100	

Table a2: Users Education

Users Education	Frequency	Percent	Valid Percent	Cumulative Percent
Basic	2	0,72	0,7	0,7
College	60	21,66	21,7	22,5
Comunity College	26	9,39	9,4	31,9
University	137	49,46	49,6	81,5
Master	48	17,33	17,4	98,9
PhD	3	1,08	1,1	100
Total	276	99,64	100	
System	1	0,36		

Table a3: Users by Working Activity

Users byWorking Activity	Frequency	Percent	Valid Percent	Cumulative Percent
Self Emplyoed	17	6,1	6,2	6,2
Company Employed	187	67,5	67,8	73,9
Unemployed	14	5,1	5,1	81,5
Retired	7	2,5	2,5	76,4
Student	45	16,2	16,3	97,8
Other	6	2,2	2,2	100
Total	276	99,6	100	
System	1	0,4		

Table a4: Time of use of IB

Time of use of IB	Frequency	Percent	Valid Percent	Cumulative Percent
Doesn´t use	40	14,4	14,4	14,4
less then 6 meses	20	7,2	7,2	21,7
From 6 moths - 2 years	49	17,7	17,7	39,4
More then 2 -years	168	60,6	60,6	100
Total	277	100	100	

Table a5: Percentage of operatin using IB

Percentage of operatin using IB	Frequency	Percent	Valid Percent	Cumulative Percent
Less then 15%	69	24,9	24,9	24,9
15% - 29%	23	8,3	8,3	33,2
30% - 44%	23	8,3	8,3	41,5
45% - 55%	18	6,5	6,5	48,0
55% - 69%	23	8,3	8,3	56,3
70% - 94%	58	20,9	20,9	77,3
More then 95%	63	22,7	22,7	100
Total	277	100	100	

Using Learning Analytics to Understand the Design of an Intelligent Language Tutor – Chatbot Lucy

Yi Fei Wang
Department of Curriculum and Pedagogy
University of British Columbia
Vancouver, Canada

Stephen Petrina
Department of Curriculum and Pedagogy
University of British Columbia
Vancouver, Canada

***Abstract*—the goal of this article is to explore how learning analytics can be used to predict and advise the design of an intelligent language tutor, chatbot Lucy. With its focus on using student-produced data to understand the design of Lucy to assist English language learning, this research can be a valuable component for language-learning designers to improve second language acquisition. In this article, we present students' learning journey and data trails, the chatting log architecture and resultantapplications to the design of language learning systems.**

Keywords—learning analytics; intelligent tutor; chatbot; second language acquisition; learning design

I. Introduction

The past decade has witnessed a great deal of interest in technology driven language learning. Various technologies, such as interactive websites, artificial intelligence, synchronous chat, and virtual environments, have been developed in many settings and environments to provide assistance to language learners. Among them, artificial intelligence agents such as the chatbot has tremendous potential but "is least explored in regard to its efficacy in second language learning due to the fact that the technology in this function is still under development and has not been widely applied yet" [1].

Based on the belief that chatbot technology is distinguished from other types of computer applications through simulating an intelligent conversation with human users via auditory or textual methods, language learning can take advantage of chatbots that may offer "intelligent conversational agents with complex, goal-driven behavior" [2].

This article presents how student-produced data can be used to understand the design of an intelligent language tutor, chatbot Lucy, to assist English language learning.

II. BACKGROUND INFORMATION

A. *Communicative Approach to Second Language Acquisition*

In an effort to improve English language learning in British Columbia Canada, there has been a renewed pedagogical emphasis on the communicative approach towards teaching English throughout the province. This communicative approach requires natural communication and meaningful interaction in the target language, in which speakers are concerned not with the form of their utterances but with the messages they are conveying and understanding.

One way to foster positive English learning outcomes is to provide learners comprehensible input in low anxiety situations, containing messages that learners really want to hear[3]. This suggestion of comprehensive input lays a solid foundation for the instructional model that is now commonly known as the communicative approach to language acquisition.

This communicative approach to second language acquisition does not force early production in language learning, but allows learners to produce sentences when they are ready. It recognizes that improvement comes from supplying communicative and comprehensible input instead of forcing and correcting production [3]. Unlike the behaviourist-centered perspective of the 1960s that emphasizes stimulus and responses such as the audio-lingual method, the communicative approach "stresses the importance of authentic and meaningful practice in reality-based simulative environments, with the ultimate goal of communicative competence in mind, rather than knowledge of grammar rules" [4].

B. *Artificial Intelligence: Chatbot Technology*

Chatbots are computer programs that simulate a human conversation using natural language. A wide variety of terms have been used, including chatterbots, virtual assistants, virtual agents, intelligent agents or web-bots. Chatbot architecture integrates a language model and computational algorithms to emulate informal chat communication between a human user and a computer using natural language. Users can chat through text or voice input over a computer screen with chatbot text output or audio/voice output.

Chatbots are developed for a variety of reasons. They can be created for fun such as virtual characters and entertainers, or as part of interactive games such as game player. They can be designed to provide specific information and direct dialogue to specific topics such as website guide, frequently asked questions (FAQ) guide, virtual support agent, virtual sales agent, survey taker, quiz host, learning tutor and chat-room host.

Among hundreds of ways of using a chatbot, its potential role as a language tutor has been widely explored in the computer-assisted language learning (CALL) field. As a language learning tutor/facilitator, a chatbot may re-create the learner-teacher bond through providing learners a character that does not get bored or lose patience.

C. Learning Analytics

Learning analytics has it close ties to the field of business intelligence, web analytics, educational data mining and academic analytics [5]. As an emerging field in the intersection of learning and information technology, learning analytics uses student-produced data and analysis models to discover information and social connections, and to predict and advise on learning [6].

The interpretation of a wide range of data produced by and gathered on behalf of students can not only be used to assess academic progress, predict future performance, and spot potential issues [7], but also can be used to predict and advise the design of innovative learning technologies.

III. REVIEW OF RELATED WORKS

Initially, chatbots were developed for fun. They were designed to use simple keyword matching techniques to find a match of users' input [8]. ELIZA was one of a type of chatbot that could extract keywords from users' input, rephrasing users' statements as questions and post them back to users based on Rogerian analysis, a 1960's innovation in counselling.

After ELIZA, other chatbot systems were developed using different algorithms of pattern matching [9] to simulate fictional or real personalities such as PARRY, which used simple internal affective state – fear, anger and mistrust matching, or MegaHAL that used Markov Model, a more linguistically sophisticated model [10].

The exponential growth in text and natural-language interface research in the late 80s encouraged the creation of many new chatbot architectures such as Jabberwacky and ALICE [11].

Jabberwacky, a chatbot that is operated entirely through user interaction, is designed on the principle that the system learns from all its previous conversations with human users. There are no fixed rules or principles programmed into the system. Jabberwacky stores everything that is said to it and uses contextual pattern matching techniques to select the most appropriate response. Hence, Jabberwacky relies entirely on previous conversations [7].

The widely used ALICE was the winner of the 2000, 2001, and 2004 Loebner competition. Developed by Dr. Richard Wallace using an XML-based language called AIML (Artificial Intelligence Markup Language), ALICE aims to entertain users. ALICE is one of Pandorabots, the largest free open-source chatbot community on the Internet. ALICE-style chatbot stores its knowledge of conversation pattern in AIML files. AIML is a derivative of Extensible Mark-up Language (XML) [9]. AIML consists of data objects called AIML objects, which are made up of units called topics and categories [9]. The topic is an optional top-level element, which contains a name attribute and a set of categories related to that topic [9]. The basic unit of knowledge in AIML is called a category. There are three types of categories, namely, atomic categories, default categories and recursive categories. Each category is a rule for matching an input and converting an input to an output. It consists of a pattern that contains words or sentences provided to chatbot, and a template, which is used in matching to find the most appropriate response to users' input and generating the ALICE chatbot answer [9].

Chatbots used for language education are not new. Fryer and Carpenter [11] presented six potential advantages and applications of Jabberwackychatbots for foreign language learning and teaching. According to Fryer and Carpenter [11], chatbots can help language learners through six ways: (1) students tend to feel more relaxed talking to a computer than to a person; (2) The chatbots are willing to repeat the same material with students endlessly; they do not get bored or lose their patience; (3) many bots provide both text and synthesized speech, allowing students to practice both listening and reading skills; (4) bots are new and interesting to students; (5) students have an opportunity to use a variety of language structures and vocabulary that they ordinarily would not have a chance to use; (6) chatbots could potentially provide quick and effective feedback for students' spelling and grammar.

Jia [12] described the CSIEC system that had advantages over the old ELIZA-like keyword matching mechanism. According to Jia [12], the CSIEC system was developed based on logical reasoning and inference directly through syntactical and semantic analysis of textual knowledge. His paper explored an NLML approach to generate communicative responses. In the paper, Jia presented the CSIEC system architecture and underlying technologies as well as its educational application results. His statistical analysis of the experiment indicated that users preferred the unique chatting function in the CSIEC system, which was lacked in other chatbot systems [12].

Wang [13] reported an ethnographic study that investigated ESL learners' experiences with a commercial chatbot English tutor. Her study identified four conditions for effective chatbot-supported English learning, namely, communicative practice, multimodal interface, emotional design and individualized content. Her findings revealed the promises of chatbot technology in terms of its communicative function for creating an optimal interactive English learning environment.

Lehtinen [14] discussed a research study that used Jabberwacky, God, ALICE and George to learn English. His findings showed an overall positive outlook in interacting with chatbots. His research demonstrated that regardless of the structured or unstructured use, AI chatbots had great potential to be used inside and outside a language classroom as they might allow language learners to practice language and develop confidence in an individualized stress-free manner at their own pace and preference.

Coniam [15] evaluated six chatbots available either online or for purchase – Cybelle, Dave, George, Jenny, Lucy and Ultra Hal Assistant. His evaluation examined chatbots from the perspective of interfaces as a human-looking or sounding partner to chat with, and the usability as pieces of software suitable for ESL learners. Coniam concluded that chatbots had matured considerably since the early days of ELIZA, but they still had a long way to go before they could interact with students in the way that researchers such as Atwell [16] envisaged.

Williams and Compernolle [17] investigated interactions between a chatbot and French learners at various levels of proficiency as well as a native speaker of French. Their study responded to Fryer and Carpenter's [11] six potential advantages and applications of chatbots for foreign language learning and teaching by arguing that the discourse of the particular chatbot represented a less-than-ideal communicative model for learners. Chatbots, as peers/tools for language learners might offer some potential for language learning, but at present, post-interaction tasks based on transcripts appear to hold the most promise for language awareness and development.

IV. MATERIALS AND METHODS

Building on the research literature of using chatbot for language education, this research explored the instructional design process of an intelligent language tutor, chatbot Lucy through critical analysis of student-produced data. In particular, guided by findings of Williams and Compernolle [17], this research responded to and built upon Williams and Compernolle's response to Fryer and Carpenter's six potential advantages and applications of chatbots for foreign language learning and teaching.

A. *Commercial Chatbot Lucy*

The commercial chatbot Lucy is a digital language tutor that can carry on extensive conversations with learners as they speak into their computers through a microphone. Using an advanced speech recognition system, Lucy can give learners feedback on their pronunciation and guide them through useful exercises to improve their pronunciation and accuracy. Lucy's world is where learners meet Lucy. In each world, Lucy offers users over 1000 sentences on a specific subject. Each of Lucy's worlds focuses on a different topic including helping visitors, hotel English, giving directions, English for traveling and restaurant English (Figure 1).

Fig. 1. Lucy's world

Learners need a microphone connected to a laptop. Lucy's learning materials are translated into seven languages including simplified Chinese, traditional Chinese, Japanese, Korean, Vietnamese, Russian, Spanish and Portuguese (Figure 2).

When learners enter into Lucy's world, Lucy greets them and starts the conversation. If learners do not hear or understand what she says the first time, they can click on her to make her repeat.

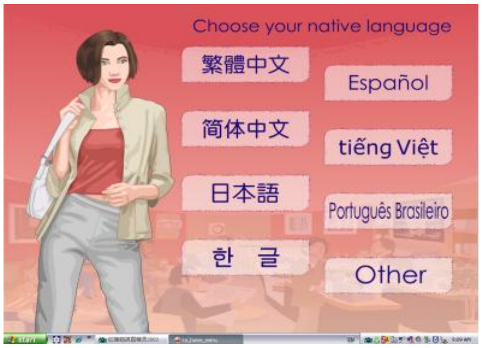


Fig. 2. Lucy's translation interface

If there are some words that learners don't understand, they can just mouse over and Lucy shows the translated languages. If learners want to challenge their listening skills, they can close the translation window so what Lucy says won't appear on the computer screen. Lucy's worlds: Travel English, Helping Visitor, Restaurant English, Hotel English and Small Talk are used in the study (Figure 3).

Fig. 3. Examples of Lucy's world

B. *Intelligent Chatbot Lucy*

Intelligent chatbot Lucy, hosted on Pandorabots website[1], is an online language robot created to help English 101 learners review English grammar and vocabulary learned from Lucy's world. It is an offshoot of "Dr. Wallace's A.L.I.C.E. – March 2002" ALICE artificial intelligence program. Lucy is designed to be more "language tutor" than ALICE. She is trained based on the commercial chatbot Lucy's world (Figure 1). Besides this, a default response category is built into Lucy as an Input Pattern. As well, a recursive category is built in to allow learners to express the same meaning using different sentence structures.

C. *Method – Discourse Analysis*

Language is structured according to different domains of social life [18]. Discourse analysis is the analysis of these patterns [18].

Computer-mediated discourse is "the communication produced when human beings interact with one another by transmitting messages via networked computers" [19]. Computer-mediated discourse uses discourse analysis to address the focus of language and language use in computer networked environments [19].This research focused on the discourse between language learners and chatbot Lucy. The analysis of the conversational patterns saved in Lucy's logs is a key to understanding how language learning happens when

[1] http://www.pandorabots.com/botmaster/en/home

using a chatbot and how it can be better designed based on language learning trials.

Drawing on computer-mediated discourse analysis (CMDA), we examine the conversation logs of these interactions. CMDA in this study aims to understand the learning nature of the online communication between language learners and Lucy. Such an understanding is facilitated by the fact that language learners engage in meaningful learning activities in an online conext in a way that they typically leave a textual trace, making the interactions accessible to scrutiny and reflection and enabling researchers to employ empirical, micro-level methods to shed light on macro-level phenomena [20].By critically analyzing learning dialogues, we identify patterns of learning activities that correspond to meaningful learning and knowledge construction. The approach to analyzing logs of verbal interaction [20], in search of indicators of learning and design clues, allows us to transform student-produced data into a new and coherent depiction of the affordances of chatbot for language education and how we should design chatbot's response and feedback to engage language learners.

V. Results and Discussion

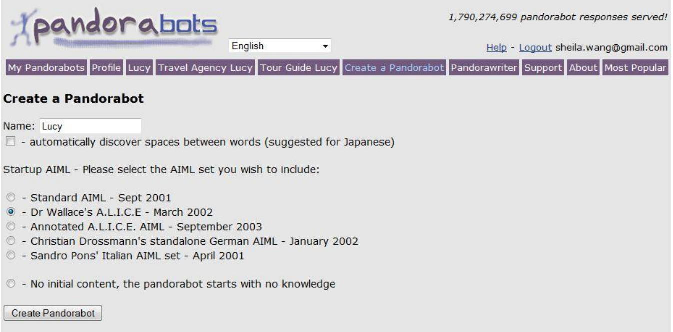
Results of understanding the design of Lucy to assist English language learning are presented in parallel with the discourse analysis of communication logs saved in Lucy.

A. *The Design of Chatbot Lucy*

Intelligent chatbot Lucy is initially designed as an offshoot of "Dr. Wallace's A.L.I.C.E. – March 2002" artificial intelligence program (Figure 4). She is trained to play five characters – travel agency assistant, hotel assistant, tour guide, waitress and call center assistant. Conversations from Lucy's world are converted to AIML using Pandorawriter[2] (Figure 5). AIML files are then uploaded onto Lucy's AIML file logs (Figure 6).

B. *Learning Procedure*

English101 language learners are asked to interact with the commercial chatbot Lucy first. They are required to learn vocabulary, grammar and sentences in Lucy's world. Learners are then asked to communicate with Lucy online with the focus on reviewing vocabulary, grammar and sentences learned from Lucy's world.



Create a Pandorabot

Name: Lucy

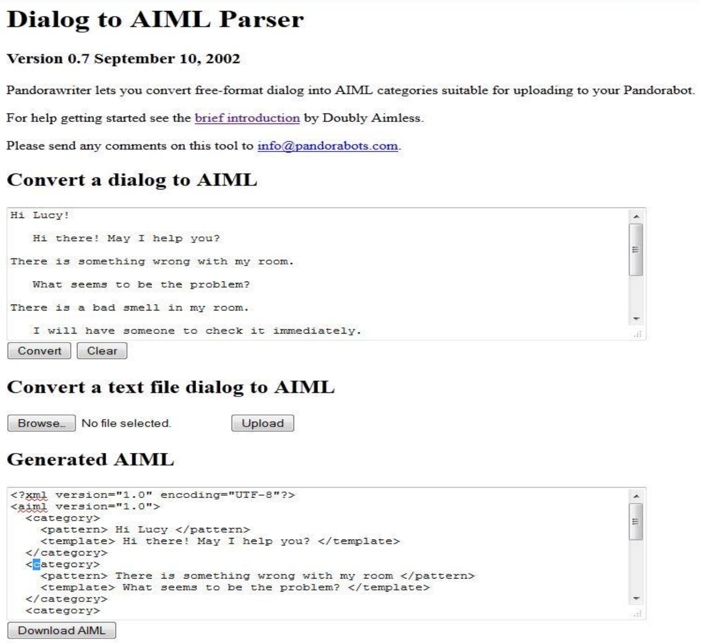
□ - automatically discover spaces between words (suggested for Japanese)

Startup AIML - Please select the AIML set you wish to include:

○ - Standard AIML - Sept 2001
● - Dr Wallace's A.L.I.C.E - March 2002
○ - Annotated A.L.I.C.E. AIML - September 2003
○ - Christian Drossmann's standalone German AIML - January 2002
○ - Sandro Pons' Italian AIML set - April 2001

○ - No initial content, the pandorabot starts with no knowledge

Fig. 4. Lucy's creation interface



Dialog to AIML Parser

Version 0.7 September 10, 2002

Pandorawriter lets you convert free-format dialog into AIML categories suitable for uploading to your Pandorabot.

For help getting started see the brief introduction by Doubly Aimless.

Please send any comments on this tool to info@pandorabots.com.

Convert a dialog to AIML

```
Hi Lucy!

   Hi there! May I help you?

There is something wrong with my room.

   What seems to be the problem?

There is a bad smell in my room.

   I will have someone to check it immediately.
```

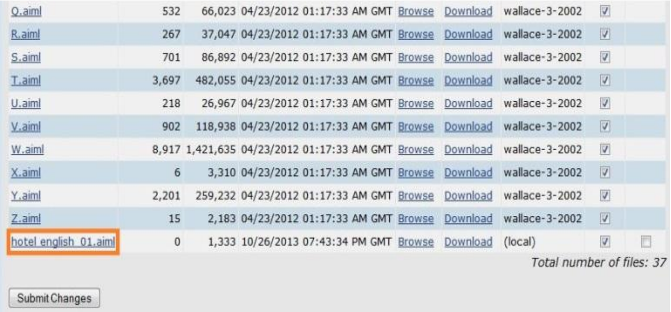
Convert a text file dialog to AIML

No file selected.

Generated AIML

```
<?xml version="1.0" encoding="UTF-8"?>
<aiml version="1.0">
  <category>
    <pattern> Hi Lucy </pattern>
    <template> Hi there! May I help you? </template>
  </category>
  <category>
    <pattern> There is something wrong with my room </pattern>
    <template> What seems to be the problem? </template>
  </category>
  <category>
```

Fig. 5. AIML converter



Q.aiml	532	66,023	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
R.aiml	267	37,047	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
S.aiml	701	86,892	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
T.aiml	3,697	482,055	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
U.aiml	218	26,967	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
V.aiml	902	118,938	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
W.aiml	8,917	1,421,635	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
X.aiml	6	3,310	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
Y.aiml	2,201	259,232	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
Z.aiml	15	2,183	04/23/2012 01:17:33 AM GMT	Browse	Download	wallace-3-2002	☑	
hotel english 01.aiml	0	1,333	10/26/2013 07:43:34 PM GMT	Browse	Download	(local)	☑	□

Total number of files: 37

Fig. 6. Lucy's AIML file logs

C. *Discourse Analysis of Verbal Interaction Logs and Its Application in Training Lucy*

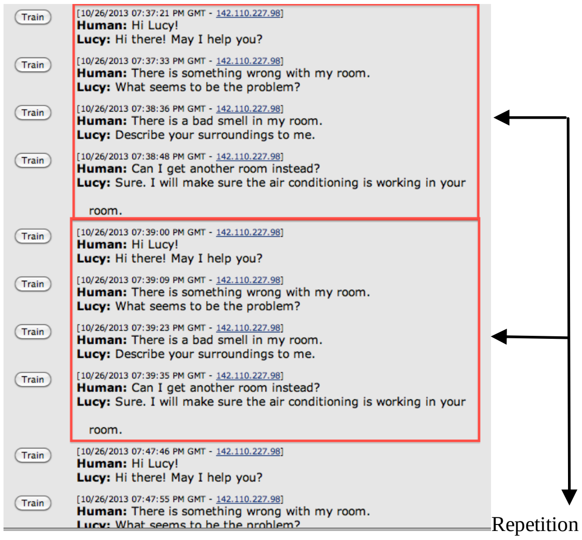
Logs of verbal interaction reflect language learners' learning journey through interacting with intelligent chatbot Lucy. In coding logs of language learners' discourse, we found that Lucy needs to be trained to not only provide language learners with meaningful responses but also with feedback that can target on language learners' common errors.

1) *Intelligent Chatbot Lucy's Ability to Repeat*

One of Lucy's important features in this study is her ability to repeat sentences. English 101 is designed for intermediate level language learners. Logs of verbal interaction show a repetition pattern used by learners. Lucy is willing to repeat the same materials with students endlessly; literally, chatbots do not get bored or lose their patience [11] (Figure 7).

Intermediate or lower level learners benefit from this repetition, which may provide them an opportunity to understand sentence structures thoroughly.

[2] http://www.pandorabots.com/botmaster/en/aiml-converter.html



Train [10/26/2013 07:37:21 PM GMT - 142.110.227.98]
Human: Hi Lucy!
Lucy: Hi there! May I help you?

Train [10/26/2013 07:37:33 PM GMT - 142.110.227.98]
Human: There is something wrong with my room.
Lucy: What seems to be the problem?

Train [10/26/2013 07:38:36 PM GMT - 142.110.227.98]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

Train [10/26/2013 07:38:48 PM GMT - 142.110.227.98]
Human: Can I get another room instead?
Lucy: Sure. I will make sure the air conditioning is working in your room.

Train [10/26/2013 07:39:00 PM GMT - 142.110.227.98]
Human: Hi Lucy!
Lucy: Hi there! May I help you?

Train [10/26/2013 07:39:09 PM GMT - 142.110.227.98]
Human: There is something wrong with my room.
Lucy: What seems to be the problem?

Train [10/26/2013 07:39:23 PM GMT - 142.110.227.98]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

Train [10/26/2013 07:39:35 PM GMT - 142.110.227.98]
Human: Can I get another room instead?
Lucy: Sure. I will make sure the air conditioning is working in your room.

Train [10/26/2013 07:47:46 PM GMT - 142.110.227.98]
Human: Hi Lucy!
Lucy: Hi there! May I help you?

Train [10/26/2013 07:47:55 PM GMT - 142.110.227.98]
Human: There is something wrong with my room.
Lucy: What seems to be the problem?

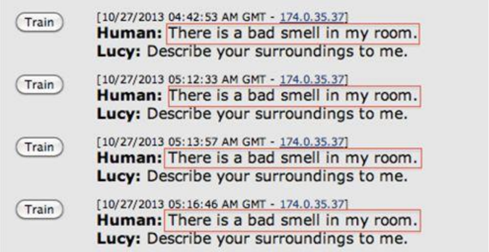
Repetition continues

Fig. 7. Repetition of verbal interaction

2) Chatbot Lucy's Ability to Match

Lucy conducts conversations with learners by matching patterns to find the most appropriate response to input. Hence, learners may get confused by responses that differe from the commercially trained examples. When this happens, some learners retype the same sentence into Lucy but get the same response (Figure 8).

As shown in Figure 8, when a learner encounters responses different from those in Lucy's world, he/she may repeat the same sentence many times. The learner assumes that Lucy should generate the same chat response as the commercial example by replying to the learner: "I will have someone to check it immediately."



Train [10/27/2013 04:42:53 AM GMT - 174.0.35.37]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

Train [10/27/2013 05:12:33 AM GMT - 174.0.35.37]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

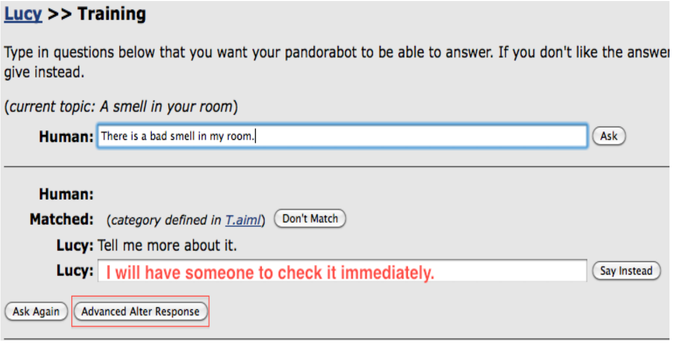
Train [10/27/2013 05:13:57 AM GMT - 174.0.35.37]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

Train [10/27/2013 05:16:46 AM GMT - 174.0.35.37]
Human: There is a bad smell in my room.
Lucy: Describe your surroundings to me.

Fig. 8. Repetition of the same sentence

In order to reduce learners' stress and confusion in communicating with Lucy, we trained Lucy to respond to learners in the same way as the commercial examples in chatbot Lucy.

As shown in Figure 9, the Input words are "There is a bad smell in my room". This Input Pattern should be matched by Lucy's output response. We redefined the output response by typing "I will have someone to check it immediately" into Lucy's training interface.



Lucy >> Training

Type in questions below that you want your pandorabot to be able to answer. If you don't like the answe give instead.

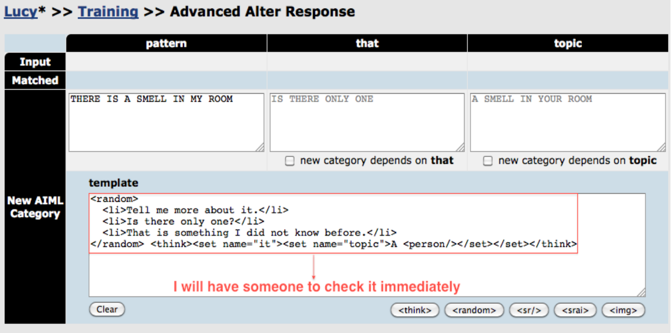
(current topic: A smell in your room)

Human: There is a bad smell in my room.

Human:
Matched: (category defined in T.aiml)
Lucy: Tell me more about it.
Lucy: I will have someone to check it immediately.

Fig. 9. Lucy's training interface

Lucy searches for a path of linked nodes that matches the Input Pattern. We used the Advanced Alter Response as shown in Figure 10 to add a new template to the AIML category. We changed the labeled template into "I will have someone to check it immediately" and saved our change. When we return to the training interface, we click on the Ask Again button to cycle through the complete set of responses.



Lucy* >> Training >> Advanced Alter Response

	pattern	that	topic
Input			
Matched			
New AIML Category	THERE IS A SMELL IN MY ROOM	IS THERE ONLY ONE □ new category depends on **that**	A SMELL IN YOUR ROOM □ new category depends on **topic**

template

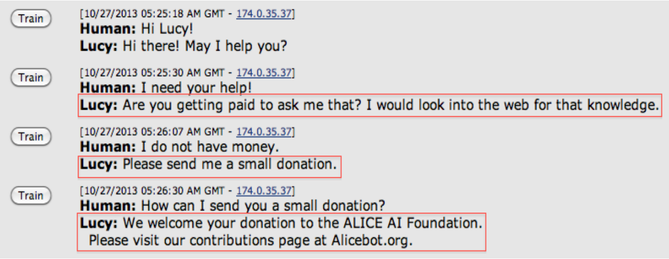
```
<random>
  <li>Tell me more about it.</li>
  <li>Is there only one?</li>
  <li>That is something I did not know before.</li>
</random> <think><set name="it"><set name="topic">A <person/></set></set></think>
```

I will have someone to check it immediately

Fig. 10. Lucy's Advanced Alter Response

The potential variation such as the above example in this study is immense. Like Williams and Compernolle [17], we also discovered that "the lexicon is determined by the amount of time spent by the botmaster entering data and the level of sophistication of the software"

Figure 11 shows that when a learner does not get a response from Lucy in the way that he/she expects, he/she may stop the interaction.



Train [10/27/2013 05:25:18 AM GMT - 174.0.35.37]
Human: Hi Lucy!
Lucy: Hi there! May I help you?

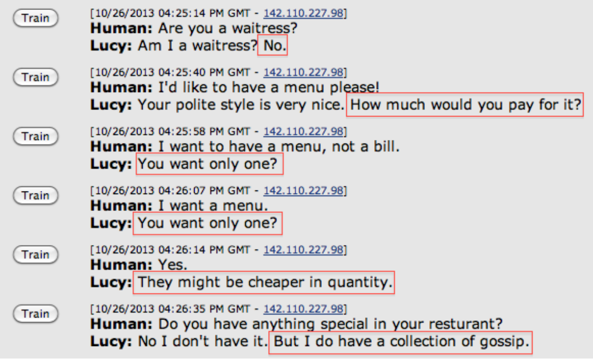
Train [10/27/2013 05:25:30 AM GMT - 174.0.35.37]
Human: I need your help!
Lucy: Are you getting paid to ask me that? I would look into the web for that knowledge.

Train [10/27/2013 05:26:07 AM GMT - 174.0.35.37]
Human: I do not have money.
Lucy: Please send me a small donation.

Train [10/27/2013 05:26:30 AM GMT - 174.0.35.37]
Human: How can I send you a small donation?
Lucy: We welcome your donation to the ALICE AI Foundation. Please visit our contributions page at Alicebot.org.

Fig. 11. Lucy asking for donation

When the learner says "I need your help", Lucy presents him/her a response regarding a donation to the ALICE AI Foundation. The learner continues his/her response to Lucy by saying – "I do not have money". This conversation begins with a topic off the learning track; hence, the learner stops the conversation with Lucy.



[10/26/2013 04:25:14 PM GMT - 142.110.227.98]
Human: Are you a waitress?
Lucy: Am I a waitress? No.

[10/26/2013 04:25:40 PM GMT - 142.110.227.98]
Human: I'd like to have a menu please!
Lucy: Your polite style is very nice. How much would you pay for it?

[10/26/2013 04:25:58 PM GMT - 142.110.227.98]
Human: I want to have a menu, not a bill.
Lucy: You want only one?

[10/26/2013 04:26:07 PM GMT - 142.110.227.98]
Human: I want a menu.
Lucy: You want only one?

[10/26/2013 04:26:14 PM GMT - 142.110.227.98]
Human: Yes.
Lucy: They might be cheaper in quantity.

[10/26/2013 04:26:35 PM GMT - 142.110.227.98]
Human: Do you have anything special in your resturant?
Lucy: No I don't have it. But I do have a collection of gossip.

Fig. 12. Restaurant English

Figure 12 is another example where Lucy responds to a learner through a random matching. The learner aims to practice Restaurant English in this conversation. He/she asks Lucy - "Are you a waitress?" which is different from the way that Lucy is trained. So Lucy questions the learner by replying "Am I a waitress? No".

In spite of being refused, the learner now starts to use the sentence example from Lucy's world – "I'd like to have a menu please".

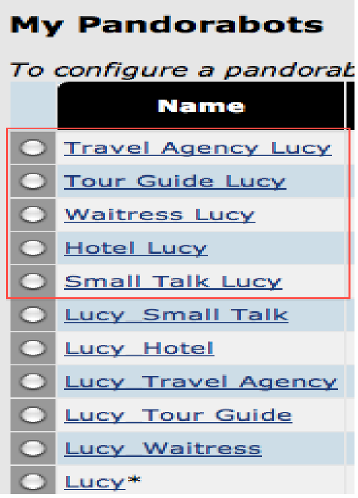
Lucy does not respond to the learner with something meaningful. Lucy's response – "How much would you pay for it? – leads to an incoherent response to the learner's illocutionary act.

The learner continues his/her turn by saying – "I want to have a menu". In response to Lucy's question – "You want only one?", the learner repeats –" I want a menu." Lucy asks again by saying - "You want only one?"

The learner seems tired of this conversation. So he/she says "Yes". Again, Lucy does not reply to the learner with a meaningful response.

The learner decides to have another try by starting over the conversation using exactly the same sentence from Lucy's world with the hope of continuing the Restaurant English conversation. Unfortunately, Lucy fails to respond to the learner. As a result, the learner stops the conversation.

Lucy's random match to the Input Pattern is problematic. Although Lucy has five characters built into the system, the Default Responses in the Knowledge Web randomly select something meaningful as the chatbot's response to learners. In order to avoid this problem, we redesigned Lucy to be Small Talk Lucy, Hotel Lucy, Waitress Lucy, Tour Guide Lucy and Travel Agency Lucy as shown in Figure 12.



My Pandorabots

To configure a pandorab

Name

○ Travel Agency Lucy
○ Tour Guide Lucy
○ Waitress Lucy
○ Hotel Lucy
○ Small Talk Lucy
○ Lucy_Small Talk
○ Lucy_Hotel
○ Lucy_Travel Agency
○ Lucy_Tour Guide
○ Lucy_Waitress
○ Lucy*

Fig. 13. Lucy on Pandorabots website

The redesigned Lucy aims to help language learners review exactly the same sentence structures learned from Lucy's world. Besides this, we open up some options for our learners to learn more expressions with the similar meaning of those sentences in Lucy's world.

There is no initial content built into Small Talk Lucy, Hotel Lucy, Waitress Lucy, Tour Guide Lucy and Travel Agency Lucy. We converted learning content from Lucy's world into AIML and uploaded into each corresponding Lucy. The example below is the AIML file generated using Pandorawriter[3]. This example only uses atomic categories, which only contain a Pattern and Template and do not have wildcard symbols, _ or *.

```
<?xml version="1.0" encoding="UTF-8"?>
<aiml version="1.0">
<category>
<pattern> Hi Lucy </pattern>
<template> Hi there! May I help you? </template>
</category>
<category>
<pattern> There is something wrong with my room </pattern>
<template> What seems to be the problem? </template>
</category>
<category>
<pattern> There is a bad smell in my room </pattern>
<template> I will have someone to check it immediately. </template>
</category>
<category>
<pattern> Can I get another room instead </pattern>
<template>
        Sure. I will make sure the air conditioning is working in your
room.
</template>
</category>
</aiml>
```

We also designed default categories to allow Lucy to respond to learners if the Input Pattern is not found in the Knowledge Web. We used Lucy's training interface and Advanced Alter Response to add some randomly possible meaningful responses. Besides default categories, we designed

[3] http://www.pandorabots.com/botmaster/en/aiml-converter.html

recursive categories, which may allow learners to experience some different ways to express the same meaning.

The five modules of Lucy's logs show that learners at lower levels of language proficiency benefit from the interaction with Lucy. Learners in this study seem better suited to communicate with Lucy due to the fact that learning outcomes in this study do not require learners to use a variety of language structures and instead require them to practice and review exactly the same sentence structures and grammar as what they learn from examples in Lucy's world.

3) Chatbot Lucy's Ability to Provide Feedback

Another important feature that we designed in Lucy is her ability to provide spelling and grammar correction feedback. Continuous feedback is difficult to be mimicked, much less produced in a random fashion. The main difficulty for a chatbot to check spelling and grammar is that an optional list of candidate words cannot be built in the system. Hence, we used logs to identify learners' spelling/grammar errors and entered data into Lucy. This means the more learners use Lucy and the more spelling/grammar feedback data entered into Lucy, the more robust Lucy becomes. For example:

```
    <category>
    <pattern>I WOULD LIKE TO HAVE A SMKING ROOM</pattern>
    <template>Do you mean smoking? <think>
    <set name="it">
    <set name="want"><set name="topic">to have <person/></set></set>
    </set>
    </think></template>
</category>
```

D. Discussion

The design of Lucy aims to help learners review and practice exactly the same sentence structures learned from Lucy's world. In response to Fryer and Carpenter [11] and Williams and Compernolle [17], we find that chatbot can be designed to repeat the same material with learners, endlessly. We believe that this is one of the affordances that the chatbot may provide to intermediate levels or lower levels of language learners. Lucy has a speech recognition system installed, which aims to help learners practice both listening and reading skills. Lucy does not help learners review spellings and sentence structures.

In our case, we redesigned Lucy to provide learners an opportunity to use a variety of language structures and vocabularies that they ordinarily would not have a chance to use. Learners do not have opportunities for generating their own output due to the fixed sentence structures designed in Lucy's world. Lucy's world doesn't provide affordances for learners to negotiate Lucy's expressions. We opened up some opportunities for learners to try limited language structures with the hope of engaging them in active conversation. But Lucy is very limited in providing learners an opportunity to use a variety of language structures and vocabulary because of the amount of time it requires to enter data into the system and the level of sophistication of the software.

Furthermore, "feedback is a classical concept in learning, whose importance is acknowledged across different learning theories" [19]. Data analyzed in our study shows that chatbotscan provide effective feedback for learners' spelling and grammar, but it depends on extensive entry of error data into the system.

VI. Conclusion

Discourse analysis of learning trails plays an important role in designing interactive intelligent language tutor systems for language learners at intermediate or lower levels.

We found great advantages in chatbot technologies in that they offer language learners realistic opportunities for individual tutoring. Language learners can tailor a chatbot for their own pace of learning: They can enter an answer to each question, repeat a sentence without pressure, or skip sentences that do not make sense to them or are difficult to identify.

The potential use of chatbots can simulate human-like communication. Language learning implies corresponding cultural learning. Understanding culture is a key to understanding the language use in contexts. Chatbot Lucy does not contain this feature. Language learners who eventually can communicate with native speakers require cultural knowledge of the target language.

Another issue that we experienced in this study is continuous feedback. Continuous feedback requires very fast interpretation of learners' input on the fly. We found that chatbot technology has a limitation in how to quantify and model continuous feedback and handle the fast integration and interpretation.

By applying learner analytics for understanding the design of the intelligent chatbot Lucy, this study generates important findings for scrutinizing student-produced data and learning trials for the design of learning technologies. This study opens up possibilities for connecting and analyzing students' data trials. Approaches developed in this study can be useful in studying an instructional innovation through the lens of text-based messages [18]. Insights gained from this study can also inspire additional learning technology research.

Acknowledgment

We thank Drs. Ling Shi and Franc Feng for expertise, support and incisive comments. We also thank colleagues in the How We Learn Lab for help and support with data collection.

References

[1] Y. Zhao and C. Lai, Technology and second language learning: Promises and problems, in L. Parker (Ed.) Technology-mediated Learning Environments for Young English Learners, London, UK: Lawrence Erlbaum Associates, 2008, pp.167-205.

[2] A. Kerly, P. Hall and S. Bull, "Bring chatbots into education: Towards natural language negotiation of open learner models," Knowledge-based Systems, vol. 20, pp.177-185, 2006.

[3] S. Krashen, Second Language Acquisition and Second Language Learning, Park, CA: Sage, 1981.

[4] S. Prizeman, "Towards technology best practices: The future of second language instruction in British Columbia public schools," unpublished.

[5] P. Prinsloo, S. Slade and F. Galpin, "Learning analytics: Challenges, paradoxes and opportunities for mega open distance learnng institutions," Proceedings of the 1st International Conference on Learning Analytics and Knowledge. Banff, AB, Canada, 2012.

[6] G. Siemens, "What is learning analytics," unpublished.

[7] L. Johnson, R. Smith, H. Willis, A. Levine and K. Haywood, The 2011 Horizon Report. Austin, TX: The New Media Consortium, pp.28, 2011.

[8] J. Weizenbaum, "ELIZA: A computer program for the study of natural language communication between man and machine," Communication of the ACM, vol. 9 issue 1, pp.35-36, 1966

[9] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?" Band vol. 22 issue 1, pp.29-49, 2007.

[10] T. Hutchens and M. Alder, "Introducing MegaHAL," Unpublished.

[11] L. Fryer and R. Carpenter, "Emerging technologies: Bots as language learning tools," Language Learning & Technology, vol. 10 issue 3, pp.8-14, 2006.

[12] J. Jia, "CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning," Knowledge-Based Systems vol.22, pp.249-255, 2009.

[13] Y. F. Wang, "Designing chatbot technology for oral English practices: An ethnographic research into ESL learners' experiences," unpublished.

[14] B. Lehtinen, "Accessing language using online chatbots," unpublished.

[15] D. Coniam, "An evaluation of chatbots as software aids to learning English as a Second Language," The EUROCALL Review, vol. 13, 2008.

[16] E. Atwell, "The language machine: The impact of speech and language technologies on English language teaching," British Council, 1999.

[17] L. Williams and R. A. Compernolle, "The chatbot as a peer/tool for learners of French," in L. Lomicka and G. Lord (Ed.) The Next Generation: Social Networking and Online Collaboration in Foreign Language Learning, CALICO, 2009.

[18] M. Jorgensen and L. Phillips, Discourse Analysis, London: SAGE Publications, 2002.

[19] S. C. Herring, "Commuter-mediated discourse," in D. Tannen, D. Schiffrin and H. Hamilton (Ed.) Handbook of Discourse Analysis, Oxford: Blackwell, 2004, pp.612-634.

[20] S. C. Herring, "Computer-mediated discourse analysis: An approach to researching online behavior," in S. Barab, R. Kling and J. H. Gray (Ed.) Designing for Virtual Communities in the Service of Learning, New York: Cambridge University Press, 2004, pp.338-376.

[21] Y. Zhao and C. Lai, Technology and second language learning: Promises and problems. In L. L. Parker (Ed.) Technology-mediated Learning Environments for Young English Learners, London, UK: Lawrence Erlbaum Associates, 2008, pp.167-205.

Web Resources Annotation for the Web of Learning

Jawad Berri
Information Systems Department
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Semantic annotation of web resources is an essential ingredient to leverage the web of information to the semantic web where resources are easily shared and reused. In the education field, reusing hypermedia web resources can support to a great deal the design of modern instructional environments and the development of interactive and non-linear material for learning .Sharing and reusing these resources by different web applications and services presupposes that they are visible for retrieval through a semantic description of their content, function and relations with other resources. This paper presents the annotation and discovery of web resources to create learning objects that constitute the building blocks of learning sessions which are delivered to users in the Web of Learning. Semantic annotation is done by the contextual exploration method which analyzes web resources' text descriptions and metadata in order to annotate automatically resources. We present the system architecture and a case study that illustrates the proposed approach.**

Keywords—Web resources; semantic annotation; web of learning; contextual exploration

I. Introduction

Data has proliferated on the web during the last decade resulting in a huge amount of web resources. Web 2.0 technologies eased this information rise by providing tools for collaboration and sharing. The advances in mobile technologies has also facilitated for users to produce and upload with few clicks web resources that are conveniently shared on the web. Sharing and reusing these resources by different web applications and services presupposes that they are described semantically.

This task is a necessity to mutate the existing web of information to the semantic web (or Web 3.0) [1, 2]. Semantic description allows a web resource to be searched and retrieved in accordance with its substance, the function it achieves and its relationships with other resources. It will be visible through its semantic description and not simply keywords, which makes it easy for semantic web search engines to discover it. In the education field, reusing hypermedia web resources can support to a great deal the design of modern instructional environments that exploit available information and ubiquity of technologies. Hypermedia web resources such as video and audio files, images, wikis, presentations, web documents and othersare particularly interesting as they allow the development of interactive and non-linear material for learning. Reusing hypermedia resources for learning will also relief educators from the burden of systematically authoring learning material which is a major bottleneck in the design of instructional environments.

A. *Intuitive Instructional Environments*

The development of new instructional environments is a necessity as learners in the age of technology are exposed all the day to different kinds of sophisticated devices and very rich information content. They interact with their devices with ease and have intimate relations with them. In this hi-tech environment characterized by rich content hypermedia information, learners are expecting to be exposed to familiar environments when they seek information, communicate, play games and learn. Although lots of efforts have been deployed to develop learning environments in the field of education by using Web 2.0 technologies, there is still work to be done to create learning spaces where learning becomes intuitive and more adapted to the real needs of learners. In fact, information should be disseminated in such a way that users looking for needed information become learners as they will be able to deepen and diversify their knowledge through durable learning. Making use of available hypermedia web resources can contribute greatly towards the creation of intuitive learning spaces as they promote interaction and allow fluent navigation over the learning environment.

Web 2.0 technologies and hypermedia information available on the web can contribute greatly in the education field. Web resources can be reused and aggregated with other material to fit education purposes. Although the majority of web resources available are not meant to be used for education, consulting and viewing these resources by users is a form of learning as knowledge is acquired and used in their lives. For example a video which shows how wild life animals hunt a pray can be used into a biology class to illustrate the food chain principles. Users accessing the web seeking information is a form of ad hoc learning requiring the user to spend time looking for adequate resources that allow him/her to build relevant and enough knowledge about a topic of interest.

B. *The Web of Learning*

The *Web of Learning* is a learning ecosystem build on the top of the existing web (the web of information) which makes use of existing web resources and organizes them to fit education purposes [3]. It aims to reorganize web information in order to provide users with learning spaces that are generated from information and web resources. The Web of Learning is characterized by a set of features: i) the use of hypermedia resources available on the web that match the learner's needs; ii) the integration of different forms of hypermedia information to involve diverse cognitive human activities in learning; iii) it allows learners to construct their personal learning pathways among the proposed learning structure to promote active and adaptive learning; and iv) it manages learning through sessions allowing learners to learn according to their pace and constraints and can resume their learning at any time. Web of Learning promotes just-in-time-learning which means that the generation of learning material is done when needed. This paradigm supports education on demand and allows users to control the pace and course of learning. In this paper we address mainly the annotation of hypermedia web resources as a requirement to reuse them in the web of learning. We present also the semantic annotation architecture which creates learning spaces on-demand. The architecture relies on the annotation and discovery of web resources through their semantic description to generate learning objects that constitute the building blocks of learning sessions which are delivered to users in the Web of Learning. Semantic annotation is done by the contextual exploration: a linguistic method which analyzes web resources' text descriptions and metadata in order to annotate them automatically.

This paper is organized as follows: next section presents similar approaches that have been proposed in semantic annotation. Section 3 explains how automatic semantic annotation of resources can serve the education field. Section 4 presents the contextual exploration method as a computational system that can carry out resource annotation automatically. Section 5 exposes the system architecture and the following section illustrates the approach through a case study. In the last section we conclude this work and propose research work to be undertaken in the near future.

II. Related Work

Research about Semantic Annotation is abundant specifically with Web 2.0 wave which have favored social interactions between web users. Social tagging or social annotation is the particular activity by which a web user associates a semantic tag (text string) to a web resource (video, image, web page, web application, web service, etc.). Most of the Web tools and systems offering this facility do no restrict the tag selection and give the user the freedom to associate any tag to any resource. Some others use a controlled vocabulary or a thesaurus from which the user chooses the tag. Semantic annotation is very practical for organizing web resources as it improves the search of resources on the web and provides useful recommendations based on the resource content and correlations between resources, users, interests etc. sharing similar tags. In [4] Andrews et al. present a semantic annotation classification of tools and models based on three criteria: i) the structural complexity which indicates the amount of information associated with the annotation. Four types of annotations are proposed based on their complexity: *tags* describing a particular resource property such as the name of a place in a picture ,*attributes* which define a property of the annotated resource such as location or starting date; *relations* which relate a resource with another one; and *ontologies* which allows to describe the resource with respect to an ontology whereby relations, properties, and restrictions will hold among resources.; ii) the vocabulary type which describes the vocabulary used for annotation: free-form natural language text, controlled vocabulary or ontology; and iii) the user collaboration (single-user or community models) which denotes the way users contribute to create different types of annotations. Zubiaga et al. proposed in [5] an approach to produce an automated classification of resources based on existing social tag sets from the social tagging systems: Delicious[1], Library Thing[2] and Good Reads[3]. The approach uses the Support Vector Machines classification algorithm to derive two main results: the way users tag, and the way tags are used by the social tagging system to automatically classify resources. The success of the approach is mainly dependent on factors related to the tagging system itself such as whether the tags are suggested by the system or freely added by the user. Yu and al. in [6] present an approach to annotate educational video resources for distance learning. The approach uses the Linked Data [7] technology and ontologies to annotate videos. The authors developed a tool (Sugar-Tube) that searches resources based on the annotations associated to videos. Moreover, the tool allows to link videos with other educational resources from the Linked Open Data cloud and the web. Another work proposed by Lau and Lee [8] considers annotating educational resources using social tags. The authors present a system that uses folksonomy tags to filter, rank and recommend learning resources which are annotated to fit with the users' needs on a social learning environment. Smine et al. [9] propose a semantic annotation approach based on the contextual exploration method to annotate learning objects. The tags are then used to create a semantic inverted index in order to retrieve learning objects.

Much research has been dedicated to web resource annotation for reuse and sharing. In this paper we are interested to annotate web resources for use in a learning environment. Our approach considers the web as a learning space where resources can be annotated to fit with specific learning scenarios for users in context.

III. Semantic Annotation

Semantic Annotation is the process that associates attributes, comments, descriptions or any other metadata to a resource. This task represents one of the objectives stated by the semantic web initiative of having data on the web defined and linked in such a way that it can be used by machines for automation, integration and reuse across various applications [10]. Lots of works [11] have been done in this direction leaded by the W3C initiative[4] resulting in the definition of web standards [12] for semantic annotation(RDF, FOAF, etc.) and tools (Semantic Media Wiki [13], Annotea [14], KIM [15]) which allow semantic annotation of resources. These efforts have settled a web environment ready for dealing with semantic annotation of resources where users can post their authored resources along with semantic descriptions for public use; however, less focus has been dedicated to develop computational models, architectures and tools that can annotate automatically web resources.

Automatic annotation of web resources is an important web ingredient to leverage existing information and to foster the web mutation. It is essential for resource discovery and reuse. Resources will be produced and then described automatically

[1] http://delicious.com
[2] http://www.librarything.com
[3] http://www.goodreads.com
[4] http://www.w3.org/standards/

by semantic annotation tools in order to be reused in many fields and by different web services and applications. The development of semantic annotation tools that can annotate automatically web resources has many advantages: i) it decouples resources authoring from their semantic description and use. Resources can be produced for a specific purpose and can be reused by applications and services for other purposes if their semantic description fits with the application needs and context; ii) it facilitates resources reuse and sharing as resources will be visible through their semantic description which makes it easy for semantic web search engines to discover them; and iii) it will relief authors from the burden of describing all the features of their resources in detail. In this work we present a semantic annotation system that is able to annotate resources automatically from their text descriptions.

C. Web Resources

The concept of *Web Resource* is fundamental in the web architecture. It is the primitive element that constitutes the web and that can be addressable. A web resource is identified by its Uniform Resource Locator (URL) which has been used originally to address documents and files on the web. The concept has evolved with the diversity of schemes for web resources to encompass any “entity” or “thing” that can be identified in a networked information system. Therefore, the identification of web resources has been extended to Uniform Resource Identifier (URI) specification which provides a simple and extensible means for identifying a resource [16]. Web resources can be described semantically using the Resource Description Framework (RDF)[5] language. RDF is based on XML which facilitates processing, exchange and reuse of web resources and their associated descriptions.

D. Web Resources for Learning

Using web resources in learning is an interesting field to investigate. Modern instructional environments such as those developed in E-Learning or M-Learning are using Hypermedia Web Resources (HWR) embedding multimedia and hypertext medium of information to create interactive non-linear material for learning. The availability of multimedia development software (such as animation tools, presentation tools, web authoring tools, and others) has facilitated the production of HWR such as videos, audios, images, presentations, web documents, maps, news, emails, web services, mobile applications, wikis, blogs, podcasts, etc. HWR are very suitable for learning as they provide more control and adaptation on the learning flow. They also promote interaction and collaborative learning as they ease navigation over the instructional environment. In this research we focus on the description of HWR for resources annotation and discovery to create integrated learning objects. These learning objects are the building blocks of learning sessions that are delivered to users in the Web of Learning.

E. Web Resources Annotation

Automatic annotation of HWR is not a simple task considering the masses of resources available on the web and the difficulty of apprehending these resources for semantic analysis. This task is challenging the research community requiring efforts from many disciplines to describe semantically web information. On the technical side, the network infrastructure needs an extension in terms of storage and access to be able to store semantic descriptions and retrieve them with no delay. The design and implementation of effective semantic annotation systems is fashioned by the resource content to analyze for annotation. The digital content of HWR is naturally the most appropriate data to analyze as it contains the essence of the resource. For instance, the analysis of video files involves video content analysis and an interpretation of the scene data [17]. It requires the extraction of contours and features, contrasting colors between regions and comparing frames in order to identify objects, individuals and motion [18]. This field of research is focused on analyzing the physical data of the resource which are the image pixels represented as bits. Tremendous research efforts have been deployed in this area resulting in the development of sophisticated algorithms and software tools that are able to detect for instance human intruders in video surveillance systems, recognize car plate numbers in traffic monitoring systems, kinetic based video games, etc. Although these systems and tools can recognize objects and their motion with precision they cannot infer basic semantic features from a scene in real life such as names of individuals, their roles, the relation between objects, etc. [19]. These semantic features are essential ingredients for semantic annotation. Relying only on the resource’s digital content analysis does not help much in semantic annotation.

Web resources include generally other data added by the resource creator or by users who viewed the resource. This information can be very rich including the name of the resource, a text description, hyperlinks to other related resources and tags associated to the resource. Also when the resource triggers a specific interest, users add their own comments that can be very interesting to analyze for semantic annotation. In addition, social interaction generates metadata that is added to describe the relevance of the resource such as the number of views, the percentage of users who liked the resource, and personal tags that might be added by users. In this work we focus on analyzing the text and metadata related to the resource for semantic annotation. Analyzing text data requires linguistic analysis tools which need to be robust and flexible in order to deal with unrestricted text on the web. Moreover, dealing with resource metadata necessitates decision models to sort resources based on relevance. In the next section we present a linguistic method – the Contextual Exploration Method, that is able to annotate automatically resources based on the text associated with resources and we present a system that has been developed to annotate videos on the web.

IV. Contextual Exploration

The Contextual Exploration Method (CEM) can significantly contribute to web resource annotation. CEM is a computational linguistics method that is suitable to analyze unrestricted text. CEM is the result of many years of research which has led to the development of a framework which has been applied to solve many problems related to language processing [20, 21, 22]. It is a decision-based method that involves grammatical and lexical knowledge regarding a decision making task when solving a linguistic problem. The

[5] http://www.w3.org/TR/rdf-schema/

method simulates the behavior of a human who is reading a text to analyze it in view of taking a decision. CEM scans the linguistic context looking for linguistic markers that can trigger decisions. Markers can be any word occurrence, morpheme, lexeme, or lexical unit. Once a marker is found, then the context is further analyzed to find linguistic contextual clues surrounding the marker to support taking an unambiguous decision. Linguistic markers and clues are organized into a database and are used by decision rules which annotate text passages.CEM is a flexible and robust method that can deal with web resources repositories which include resources that are freely described by users. Unlike classical natural language processing architectures, CEM does not rely on rigorous parsing and language dictionaries; instead, it uses the linguistic expertise related to the problem at hand and is able to take decisions when annotating texts by mean of heuristics and strategies. CEM can deal with the inherent variations of texts associated with web resources. These texts are written by users who use an open language that is influenced by user's social cultural factors such as the community to which the user belongs, the age, the region and specific customs. For instance, Arabic speaking mobile users refer to the BlackBerry Smartphone by its initials (البي بي*The BB*). This particular naming is used in general by young Arabic speakers. Arabic speaking fans of the Barcelona Football Club refer to their club using many community specific words such as (البارشا، بارسا، البرشا*Albaarshaa, Baarsaa, Albarshaa*) [23]. Social factors play a major role in the expression of the resource description and any social interaction related to the resource; they must be taken into account to process data associated to web resources for semantic annotation.

The Contextual Exploration Method involves a set of resources and development tools to develop contextual exploration modules and knowledge components to annotate web resources. Figure 1shows the different components involved in developing a Contextual Exploration Module for semantic annotation.

F. *Resources and Development Tools*

The Linguistic Expertise is represented in the knowledge base; it includes all the linguistic markers and clues which are first detected in the text to be analyzed. Linguistic markers and clues are organized into equivalence lists that are invoked by decision rules. Decision rules check the presence of a specific marker and clues inside a particular context consisting of a text passage and accordingly assign a tag. Decision rules are hierarchically organized in the knowledge base in order to solve inherent languages' ambiguities that are due to polysemy or equivocal contexts.

Linguistic tools are all the tools that are essential for preparing texts to be analyzed. This component includes preprocessing, tokenizing, stemming and morphological tagging tools to handle texts and to prepare them for annotation by the contextual exploration module. Preprocessing the text is a necessary initial step which objective is to remove noise and filter out data that is not used in the linguistic analysis such as special characters, Xml tags, encodings, etc. Tokenizing splits the text into tokens and associates to each one a set of data such as its offset position in the text, its sentence, and other data related to the text physical structure. Stemming allows to reduce words into a canonical form so that to avoid the differences due to word affixes. Stemming is a light process that does not require a language dictionary as words are simply chopped resulting in word stems. This rough process is suitable sometimes when there is no need for words' part of speech and morphological categories. In case the latter information is needed, then contextual exploration uses Morphological Tagging tools to associate morphological information to the text tokens.

The Corpus is the linguistic repository that helps acquiring and validating the linguistic expertise; it is mainly used in the knowledge acquisition phase to gather linguistic markers and clues and to identify the decision rules.

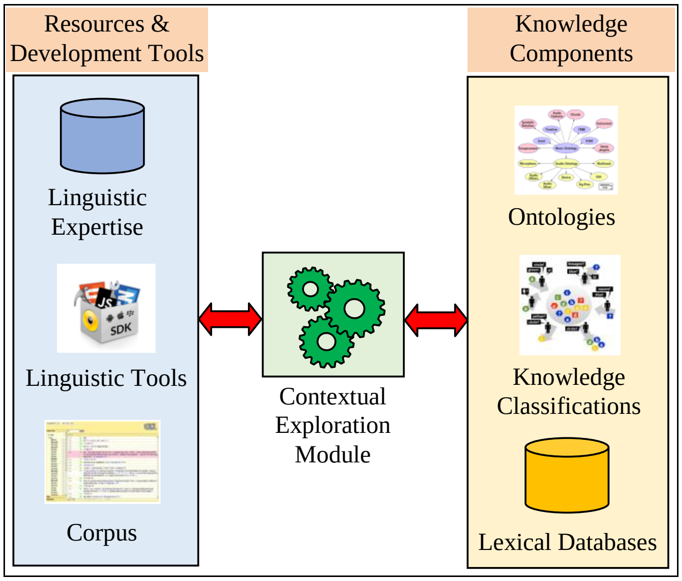


Fig. 1. Contextual Exploration for Semantic Annotation.

G. *Knowledge Components*

CEM requires a specification of knowledge related to semantic annotation. This is done through knowledge components namely: ontologies, knowledge classifications and lexical databases. Ontologies are formal descriptions of concepts, their properties, types and relationships holding between them in a domain of knowledge. When semantic annotation is about knowledge domains for which an ontology is available, CEM uses this ontology in order to annotate texts. For instance in the linguistic field, CEM used the tense and aspect ontology as defined by [20]. When an ontology is not available, which is the case when we deal with open web repositories and unrestricted text, CEM uses available knowledge classifications that are set by domain experts such as taxonomies, thesauri, controlled vocabularies, or folksonomies which generally emerge from web social interaction. Moreover, Lexical Databases are useful to enrich existing knowledge classifications. Word Net [24] is an example of a lexical database that includes many semantic relationships such as synonymy, hyponymy, hypernymy, and antonymy to extend existing knowledge classifications and to relate existing concepts.

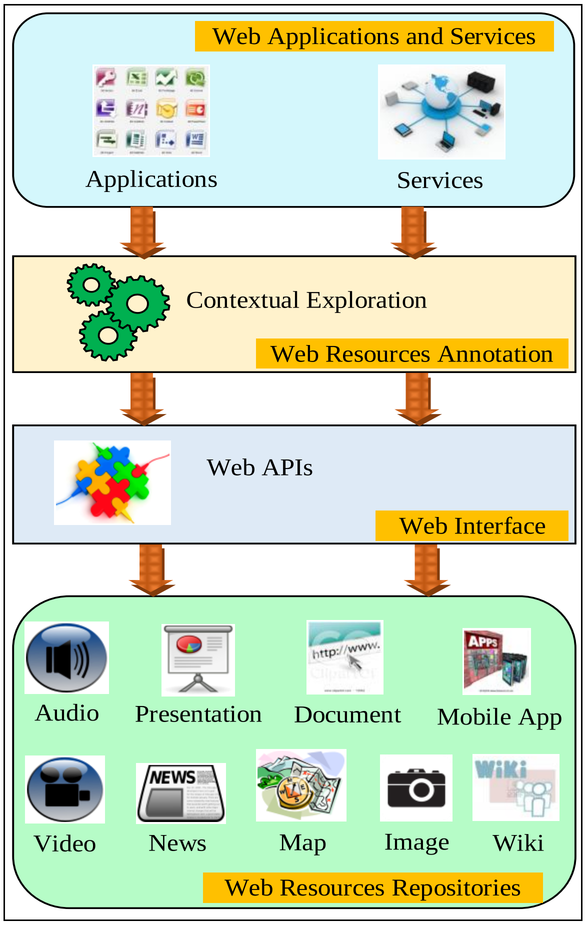


Fig. 2. System architecture

V. System Architecture

Figure 2 presents the system architecture for web resources annotation. The architecture includes four components namely: Web Applications and Services, Web Resources Annotation, Web Interface and Web Resources Repositories. Web resources annotation is a process that is triggered by a request from a web application or a web service which seek to use web resources for a specific purpose. The request is analyzed by the Web Resources Annotation component which activates a contextual exploration module specialized in annotating web resources as requested by the web application or service. CEMs include linguistic expertise for annotating specific semantic categories. The activation of the right CEM is done through a fine analysis of the request [23] which detects what type of annotation is required and the objective of the query and then selects the suitable CEM which analyzes the resource's data and annotates the resource. Accessing web resources requires specific Application Program Interfaces (APIs) to search and get relevant resource's data for annotation. Most of the popular web hypermedia repositories offer convenient web APIs publicly available to search and retrieve hypermedia resources. The Web Resources Repositories Component represents the available hypermedia resources' repositories available on the Web. These repositories include a variety of web resources such as video and audio files, web documents, images, presentations, news, wikis, maps, mobile applications, web services, podcasts, blogs, etc. A first prototype implementing some principles of the Web of Learning has been developed on a mobile platform [25]. The prototype generates learning objects for users on a mobile platform.

VI. Case Study

In order to illustrate how CEM annotates web resources let's consider the following scenario: Ayoub is a college student who is interested to learn how to graph quadratic functions. Although he has studied at school a full chapter on this topic, he would like to have a short and concise hypermedia presentation on his tablet computer about this topic. Ayoub submits the following query *"How to graph quadratic functions"*to the system which is first analyzed and forwarded to many hypermedia web repositories in order to search and retrieve relevant hypermedia web resources corresponding to the information requested. Contextual exploration analyzes the user's query in order to annotate it and extracts relevant words and phrases. The query objective is detected from the words "How to" hence the query is annotated as (*Objective = method*) which means that the user is looking for a method, a manner or a description of how to accomplish the specific task related to "graph quadratic functions". Accordingly, information to be retrieved should have the same objective tag so that to match the query. Table I shows part of the linguistic expertise that is used to annotate the query as "method". In the first column some linguistic expressions denoting the semantic category "method" are represented in the EBNF (Extended Backus-Naur Form) notation [26] where lower case words are terminals and capitalized words are non-terminals. For instance the non-terminal *Lmethod* refers to the list of words expressing a method such as: *method, technique, way, manner, practice, approach and procedure*. The non-terminals "Vcan" and "Vbe" represent all the possible morphological variations of the verbs "can" and "be". Other semantic categories are used for objective annotation such as: "definition" ("What are mobile agents?"), "cause" ("Why the screen is dark?"), "time" ("When Mona Lisa has been paint") and "location" ("Where can I store my files?").

Matching between the query and the resource is done through similar annotations. CEM analyzes the resources' titles, descriptions and metadata. Resources that are annotated similarly are retrieved and used in the learning object (LO). For some resources the comments left by users may also be analyzed. The linguistic expertise presented in Table I is specific for analyzing queries. It is different from the linguistic expertise to annotate texts present in the descriptions of resources.

TABLE I. Linguistic Expertise for the Query Objective

Query Objective		
Linguistic Expression	***Example***	***Annotation***
["how" + "to"]	How to save a file?	method
["how"+Vcan+Ppronoun]	How can I copy an image?	
[Ppronoun]	I, we, someone,	
["what"+"Vbe"+Lmethod]	What is the procedure to apply for a research grant?	
[Lmethod]	method, technique, way, manner, practice, approach, procedure.	

The following resources have been tagged by CEM as fitting with the objective of the query for the task “graph quadratic functions”:

- From Wikipedia[6], the system extract the table of contents and the first paragraph’s sentence which introduces the topic: “quadratic functions”;
- From Youtube[7], the first retrieved video is selected as it has been annotated as “method” due to the presence of the following sentence in the video description: *“I outline a little recipe of things to examine when graphing a quadratic function by hand.”*The underlined words have been spotted by CEM as relevant marker and clues for annotation;
- From Yahoo Images[8], the system retrieves the third image proposed has been annotated as “method”. The following sentence describing the image in the webpage (“online math learning”[9]) has been analyzed as denoting a “method”: *“In this lesson, we shall learn how to graph of quadratic functions by plotting points”.* The underlined words have been spotted by CEM.

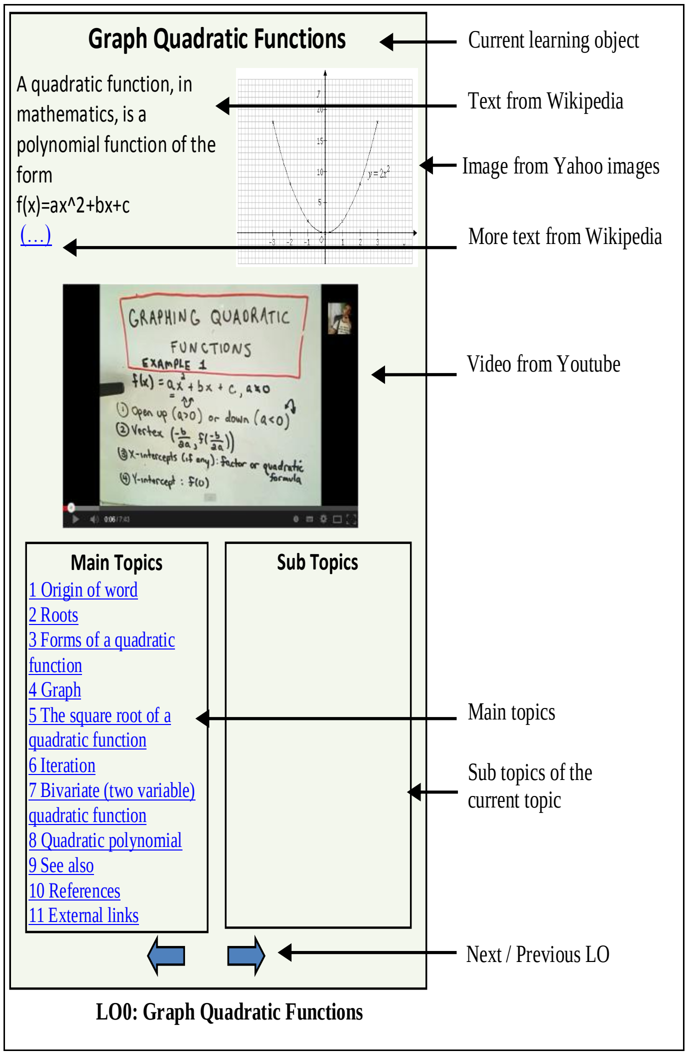
Resources are searched into the three popular websites: Wikipedia, Google Images and You tube. We have restricted the system to these websites in a first phase but we can extend itby considering additional APIs to access other websites looking for similar web resources. Once resources are annotated and retrieved, the system packages them into a LOas shown in Figure 3.

The learning object LO0 represents the first LO to display. LO0 includes the title on the top that corresponds to the phrase “Graph Quadratic Functions” which has been extracted from the query. LO0 includes a short text description about the topic that has been extracted from Wikipedia webpage. Beside the text the learning object displays an image from Yahoo images corresponding to the topic. In the middle the learning object includes a video that has been retrieved from You tube about graphing quadratic functions.

Moving from a LO to another one is possible from the navigation map represented into the lower rectangles “Main Topics” and Sub Topics” which include hyperlinks to all possible objects in the learning web of Figure 4. Navigation is not constrained on a specific pathway; the user is free to follow the normal sequence of LOs as suggested by the learning web in which case he should use the “Next” button (represented as an arrow).

He can also navigate randomly by clicking on any topic he wishes to view in which case he will be directed to the specific topic. When the user clicks on any topic or the “Next/Previous” button, he requests to view a LO corresponding to the topic in question. In such case the system generates the LO for the topic/subtopic requested by the user, displays it on the screen and updates the navigation status in the “Main Topics” and “Sub Topics” rectangles.“Main Topics” contain the topics which correspond to the main sections of Wikipedia Webpage. “Sub Topics” displays any sub sections of the currently displayed topic.

For instance, when Ayoub clicks on the topic “4. Graph”, LO4 is generated the same way as LO0 has been generated and is displayed on the screen. In this case “Sub Topics” will include the sub topic “4.1. Vertex” (Fig. 4).



Graph Quadratic Functions

A quadratic function, in mathematics, is a polynomial function of the form f(x)=ax^2+bx+c (…)

Main Topics

1 Origin of word
2 Roots
3 Forms of a quadratic function
4 Graph
5 The square root of a quadratic function
6 Iteration
7 Bivariate (two variable) quadratic function
8 Quadratic polynomial
9 See also
10 References
11 External links

Sub Topics

Current learning object
Text from Wikipedia
Image from Yahoo images
More text from Wikipedia
Video from Youtube
Main topics
Sub topics of the current topic
Next / Previous LO
y=2x²
GRAPHING QUADRATIC FUNCTIONS
EXAMPLE 1
f(x) = ax² + bx + c, a≠0
① Open up (a>0) or down (a<0)
② Vertex (-b/2a, f(-b/2a))
③ X-intercepts (if any): factor or quadratic formula
④ Y-intercept : f(0)
LO0: Graph Quadratic Functions

Fig. 3. Learning Object LO0: Graph Quadratic Functions

[6] http://en.wikipedia.org/wiki/Quadratic_function (accessed on 1st Nov. 2013).
[7] http://www.youtube.com/watch?v=mDwN1SqnMRU(accessed on 1st Nov. 2013)
[8] http://www.google.com (accessed on 1st Nov. 2013)
[9] http://www.onlinemathlearning.com/quadratic-graphing.html (accessed on 1st Nov. 2013)

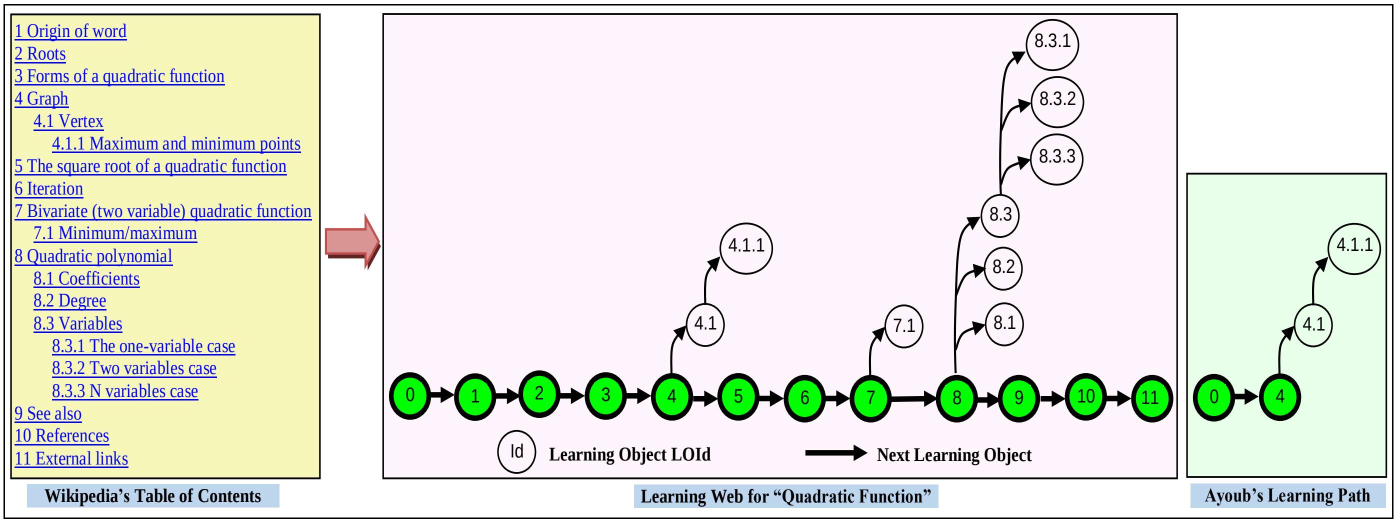


Fig. 4. Learning Web Generation and Learning Path

The system builds the learning web from Wikipedia[10] (Fig. 4). The table of contents of Wikipedia is a good learning structure that organizes the depth and breadth knowledge of a given topic. The Learning Web represents all the possible pathways where the user may navigate in while learning. In our example, Ayoub has visited LO0 then he requested LO4 and as he is interested in Graphing Quadratic Functions. Then he wanted to investigate deeply the topic "4. Graph", he requested the subtopic "4.1 Vertex" then its subtopic "4.1.1 Maximum and minimum points". Those LOs requested and visited by Ayoub constitute his personal learning path (Fig. 4 – right graph).

VII. CONCLUSION

In this paper we presented an approach for reusing available hypermedia web resources to design modern instructional environments that contribute towards the establishment of the Webof Learning. The approach sustains the generation of on-demand interactive and non-linear learning spaces for learners. Prior to be reused, web resources are annotated semantically by the contextual exploration method which analyzes the resources' text descriptions and associates semantic annotations that denote the role and function of the web resource. The implemented system generates on-demand learning material by packaging semantically correlated web resources into learning objects which are plugged into a course map that represents all the possible pathways a learner may navigate in. The system tests done on a set of different topics are encouraging, they show that automatic semantic annotation is more accurate than classical information retrieval in retrieving correlated hypermedia resources [23]. Besides, organizing the learning content into a learning web and allowing users to learn through sessions aremuch appreciated as they allows users to deepen and diversify their knowledge through durable learning.

Future research aim to extend the linguistic expertise to encompass more semantic categories to have an accurate matching between the user query and the web resources multimedia descriptions. As for the future extensions of the system, it is important to have a large web coverage by considering more hypermedia websites from where web resources will be considered for annotation and reuse. Furthermore, we would like to develop diversified and more intuitive LO layouts to adapt the packaged hypermedia content to fit the user profile and needs.

REFERENCES

[1] T. Berners-Lee, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web, HarperCollins, New York, 2000.

[2] E. Turban, D. King, Electronic Commerce: A Managerial Perspective, Prentice Hall, 7th Edition, 2012.

[3] J. Berri, The Web of Learning: Evolving the Web into a Learning Ecosystem, submitted for publication, 2013.

[4] P. Andrews, I. Zaihrayeu, and J. Pane, A Classification of Semantic Annotation Systems, Semantic Web, vol. 33, pp. 223-248, 2012.

[5] A. Zubiaga, V. Fresno, R. Martınez, and A. Pérez Garcıa-Plaza, Harnessing Folksonomies to Produce a Social Classification of Resources, IEEE Transactions on Knowledge and Data Engineering, vol. 25, No. 8, pp. 1801- 1813, August 2013.

[6] H. Q. Yu, C. Pedrinaci, S. Dietze, and J. Domingue, Using Linked Data to Annotate and Search Educational Video Resources for Supporting Distance Learning, IEEE Transactions on Learning Technologies, vol. 5, no. 2, pp. 130-142, April-June 2012.

[7] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data—The Story So Far," International Journal on Semantic Web and Information Systems, vol. 5, no.3, pp. 1-22, 2009.

[8] S. Boung-Yew Lau, and C.-S. Lee, Enhancing Collaborative Filtering of Learning Resources with Semantically-Enhanced Social Tags, 12th IEEE International Conference on Advanced Learning Technologies, pp. 281-285, Rome, Italy, 4 - 6 July, 2012.

[9] B. Smine, R. Faiz and J.-P. Desclés, Extracting Relevant Learning Objects Using a Semantic Annotation Method, International Conference on Education and e-Learning Innovations (ICEELI'2012), Sousse, Tunisia, 1-3 July, 2012.

[10] J. Hendler, T. Berners-Lee, and E. Miller, Integrating Applications on the Semantic Web, Journal of the Institute of Electrical Engineers of Japan, vol 122, no. 10, pp. 676-680, October 2002.

[11] M. Nagarajan, Semantic Annotations in Web Services, Semantic Web Services, Processes and Applications, J. Cardoso and A. P. Sheth Eds., Springer, pp. 35-61, 2006.

[10] http://www.wikipedia.org

[12] L. Causton, Identifying and describing Web resources, European Commission DGXIII/E-4, Interactive Electronic Publishing, 1998. Available at http://www.elpub.org/brochures/

[13] M. Krötzsch, D. Vrandecic, M. Völkel, H. Haller, and R. Studer. Semantic Wikipedia. Journal of WebSemantics, 5(4):251–261, 2007.

[14] J. Kahan, M.-R. Koivunen, E. Prud'hommeaux, andR. R. Swick. Annotea: an open RDF infrastructure forshared web annotations. Computer Networks, vol. 39, no. 5, pp. 589–608, August 2002.

[15] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov, KIM - a semantic platform forinformation extraction and retrieval. Journal of Natural LanguageEngineering, vol. 10, no. 3-4, pp. 375–392, September 2004.

[16] T. Berners-Lee, R. Fielding, and L. Masinter, RFC 3986, Uniform Resource Identifier (URI): Generic Syntax, 2005. Available at http://tools.ietf.org/html/rfc3986

[17] J. K. Aggarwal and Q. Cai, Human Motion Analysis: A Review, Computer Vision and Image Understanding, vol. 73, no. 3, pp. 428–440, March1999.

[18] G. Papari, N. Petkov, Edge and line oriented contour detection: State of the art, Image and Vision Computing, vol. 29,pp. 79–103, 2011.

[19] A.K. Jain, L. East, R.P.W.Duin, and M. Jianchang, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 , no. 1, pp. 4-37, January 2002.

[20] Desclés, J. –P.: Langages applicatifs, Langues naturelles et Cognition, Hermès, Paris, 1990.

[21] J. Berri, D. Maire-Reppert, H.-G. Oh-Jeong, Traitement informatique de la catégorie aspecto-temporelle", T.A. Informations, vol. 32, no. 1, pp. 77-90, 1992.

[22] J. Berri, M. Al-Khamis, Information Exploration Using Mobile Agents, WSEAS Transactions on Computers, vol. 3, no. 3, pp. 706-712, 2004.

[23] A. AlAwajy, J. Berri, Combining Semantic Techniques to Enhance Arabic Web Content Retrieval, International Conference on Innovations in Information Technology (IIT'13), Al Ain, UAE, pp. 141-147, March 17-19, 2013.

[24] G. Miller, WordNet: A Lexical Database for English. In Communications of the ACM., vol. 38, no. 11, pp.39-41, 1995.

[25] M. Alzaabi, J. Berri, and M. J. Zemerly, Web-based Architecture for Mobile Learning, International Journal of Infornomics, vol. 3, no. 1, pp. 213–222, 2010.

[26] R. W. Sebesta, Concepts of Programming Languages, 10th Edition, Addison-Wesley, 2012.

An Assessment of 3G Mobile Service Acceptance in Bangladesh

Tajmary Mahfuz and Subhenur Latif
Department of Computer Science and Engineering
Daffodil International University
102, Shukrabad, Dhaka, Bangladesh

***Abstract*—This paper aims to find out the key factors influencing mobile users to adopt 3Gtechnology and affecting the subscriber's feedback while using third generation (3G) mobile services that are available for one year in Bangladesh. An interesting fact that motivated this research was the significant low rate of 3G service usage among mobile operators in Bangladesh though we get the completely opposite picture worldwide. To examine the user acceptance and to depict user behavioral pattern, data were collected from 200 respondents through a survey. The analysis was done into two categories: one was in general and the other one was department based. The results of the study revealed the user intention, awareness, attitude, expectation, key 3G service usage etc. The findings have future implications for existing as well as newly arrived service providers who have very recently started their journey. Considering these identified factors would provide the directions for telecom operators to achieve high rate of 3G service adoption and to provide more successful 3Gservices.However, the study covered a limited area where those findings are applicable. The result of this study might be helpful for the telecom operators while targeting the 3G subscriber market and also for the future research on this field.**

***Keywords*—Awareness; Adoption; 3G mobile service; usage pattern**

I. Introduction

3G was a long awaited thing that created high expectations among the mobile operators before its arrival in Bangladesh. There has been a steady growth in worldwide 3G mobile adoption. To shine with the 3G growing world, Bangladesh had started its journey of 3G on 14 October, 2012 by Teletalk, the government-owned telecom operator. About a year later, the number of subscribers has not been increased significantly as expected (around 4lac).Despite the availability of 3G services, basic mobile services are still the most popular services. Bangladesh held its first 3G spectrum auction in September, 2013 where four more telecom operators had been awarded 3G spectrum (Grameen Phone, Airtel, Banglalink and Robi) along with Teletalk.

The development of mobile services – or mobile commerce or mobile Internet – has been intense for years but adoption has not progressed as expected [13]. Many studies have investigated the user acceptance and success factor of mobile services in general, and 3G in particular [1]. Research on 3G technology acceptance in Bangladesh will therefore be extremely worthy in providing useful information, especially at this early stage of 3G mobile internet development and implementation in this country. So far, no such research has been done on this area. Hence, the purpose of this study is to examine factors affecting subscribers' acceptance towards using 3G mobile service. From this study, the mobile service providers could use the findings to understand user demand and behavior.

II. Literature Review

Recently 3G services are tremendously developing. This section reviews literature related to current usage, customer behavior, customer segmentation, acceptance and features that affect usage in various countries, although some authors have presented their interpretations regarding the future of 3G and its prospective.3G services were first adopted in Japan in 2001. Deepti and Ajay present the patterns, awareness and adoption of 3g users among young generations in Botswana [1]. Research conducted in Malaysia by suki [2], suggests that, to adopt 3G mobile services,the 3G mobile telecommunication companies need to lift consumer's intention.

User's of 3G mobile services need to be offered with more diverse and entertaining ways of communicating, which are at the same time easily accessible and convenient to use. Kim [3] recognize various services of 3G like video calling, on line TV, global roaming and advance services via the mobile multimedia Internet for magnetizing mobile phone subscribers. Pagani [7] focused the reasons of adoption 3G and ranked "Price" as third after "usefulness" and "ease of use",

According to Greek market perspective, about the charges of 3G, Indrawati, S. Murugesan, and M. A Raman Chatziagapis [8] infer that mobile services may pledge revenue growth for the operators, but the features of usefulness, security and especially the price of the mobile services have to be considered seriously in order for future adoption. Pagani [7], Indrawati et al [8] have also found price as a determinant factor for 3G mobile services adoption.

Moreover, modern services are enhances by using 3G services. Like smart home, wireless intelligence video system. [5], [6].

Several facilities have been provided by 3G users. It allows simultaneous use of speech and data services [4]. The main services of 3G like high speed data transmission, entertainment and e-payment are interrupted due to lack of infrastructure in Japan, developing countries in Asia, Africa or even some parts of US. High subscription charges, earnings affordability, mobile network coverage and telecommunication transportation to maintain all these activities classified with regard to findings is difficult for developing or even some developed countries in the world [9]. Li-Chen Cheng, Li-Min Sun, [10] proposed some diverse varieties of brand new application services to attract the new 3G subscribers. Despite of various benefits provided by the 3G services, it has not received great adoption rate as expected.

Margheaita [11] sketches a model of consumer adoption of third generation mobile multimedia services, by a qualitative exploratory study and empirically test the proposed model on the Italian market. Moreover Ong [12] investigates the factors affecting the purpose to adopt 3G services among the university students in Malaysia as they expected to be the group with great potential to adopt 3Gservices.

III. Methodology

In this paper we have focused on the grounds, demands of young generation who are actually concerned on new technology like 3G. Our work suggests that there is great deal of research on adoption of mobile phones and mobile related services. This study intended at reviewing the a wareness and usage of new 3G mobile services like high speed internet [14], mobile device features and services [15], like video calling, online TV, etc., usage of mobile applications [8, and 16] and usages of mobile data services [17].

Also emphasis on behavioral intention to use [13], and approach towards3G mobile services in Bangladesh. A large scale of subscribers have adopted some basic mobile services such as SMS, ring tones, icons, wall papers, logos, caller tunes[18] and these services became their everyday's life styles.

Adoption patterns, present situation, problems and requirements of the 3G subscriber's have been tried to discuss in this paper and have sketched the overall scenario by reviewing the responses of the 3G usages.

IV. Data Collection And Analysis

In total, 200 students of Faculty of Science and IT from first to fourth year of the graduation participated in this study, by completing the questionnaire. The age range was from 20 to 25 years. Each of the participants was 3G user having a 3G supported cell phone. Since for providing 3Gtechnology, there was only one available telecom operator during the time of concerned research, each respondent was a subscriber of that operator.

There were two aspects of the research work. One part of focus was the general data analysis where the concerned area was the whole user domain. The other aspect was department wise analysis where each department was individually taken as area of interest and data were organized according to the departments. Through questionnaires and the statistically analysis has been by using SPSS.

A. *General data analysis*

70% users used mobile internet before adopting 3G while 30% users started using mobile internet with 3G.

Based on the use of internet connection, two categories of 3G users found. **Category-1** users have both internet connection at home and 3G connection (65%) and **Category-2** (35%) users do not have other internet connection that is only 3G users. Among **Category-1** users, 38.5% do browsing, 76.9% do uploading and 92.3% do downloading (Fig:1).

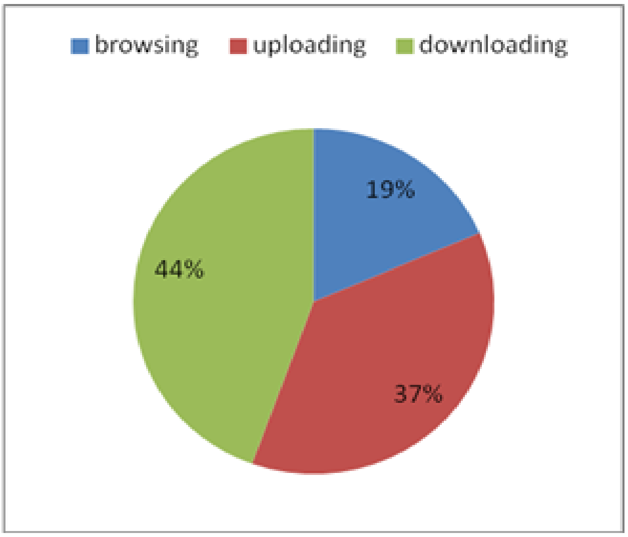


Fig. 1. using different services at home by using 3g connection.

Among the reasons behind adopting 3G, 95% users use 3G for better services, 45% respondents use 3G influenced by their friends and the rest 25% had other reasons. Of all 3G services, the most used service was *speed of data transfer(89.5%)*, followed by *gaming application(78.9%)*, then *video calling(36.8%)*and lastly *mobile TV*(31.6%). For barriers, *poor network coverage* came out as the main problem (100%) of using 3G.*High rate of charging* rated as second (90%) followed by *insufficient service provider* (50%) and *lack of high speed* (35%).Another point was user expectations or desired facilities of 3G. Most desired service was *strong network coverage* (100%), followed by *low charge rate*(95%) and *high speed*(50%).

B. *Department-wise analysis:*

Four departments were considered for analysis: Computer Science and Engineering (CSE), Electrical and Electronic Engineering (EEE), Textile Engineering (TE) and Software Engineering (SWE. The scenarios of department wise 3G users were: CSE-40%, TE-30%, EEE-15% and SWE-15%.Onthe point of use of mobile internet before 3G, department wise user behavior had been recognized. The results were: For CSE department, 42.9% had used, 33.3% had not used. For TE department, 21.4% had used, 50% had not SWE and TE. The scenarios of department wise 3G users used. For EEE, used-14.3%,not used-16.7%. For SWE, 21.4%hadused. Then we figured out department wise most frequently used services(Table 3,4). Table 3 shows most frequently used services of individual departments. For CSE department, the highest used service was both speed of data transfer(75%) and gaming application(75%). For all the three departments of TE, EEE and SWE, speed of data trans ferranked as the top most used service.

TABLE I. DEPARTMENT-WISE MOST FREQUENTLY USED SERVICES ON 3G

Department	*Most used Services*	*Response %*
CSE	Speed of data transfer	75
	Mobile TV	12.5
	Video calling	25
	Gaming application	75
TE	Speed of data transfer	100
	Mobile TV	50
	Video calling	50
	Gaming application	83.3
EEE	Speed of data transfer	100
	Mobile TV	33.3
	Video calling	33.3
	Gaming application	100
SWE	Speed of data transfer	100
	Mobile TV	50
	Video calling	50
	Gaming application	50

TABLE II. DEPARTMENT-WISE PROBLEMS ENCOUNTERED BY THE USERS

Department	*Most used Services*	*Response %*
CSE	High charging rate	87.5
	Poor network coverage	100
	Lack of high speed	50
	Insufficient service provider	75
TE	High charging rate	83.3
	Poor network coverage	100
	Lack of high speed	33.3
	Insufficient service provider	33.3
EEE	High charging rate	100
	Poor network coverage	100
	Lack of high speed	33.3
	Insufficient service provider	33.3
SWE	High charging rate	100
	Poor network coverage	100
	Insufficient service provider	33.3

V. FINDINGS AND DISCUSSION

From the empirical analysis, we have found some interesting results leading to informative facts.

Though many people in Bangladesh have heard about 3G from media and other sources, the total number of 3G users is very poor. The structured and unstructured interviews with 3G nonusers revealed the reasons. One main factor for this is the insufficiency of service provider since at first only one telecom operator was permitted for providing 3G services. Another strong reason is; the short time period for 3G service availability. Therefore, it is understandable that as like as the arrival of any new technology, the rate of 3G adoption that is the number of subscribers may increase with time.

In the case of adopting 3G,variations in user background showed different outcomes. The rate of 3G adoption is quite high for those who previously used mobile internet comparing to the non users. While digging for reason, it came out that upgrading to higher standards or versions happens naturally for an existing system user. On the contrary, this is not the case for a fresher as it demands to deal with a completely new thing.

Another observation of user behavioral analysis is that the users who were using other internet connection along with 3G had chosen ‘downloading’ as their most prioritized activity compared to others. Now the question arises that why those users are keeping additional internet connection when they can use 3G.Price is an issue here. Users have to pay more for unlimited data volume in 3G where there are more options in cheaper rate.

3G provides many attractive and additional features like as video calling, mobile TV and of course, better services than the existing system e.g. high speed internet. All these things allure users to subscribe to 3G. The users of 3G voted high speed data transfer as their most used service. Poor network coverage was the biggest barrier that interrupts the enjoyment of using 3G services at great extent. Therefore, it is no wonder that the most desired service of 3G subscribers is the strong network coverage.

VI. LIMITATIONS AND FUTURE SCOPE

Our study has some limitations. All the respondents were all most same age group and came from almost same background, which is one of the limitations. The study is based on a limited number of respondents which is a limitation. To interpret the behaviors of all mobile phone users, the result cannot be generalized. Despite these limitations, however, this study provides insights into the adoption behavior of 3G services. For future analysis, bigger sampling data would be considered. Comparison based study could be conducted on the pattern changes in 3G service usage in time. Also, uses of high technology based on 3G like smart home, could be perform in future study.

VII. CONCLUSION

In our research paper, we have sketched the adoption scenario of a new technology called 3G. The penetration of the usage and adoption of 3G mobile services has been done in this research. The recent addition of four new telecom operators in 3G market of Bangladesh has created the competitive and challenging field in service providing. This phenomenon demands to adopt the correct marketing strategy and business model to catch on the potential customers. In this perspective, the findings of this study provide the directions. The service providers should concentrate on minimizing the negative factors at highest possible rate like poor network coverage that badly affect the user. Exploitation of cost and providing high speed constantly are some key demands that needed to be addressed.

REFERENCES

[1] Deepti Garg, Ajay K. Garg,” An Assessment of Awareness, Usage Pattern and Adoptionof 3G Mobile Services in Botswana,” *International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011.*

[2] Suki,”Third generation (3G) mobile service acceptance:Evidencefrom Malaysia,”http://www.academicjournals.org/AJBM.

[3] Y. Kim, “Estimation of consumer preferences on new telecommunications services: IMT-2000 service in Korea” Information Economics and Policy, vol.17. pp.73-84, Jan 2005.

[4] Dr. Sudha Singh, Dr. D. K. Singh, Dr. M. K. Singh and Sujeet Kumar Singh [4],"The Forecasting Of 3g Market In India Based On Revised Technology Acceptance Model," International Journal of Next-Generation Networks (IJNGN) Vol.2, No.2, June 2010.

[5] Gwang Jun Kim, Chang Soo Jang, Chan Ho Yoon, Seung Jin Jang and Jin Woo Lee,"The Implementation of Smart Home System Based on 3G and ZigBee in Wireless Network Systems,"International Journal of Smart Home Vol. 7, No. 3, May, 2013.

[6] *Qigui ZHANG, Yu CHEN,"* Design of Wireless Intelligent Video Surveillance System Based on 3G Network," http://iaesjournal.com/online/index.php/TELKOMNIKA/issue/view/169, vol 12, No 1,2013.

[7] M. Pagani, "Determinants of Adoption of Third Generation Mobile Multimedia Services," Journal of Interactive Marketing, Vol. 18(3), pp.46-59. Summer 2004 [Online] Available ,http://www.interscience.wiley.com

[8] Indrawati, S. Murugesan, and M. A Raman "New Conceptual Model of Mobile Multimedia Services (MMS) and 3G Network Adoption in Indonesia," International Journal of Information Science and Management Special Issue January / June, 2010.

[9] S.T. Abu, "Empirical analysis of global diffusion of 3G mobile phones:a cross-cultural review," Discussion Paper No. AIDP0906. Graduate School of Applied Informatics, University of Hyogo, Japan 2010

[10] Li-Chen Cheng, Li-Min Sun," Exploring consumer adoption of new services by analyzing the behavior of 3G subscribers: An empirical case study," Electronic Commerce Research and Applications 11 (2012) 89–1002011 .www.elsevier.com/locate/ecra.

[11] MargheaitaPagani,"depart of adoption of third generation mobile multimedia servies,"Journal of Interactive Marketing ,Vol. 18,No. 3,Summer 2004.

[12] J. W. Ong, Yew-Siang Poong and Tuan Hock Ng, "3G Services Adoption among University Students:Diffusion of Innovation Theory," Communications of the IBIMA Volume 3, 2008.

[13] ChristerCarlsson, KaarinaHyvönen, Petteri Repo and Pirkko Walden," Adoption of Mobile Services across Different Technologies,"18th Bled e Conference e Integration in Action, Bled, Slovenia, June 6 - 8, 2005.

[14] P. Jiang "Consumer Adoption of Mobile Internet Services: An Exploratory Study," Journal of Promotion Management, vol. 15 (3), pp.418-454, 2009.

[15] A. A.Economides, and A. Grousopoulou, "Students' thoughts about the importance and costs of their mobile devices' features andservices," Telematics and Informatics, vol. 26 (1), pp. 57-84, 2009.

[16] H. Verkasalo, C. López-Nicolás, F. J. Molina-Castillo, H. Bouwman: "Analysis of users and non-users of smartphone applications,"Telematics and Informatics, vol. 27(3): pp. 242-255, 2010.

[17] J. Harno, "Impact of 3G and beyond technology development and pricing on mobile data service provisioning, usage and diffusion," Telematics and informatics, vol. 27 (3), pp. 269-282, August 2010.

[18] C. Carlsson, J. Carlsson, K. Hyvönen, J. Puhakainen and P. Walden, "Adoption of Mobile Devices/Services – Searching for Answers withthe UTAUT", in Proc. of the 39th Hawaii International Conference on System Sciences, HICSS, 2006.

CAPTCHA Based on Human Cognitive Factor

Mohammad Jabed Morshed Chowdhury
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh

Narayan Ranjan Chakraborty
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh

***Abstract*—A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is an automatic security mechanism used to determine whether the user is a human or a malicious computer program. It is a program that generates and grades tests that are human solvable, but intends to be beyond the capabilities of current computer programs. CAPTCHA should be designed to be very easy for humans but very hard for machines. Unfortunately, the existing CAPTCHA systems while trying to maximize the difficulty for automated programs to pass tests by increasing distortion or noise have consequently, made it also very difficult for potential users. To address the issue, this paper addresses an alternative form of CAPTCHA that provides a variety of questions from mathematical, logical and general problems which only human can understand and answer correctly in a given time. The proposed framework supports diversity in choosing the questions to be answered and a user-friendly framework to the users. A user-study is also conducted to judge the performance of the developed system with different background. The study shows the efficacy of the implemented system with a good level of user satisfaction over traditional CAPTCHA available today.**

Keywords—CAPTCHA; Usability; Security; Cognitive; Psychology

I. INTRODUCTION

As more people are using Internet as a daily basis, the requirement of online services is also increasing. Many services in the internet including email, search engine, social networking are provided with free of charge. With the limited available resources, there are some cases when available services are delayed or even denied. With the expansion of web services, denial of service (DoS) attacks by malicious automated programs (e.g. web bots) is becoming a serious problem as masses of web service accounts are being illicitly obtained, bulk spam e-mails are being sent, and mass spam blogs (splogs) are being created. In order to avoid tremendous attack from malicious computer programs, CAPTCHA has been introduced to distinguish humans from computers.

Moreover, most of the online services now require users to register for identification. Most of the servers can serve a limited number of users at a given point of time. So for the consideration of performance and security, it is required to distinguish between human user and computer programs. CAPTCHA system is mainly based on the assumption that human are superior to machine to understand the images and symbols. But with the advancement of technology in text recognition and image extraction, it is now possible to extract the characters shown in CAPTCHA with satisfied accuracy. To cope up with this threat, the CAPTCHA is introduced to oblique some sequence of characters that has become really harder for a normal human being to recognize. Thus a new design concept of CAPTCHA system is a necessity.

In short, CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a class of programs that is used to differentiate between human and computer programs. This classification is done through generation and grading of tests that are supposed to be solvable by only humans [4, 5]. There are some properties defined in development of CAPTCHA [15].

- Automated: Computer programs should be able to generate and grade the tests.
- Open: The underlying database(s) and algorithm(s) used to generate and grade the tests should be public. This is in accordance with the Kerckhoffs"s Principle [25].
- Usable: Humans should easily solve these tests in a reasonable amount of time. The effect of any user's language, physical location, education, and/or perceptual abilities should be minimal.
- Secure: The program generated tests should be difficult for machines to solve by using any algorithm.

There are two major issues should be considered while designing a successful CAPTCHA system: (1) robustness (difficult to break) and (2) usability (human friendly). In this paper, an idea of human cognition based CAPTCHA is presented considering above mentioned requirements. This is based on the concept that, humans can perceive the meaning of cognitive questions and answer them. But there is no such algorithm that can be used to answer those in absolute accuracy. On the subsequent section of related works proposed schemes of CAPTCHA is discussed. In the third section implementation details and the very next section survey and analysis are discussed. Before conclusion and future work result of the proposed system is discussed.

II. RELATED WORKS

Research on CAPTCHA mechanisms has received significant attention with the aim to improve their usability and at the same time prevent adversarial attacks by malicious software. Researchers promote various CAPTCHA designs based on text and speech-recognition challenges, and image puzzle problem [1]. Nevertheless, a variety of studies have been reported that underpin the necessity for improving the usability of CAPTCHA mechanisms [4,5,6]. Result from a recent study, which investigated users' perception towards

CAPTCHA challenges; claim that current implementations do not provide an acceptable trade off solution with regards to CAPTCHA usability [2]. Another large-scale study, which evaluated CAPTCHA on the Internet's biggest websites, revealed that CAPTCHAs are difficult for humans to solve [3].

The algorithms and data used to automatically generate these CAPTCHA challenges are publicly available. But with the advancement of OCR and sophisticated image processing algorithms and tools, these text based CAPTCHAs can no longer provide the secure access to the authenticate users from malicious computer programs [16,21,24]. For instance, researchers have developed an attack against Microsoft’s Hotmail CAPTCHA that yields a 60% success rate [22]. Also more complex image distortion to make it difficult for programs to crack, makes this text based method increasingly hard for human users to recognize the text, causing usability issues [23, 29].

Thus the need for new form of CAPTCHA which is automated, open, usable, and secure is of urgent need. There are some other implementations of CAPTCHA available as of now. Mostly these can be separated in three categories. Except text based scheme, there are also sound and image based CAPTCHA schemes are available. Audio based CAPTCHA was first developed for visually-impaired people [6,8]. Audio CAPTCHA usually pronounces letters or digits in randomly spaced intervals. Background noises may be added to make the tests more robust against bots. These systems are dependent on some sort of audio hardware to produce the sound clearly, and these sounds are sometimes difficult to perceive for locality reasons. Also persons with hearing difficulties cannot use this scheme. Furthermore, the basic principle to attack this CAPTCHA remains similar as text based ones, which is to extract the feature and recognize the letters. Hence, the audio based CAPTCHA scheme does not provide any more user-friendliness or robustness against bots than text based CAPTCHA [7]. In [28], a new game theorem based CAPTCHA system is proposed.

Image based CAPTCHAs inquire users to perform some forms of image recognition tasks. These systems are developed to overcome the shortcomings of previously discussed schemes of CAPTCHAs. There are some schemes that use human ability to perceive and semantically analyze images to perform a task [19]. Users are asked to categorize distorted images using noises [10] or geometric transformations [15]. There are also some methods that ask users to adjust the orientation of 3D images or to identify semantic meaning from it [13, 20]. Microsoft’s Asirra [11] was designed to use the existing database of petfinder.com and prompts users to identify images of cats out of other pets. But the availability of the database and on the top of that as being it is a classification problem; Asirra is vulnerable to Machine learning attacks [12]. But content-based image retrieval and annotation techniques have shown to automatically find semantically similar images or naming them. These will allow an affordable mean of attacking image-based CAPTCHAs. User-friendliness of the systems is also a compromised factor when repeated responses are required [9] or deformed face images are shown [17,22]. Also people with color blindness have problems to figure out the distorted images. Some works are done on video CAPTCHAs [14]. Cognitive Psychology based CAPTCHA is presented in [27]. However requirement of bandwidth and difference in perception by users may be an issue.

There are some works available [18] where question-answer based CAPTCHA is shown. The author introduces the idea to use question answer method. The work in this paper is different as, it introduces the idea of different types of question rather than simple straight forward question. The authors only put the mathematical questions and hence, very few diversity in choosing the domain of the questions. Some people do not fond of answering the mathematical questions. Our proposed work also introduces a framework, where users will be given the scope to choose question group according to his/her preference. Again, as the color images are incorporated in the questions, this solution will not work for human being with color blindness. Our developed model overcomes these flaws and provides a well-accepted solution.

III. Proposed Model

As mentioned in the related work section different types of CAPTCHA systems are used to secure web browsing. None of them gives any choice to the user to select the types of CAPTCHA. The proposed model provides the user a big window of flexibility to choose the types of CAPTCHA. By nature, humans are very responsive to answer questions. In the proposed model, a user will be provided with 5 types of CAPTCHA questions, namely, analytical, mathematical, general, text based and image based. User can select any one of the option. This system will provide 10 minutes to solve the CATPCHA problem. We want to restrict the time to prohibit the machine to analyze any single question. If enough time is given then machine can solve complex problems using artificial intelligence and pattern recognition. In that case we have to give more complex challenges to solve. It will surely affect the usability and user friendliness negatively.

Analytical type challenges provide simple analytical problems to the user. Few predefined questions are set with answers. If someone fails to give answer within the time frame another challenge is given, on the other hand if the user thinks the question is difficult to answer then try another button helps her/him to change the question instantly and reset the time. If user wants to solve mathematical problems s/he can choose mathematical option. Few simple mathematical problems are asked. The description of the mathematical questions is such that it is very easy for human being to interpret but will be very difficult to analyze by machines. User or visitor of the website has to answer within the given time limit. General type category gives the user very simple question to answer. Similar rules like analytical and mathematical types are also applicable here.

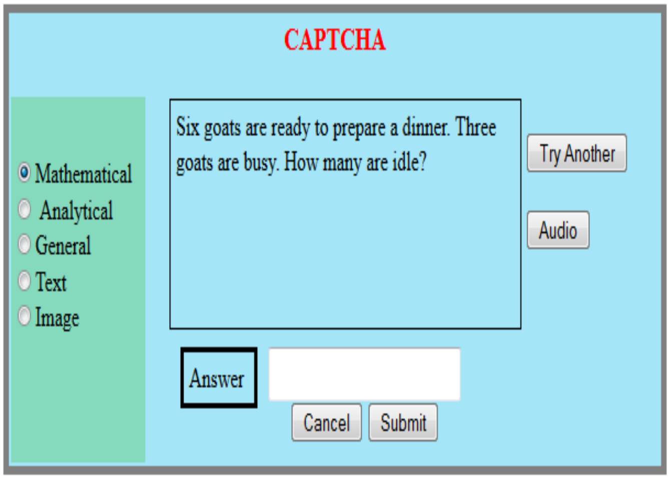
Text and image based CAPTCHA is very common now a day. Text based CAPTCHA is kept in our system because of its familiarity among users. The proposed model also provides this facility to enter text and image based CAPTCHA based on the choice from the user. For blind people, audio version of CAPTCHA is also available as an option. Analytical, mathematical and general type challenges provide the audio facility so that the blind user can choose these options.

IV. Implementation

We have developed a system with few sample questions. The sample questions are for the test purposes; these should be fine grained before used in real applications. A sample scenario of the proposed model is shown in figure 1.If the user choose mathematical category a question is appear on the screen. This type of question can be answered within couple of seconds by the user but for a bots it is very difficult to answer and also take a long period of time. For example, consider the following question of mathematical category.

“Rahim has three bananas, Karim has five apples, Sikder has seven mangos. Jamal wants to buy three apples. How many apples left to karim?”

It is clear from the above mentioned question that the user can give the correct answer easily but it is very difficult for a machine.



CAPTCHA

● Mathematical
○ Analytical
○ General
○ Text
○ Image

Six goats are ready to prepare a dinner. Three goats are busy. How many are idle?

Fig. 1. Proposed CATCHA System

System receives the answer form the user and compares it with the answer stored in the database. User can give the answer in any case (e.g., uppercase or lowercase), before comparing with the database the system makes all answers in lower case. For the incorrect answer error message is shown and it asks to try another one or change the category. Next option is given based on the user’s choice. Table 1 shows few sample questions of each category.

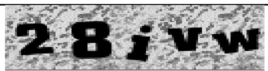
V. Survey And Analysis

A. Initialize the survey

To check the acceptance level and usability of the proposed model a session has been organized with 100 students and teachers from different departments of Daffodil International University. Along with 10 teachers, 20 Students from Computer Science and Engineering department, 15 from Electrical and Electronics Engineering department, 15 from Software Engineering department participated in the session. Other 40 students are from English, law, BBA and journalism departments. The main focus of this session is to collect the user’s opinion about the newly designed framework.

First the new framework is given to the participants to explore without giving any instructions. After playing with the system a set of questions is given to them and asked to give their feedback. Each question has four options like strongly agree, agree, partially agree and disagree.

TABLE I. Sample Questions Of Each Category

Category	Questions
Analytical	Mina had orange and mango. Mina ate orange. Which fruit is left?
Mathematical	Karim's age is one third to his father. His father age is 45. How old karim is?
General	In which direction does the Sun rise?
Text	
Image	

VI. Result Analysis

As discussed above out of 100 participants 50 are from technical background and 50 from non technical background. Responses from both groups are collected separately for better justification of the system. Concerns of the female participants are also taken separately. Following charts shows the impact of our system. Question by question analysis is given below.

B. Technical Perspective

Firstly, we will analyze the survey results from the technical perspective. We will try to find if there is any significant difference between technical and non-technical people. We are interested about the non-technical people because in today’s digital world is dominated by non-technical people.

First question was asked to the participants about whether the new framework is easy to use or not? As shows in the figure 2, 66% technical knowledge based participants strongly agree that the new framework is easy to use whereas 70% general participants strongly agree with that, 28% technical people agree in contrast with 20% non technical. 4% partially agree where none from non technical partially agree with that. Rest 2% technical people said it is not easy to use where it is 10% from non general people. So, from this perspective we can conclude that although non technical people are more in favor of this kind of system but some of them are against of this new system whereas all the technical people are in support of the new system.

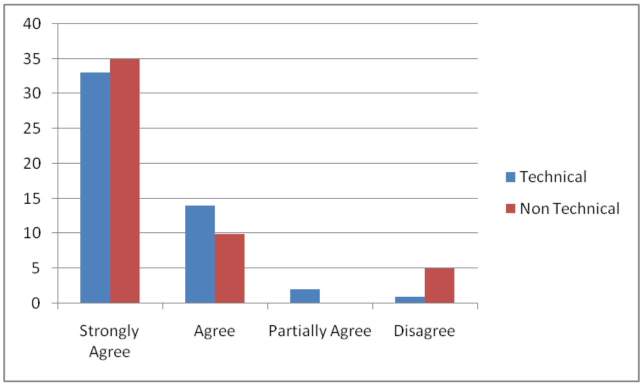


Fig. 2. : The new framework is easy to use or not

Second question was asked about the probability of making mistakes compared to existing CAPTCHA system (reCaptcha [26]). 70% of technical and 80% of non technical people think that the new system has less chance to make any error. 22% and 10% also agree. 4% of both technical and non technicalsurveyee shows their disagreement. Figure 3 reflects the opinion.

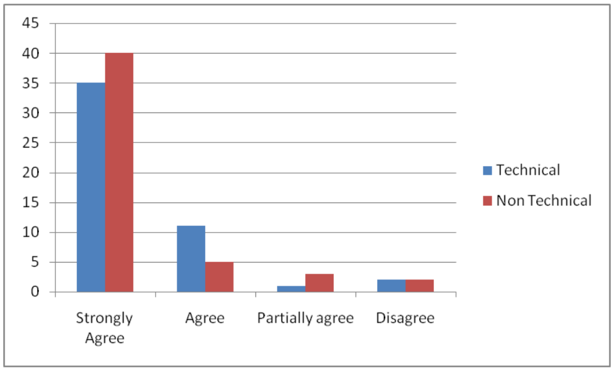


Fig. 3. This system has less chance to make error

Third question was asked about the brain work to solve any given problem. 96% participants from technical group agreed that brain work is needed to solve the given problem whereas 20% participants from non technical group agree. 60% non technical people strongly agree with that the system need brain work. Rests of the participants think differently. Figure 4 shows the result.

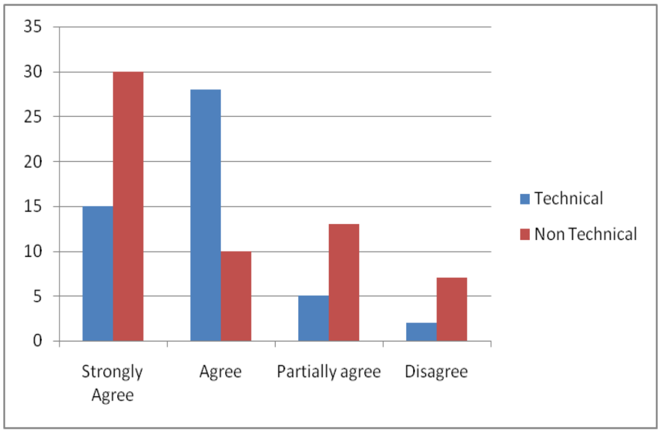


Fig. 4. Do you need any brain work to solve the problem given?

Fourth question was asked to the participants that the system is user friendly and helpful or not? From figure 5 it is clear that almost similar percentage participants from both technical and non technical group strongly agreed that the new system is very much helpful as well as user friendly. No one express their dissatisfaction. This shows strong support for the new proposed system.

Fifth Question was whether the system is better for all types of user or not? Figure 6 gives a clear idea that 80% and 90% form technical and non-technical group respectively strongly agree that the new system is better for all types of user. No one express their dissatisfaction.

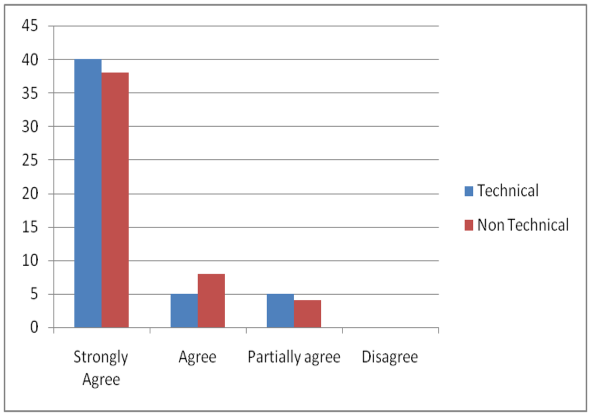


Fig. 5. The system is user friendly and helpful?

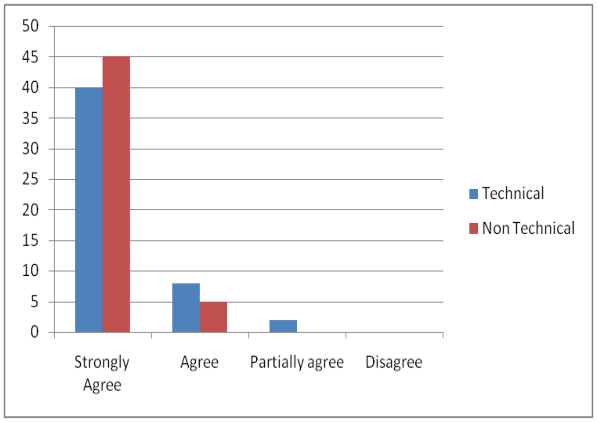


Fig. 6. System is better for all types of user or not?

Sixth and the last question was to check the satisfaction level of the user. 68% participants from technical group and 94% from non technical group strongly agree that they are satisfied with the system. 24% are agreeing on that. 6% and 4% are partially agreed. Rests of them are disagreed. Figure 7 shows the proof.

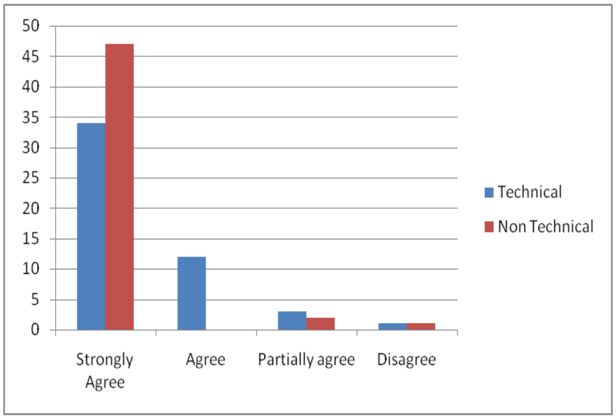


Fig. 7. Are you satisfied with the system?

So from the analysis of the technical perspective we can conclude that both technical and non-technical people are in strongly in favor of the proposed system. Specially, non-technical people are more in favor of such system.

C. Gender perspective

We have also investigated the gender factor in CAPTCHA system. 20 females have participated in the survey where 10 of them are from non-technical background. Almost 95% from non technical group express their deep satisfaction about the proposed system. More than 80% from technical group has given their consent on satisfaction. Based on the data collected form female participants it is clear that they are biased on the design and easiness of the system where as the male participants and also technical group people concern about the effectiveness of the system. Two questions were selected to test the satisfaction level of female participants.

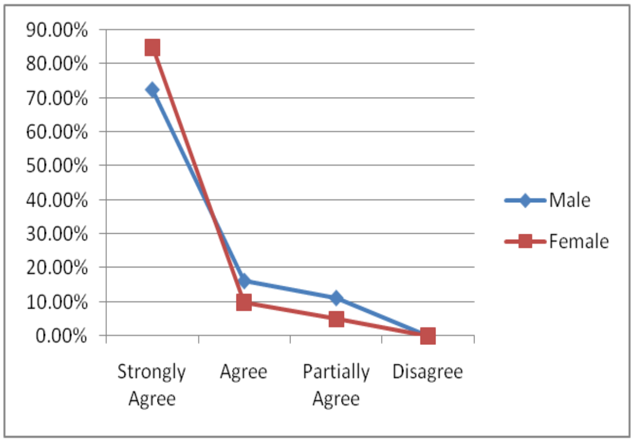


Fig. 8. System is user friendly and helpful

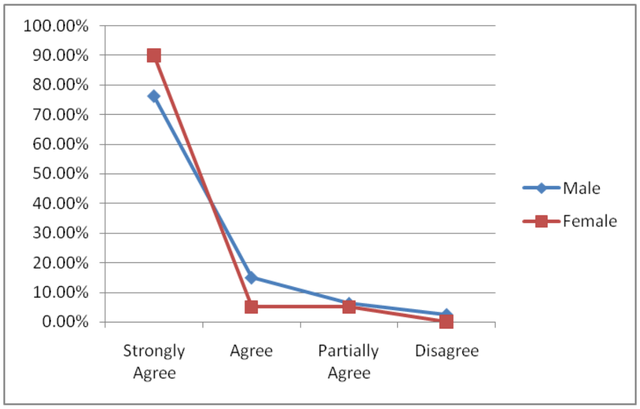


Fig. 9. Are you satisfied with the system?

From the above two figures (Figure 8 and figure 9) it is clear that female participants are more positive about the satisfaction, friendliness and usefulness of the system rather than male participants.

D. Time Perspective

We have also analyzed the time required to answer each type of question. From Table 2, we can easily see that Text based question (current reCAPTCHA) need more time compared to any other types. The response is subject to the type of questions and may vary slightly for different set of questions.

TABLE II. TIME TAKEN TO ANSWER THE QUESTIONS

	Q # 1	Q # 2	Q # 3	Q #4	Q # 5	Average time (in seconds)
Mathematical	7.99	9.67	7.51	6.73	6.48	7.76
Analytical	5.33	4.63	7.54	6.12	4.25	5.57
General	2.53	2.45	4.89	3.24	3.15	3.25
Text	9.36	8.51	11.87	7.95	10.82	9.70
Image	4.87	4.62	5.43	5.98	5.33	5.24

*** Q#1 means question 1 and so on**

E. Likeliness to attempt any category

We have identified the likeness of the user for any category of questions. From our survey it has revealed that most of the people like the general questions and least number of people attempted the text based system. Figure 10 reflects the likeliness of the participants.

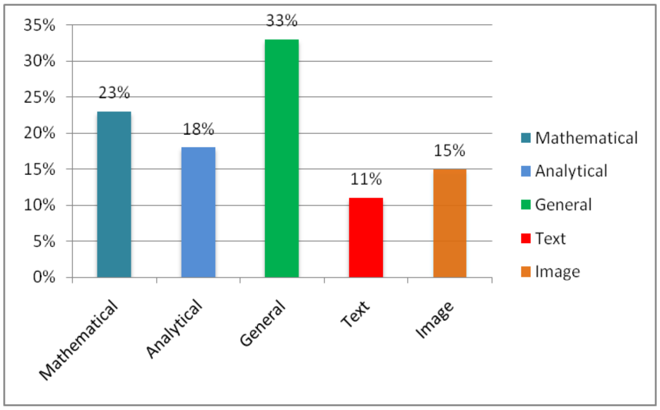


Fig. 10. Likeliness of Participants to attempt any Category

Table 3 shows the mode and means value of the selected 6 questions. The result is also categorized in technical and non-technical background to cover the diversity of the participants of the users. With analyzing the mode value for all participant data, it can be derived that most people strongly agreed on the given set of questions. Hence, it implies that the proposed model is accepted by majority of the participants. The mean value also justify this same results.

From technical perspective it is clear that the mean and mode value of all question asked in the survey is near about 3. That indicates the high acceptance and effectiveness of newly proposed system among technical people. For the non-technical group, the mean values of responses are greater than technical people all the questions. This clearly indicates the efficacy and usability of the framework among majority of the common people. The questions in the questionnaire are designed to reflect 3 criteria (Learn ability, Efficiency and Satisfaction) of usability study.

TABLE III. MEAN AND MODE OF REPLIES OF THE TECHNICAL AND NON-TECHNICAL SURVEYED.

Question number	For Technical participants		For non Technical participants	
	Mean	Mode	Mean	Mode
1	2.48	3	2.4	3
2	2.28	3	2.76	3
3	1.51	3	2.32	3
4	2.45	3	2.7	3
5	2.51	3	2.9	3
6	2.28	3	2.86	3

VII. LIMITATION

Though system has the option to listen audio for disabled person but none of them are found while taking the survey. So audio based CAPTCHA is not tested in this survey.

VIII. CONCLUSION AND FUTURE WORK

This paper illustrates a new design for CAPTCHA system based on human cognition. This model demonstrates the ability of human to find the answer that other bots and external programs fail to interpret and evaluate. The conducted survey explains the usability of this new form of CAPTCHA and provides valuable feedback to design the overall system and types of question pattern. This framework can easily be extended to specific website to include question of any particular area of interest. In future, more extensive user-study will be performed to suggest context aware questions to give the most user- friendly experience in web surfing and also combat against CAPTCHA farming.

ACKNOWLEDGEMENT

We are expressing our heartfelt thanks to all of the participants specially the students of Daffodil International University.

REFERENCES

[1] Belk, M., Germanakos, P., Fidas, C., Spanoudis, G., & Samaras, G. Studying the Effect of Human Cognition on Text and Image Recognition CAPTCHA Mechanisms. In *Human Aspects of Information Security, Privacy, and Trust* (pp. 71-79). Springer Berlin Heidelberg (2013).

[2] Fidas, C., Voyiatzis, A., Avouris, N.: On the Necessity of User-friendly CAPTCHA. In: 29th ACM Conference on Human Factors in Computing Systems, pp. 2623–2626. ACM Press, New York (2011)

[3] Bursztein, E., Bethard, S., Fabry, C., Mitchell, J.C., Jurafsky, D.: How Good are Humans at Solving CAPTCHAs? A Large Scale Evaluation. In: IEEE International Symposium on Security and Privacy, pp. 399–413. IEEE Press, Washington (2010)

[4] Ahn, L. V., Blum, M., & Langford, J. Telling Humans and Computer Apart Automatically. CACM, V47, No 2.

[5] Ahn, L., Blum, M., Hopper, N., & Langford, J. Using hard ai problems for security. Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques (pp. 294-311). Springer-Verlag.

[6] Bigham, J., &Cavender, A. Evaluating existing audio captchas and an interface optimized for non-visual use. Proceedings of the 27th international conference on Human factors in computing systems (pp. 1829-1838). New York, NY, USA: ACM.

[7] Bursztein, E., Bauxis, R., Paskov, H., Perito, D., Fabry, C., & Mitchell, J. The failure of noise-based non-continuous audio captchas. Proceedings of 2011 IEEE Symposium of Security and Privacy. Oakland.

[8] Chan, T. (2003). Using a text-to-speech synthesizer to generate a reverse turing test. IEEE International Conference on Tools with Artificial Intelligence.

[9] Chew, M., &Tygar, J. (2004). Image recognition CAPTCHAs. Proc. 7th Info. Security Conf., LNCS 3225 (pp. 268 - 279).Heidelberg: Springer-Verlag.

[10] Datta, R., Li, J., & Wang, J. IMAGINATION: A Robust Image-based CAPTCHA Generation System. MM (pp. 331-334). Singapore: ACM.

[11] Elson, J., Doucerur, J., Howell, J., & Saul, J. Asirra: A captcha that exploits interest-aligned manual image categorization. Proceedings of the 14th ACM conference on Commputer and communication security (pp. 366-374). New York, NY, USA: ACM.

[12] Golle, P. (2008). MAchine learning attacks against the Asirracaptcha. Proceedings of the 15th ACM conference on Computer and communication security (pp. 535-542). New York, NY, USA: ACM.

[13] Gossweiler, R., Kamvar, M., &Baluja, S. Whats's up captcha?: a captcha based on image orientation. Proceedings of the 18th international conference on World Wide Web (pp. 841-850). New York, NY, USA: ACM.

[14] Kluever, K., &Zanibbi, R. Balancing usability and security in a video CAPTCHA. Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS) (p. 14). Mountain View, CA USA: ACM.

[15] Mehrnejad, M., Bafghi, A. G., Harati, A., &Toreini, E. SEIMCHA:ANew Semantic Image CAPTCHA Using Geometric Transformations. International Journal of Information Security, 63 - 76.

[16] Moy, G., Jones, N., Harkless, C., & Potter, R. Distortion Estimation Techniques in Solving Visual CAPTCHAs. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[17] Rui, Y., & Liu, Z. ARTiFACIAL: Automated Reverse Turing Test using FACIAL features. Multimedia Systems, 9(6), 493 - 502.

[18] Shirali-Shahreza, M., &Shirali-Shahreza, S. Question-Based CAPTCHA. International Conference on Computational Intelligence and Multimedia Applications, (pp. 54-58).

[19] Vikram, S., Fan, Y., &Gu, G. SEMAGE: A New Image-based Two-Factor CAPTCHA. ACSAC (pp. 237 - 246). Orlando, Florida, USA: ACM.

[20] Winter-Hjelm, C., Kleming, M., &Bakken, R. An interactive 3D CAPTCHA with semantic information. NAIS.

[21] Yan, J., & Ahmed, A.. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms. Proc. of the 23rd Annual Computer Security Applications Conference (ACSAC'07), (pp. 279 - 291).

[22] Yan, J., & Ahmed, A. A Low-cost Attack on a Microsoft CAPTCHA. Proc. CCS (pp. 543 - 554). ACM Press (2008).

[23] Yan, J., & Ahmed, A. (2008). Usability of CAPTCHAs Or usability issues in CAPTCHA design. Proceedings of the 4th SOUPS.

[24] Yan, J., & Yu, S. (2009). Streamlining Attacks on CAPTCHAs with a Computer Game. IJCAI.

[25] Kerckhoffs, A. (1883). La CryptographieMilitaire. Journal des Sciences Militaires 9, 161-191.

[26] Ahn L, Maurer B, McMillen C, Abraham & Blum, reCAPTCHA: Human-Based Character Recognition via Web Security Measures, 10.1126/science.1160379

[27] Tanvee, M. M., Nayeem, M. T., &Rafee, M. M. H. (2011). Move & Select: 2-Layer CAPTCHA Based on Cognitive Psychology for Securing Web Services. International Journal of Video & Image Processing and Network Security IJVIPNS/IJENS, 11(5)

[28] Kani, J., &Nishigaki, M. (2013). Gamified CAPTCHA. In Human Aspects of Information Security, Privacy, and Trust (pp. 39-48). Springer Berlin Heidelberg.

[29] CAPTCHA, 89 - The 10th Zhejiang University Programming Contest – B, http://acm.zju.edu.cn/onlinejudge/showContestProblem.do?problemId=3714 (Accessed on 24th November, 2013)

A Pilot Study Examining the Online Behavior of Web Users with Visual Impairments

Julian Brinkley
Department of Computer Science
East Carolina University
Greenville, North Carolina

Nasseh Tabrizi
Department of Computer Science
East Carolina University
Greenville, North Carolina

***Abstract*—Thisreport presents the results of a pilot study on the online behavioral habits of 46 internet users; 26 of whom self-identified as having a visual impairment (either blind or low vision). While significant research exists which documents the degree of difficulty that users with visual impairments have in interacting with the Web relative to the sighted, few have addressed the degree to which this usability disparity impacts online behavior; information seeking and online exploratory behaviors especially. Fewer still have addressed this usability disparity within the context of distinct website types; i.e. are usability issues more pronounced with certain categories of websites as opposed to others? This pilot study was effective both in exploring these issues and in identifying the accessibility of online social networks as a primary topic of investigation with respect to the formal study that is to follow.**

Keywords—Web Accessibility; Social Networking; Human Computer Interaction

I. Introduction

Given that the modern World Wide Web is a largely visual medium, it would stand to reason that some usability disparity would exist between sighted users of the Web and users with some degree of blindness. This contention is further bolstered by the fact that these visually impaired users, by virtue of the nature of their disability, typically access the Web using a broad spectrum of specialized accessibility technologies with a number of documented deficiencies [1-4]. While the quantification of this disparity varies, several studies have indicated that the Web is roughly three times more difficult to use for individuals with visual impairments than it is for sighted users [4,5]. Few studies however have directly addressed the degree to which this problem effects the online behavior of users with visual impairments. As a result, several subsequent questions regarding the online behavioral patterns of these users are left unanswered, questions which serve as the core of our research effort:

1) Are the types of usability issues faced by individuals with visual impairments largely universal or do they vary depending upon the website type in question?

2) Do individuals with visual impairments avoid new or unfamiliar websites due to fears regarding ease of use and accessibility?

3) What impact do these issues have on these users' participation in highly interactive online social networks?

As a preliminary exercise to conducting a more formal study to investigate the aforementioned behavioral questions a pilot study was conducted using an online questionnaire. This questionnaire was administered to 46 sighted, low vision and blind internet users age 18 and older. Participants were surveyed in an attempt to provide answers to online behavioral and accessibility questions while also providing investigative direction in preparation for a more exhaustive and detailed study.

II. Background

A. Visual Impairment

For the purposes of this research individuals identified as having a visual impairment should be interpreted as those with a visual disability that is not correctable by traditional assistive devices up to and including full blindness [6]. Those individuals identified as "sighted" should be viewed as those having vision in the normal range or correctable to the normal range using traditional assistive devices like glasses or contact lenses.

B. Website Usability versus Accessibility

The terms accessibility and usability are often used inconsistently, interchangeably and in an occasionally overlapping manner in accessibility research. Given that they are not used interchangeably within this report and significant literature exists which outlines their appropriate use, a brief explanation of their use within this report is provided. Accessibility [7-9] is used to describe the ability of users with visual disabilities to functionally interact with a website whereas usability [10] is used to describe the qualitative characteristics of this interaction.

III. Pilot Study

A. Participants

A convenience sample of 46 individuals participated in the pilot study, the composition of which is outlined in Table 1. Of the 46 participants, 25 (18 men and 7 women) were categorized as having some type of visual impairment; the remaining 21 participants (6 men and 15 women) were identified as "sighted".

While research involving the general population typically requires a representative sample of 20-30 participants at a minimum, it is generally acceptable for Human-Computer Interaction (HCI) research involving users with disabilities to have as few as 5-10 participants in some cases[11]. In the case of individuals with significant visual impairments issues with physical mobility, chiefly lack of transportation and scheduling difficulties, have been well documented as barriers to study participation [11]. These individuals often utilize any number of specialized Web browsers or screen reading applications with personalized configurations, therefore the ability to precisely duplicate a specific user's configuration in its totality within a research setting has its difficulties; an issue which exacerbates the aforementioned mobility problem. Distributed research methods, like online questionnaires, have a number of attractive qualities within the context of usability research involving disabled populations. While there may exist some trade off in terms of study control relative to direct observation or interview, the added convenience of distributed research methods may potentially increase participation by eliminating roadblocks to participation. As a result, distributed research methods like diaries or surveys are often ideal tools for HCI research involving these users. These methods allow participation using accessibility configurations that are most familiar to the participants as well as participation at a time and location of the greatest convenience. Recognizing these factors an online questionnaire was chosen for this study.

Participants were recruited through the assistance of organizations for individuals with visual impairments and through posts on the social networking website Facebook (http://www.facebook.com/). In the case of the former, study information was distributed via email to the respective membership of the National Association of Blind Students (NABS), the North Carolina Association of Blind Students (NCABS) and the Massachusetts Association of Blind Students (MABS); with the assistance of each organization's respective leadership. Each e-mail contained a detailed description of the study, consent information and a clickable text hyper-link to the online questionnaire. In the latter case, distribution via online social network, this same explanatory and consent information was made available with a text hyper-link to the online questionnaire using a Facebook page administered by the authors. The study protocol was approved by the East Carolina University, University and Medical Center Institutional Review Board (UMCIRB).

While a large, more representative sample would have been preferable, it was determined that a convenience sample of as few as 10 participants with visual impairments was acceptable in this initial phase of the research effort. 25 individuals with visual impairments ultimately participated in the study however. The added complexity and expense of random sampling with a large number of participants was deemed unnecessary for this preliminary work given that this study was conducted as much for investigative direction and experience conducting HCI research with disabled participants as it was for the data to be collected. In our subsequent formal study sample sizes and selection methods will be modified with the goal of selecting a sample that is likely more representative of the target population.

TABLE I. BREAKDOWN OF STUDY PARTICIPANTS

Participants	**Sighted**		**Visually Impaired**		**Combined**	
	%	*Number*	*%*	*Number*	*%*	*Number*
Sex						
Male	29	6	72	18	52	24
Female	71	15	28	7	48	22
Age						
18 to 25	58	12	48	12	52	24
26 to 35	19	4	8	2	13	6
36 or older	23	5	44	11	35	16

B. *Procedure*

A 33 question, mixed type, online questionnaire was developed which also included user ratings with 5 point Likert-type items (1 = strongly disagree, 2 = somewhat disagree, 3 = neutral, 4 = somewhat agree, 5 = strongly agree) [12]. In addition to collecting demographic information, questions focused specifically on participants' online exploratory behavior, Web usage habits, general opinion of Web usability, opinions regarding website interactivity and use of online social networking applications. Some emphasis was given to social networking within the study given the rise of social networking applications and the increasing incorporation of social networking capabilities to all manner of websites. Five (5) questions within the study addressed social networking specifically.

IV. PRELIMINARY RESULTS

A. *Information Seeking and Exploratory Behavior*

Tables 2 and 3 provide a comparison of the information seeking and online exploratory behavior of both the sighted and visually impaired participant groups. Overall, both groups felt that they could relatively easily find what they were looking for online as expressed within a series of 5 point Likert-type items (1 = strongly disagree to 5 = strongly agree). However this feeling of ease was more pronounced in sighted participants than with the visually impaired group. While both groups felt that most websites are easy to navigate (visually impaired M =3.86, Mode = 4; sighted M = 4.14, Mode = 4), the visually impaired group expressed a strong desire for a Web that is easier to use (M = 3.59, SD =1.43, Mode = 4) whereas the sighted group expressed general satisfaction with current levels of Web usability (M = 2.38, SD =1.20, Mode = 1,2).

The visually impaired group visited a greater number of websites on a daily basis than sighted participants with 44% of the visually impaired group indicating that they visited "11 or more" websites on a daily basis compared to 5% of the sighted group.

Users with visual impairments also visited more new websites on a daily basis than the sighted group with 80% of participants with visual impairments visiting between 1 and 10 new websites daily compared to 67% of the sighted group. The visually impaired group however expressed some difficulties navigating these new websites (M = 3.04, SD = 1.33, Mode = 4) whereas the sighted group expressed minimal difficulty in this regard (M = 2.43, SD = 1.21, Mode = 1,3).

Both groups expressed minimal website avoidance behavior with fewer than 5% of both groups indicating an avoidance of new websites. While the visually impaired group indicated significantly more comfort with familiar websites (M = 4.22, SD = 1.11, Mode = 5) relative to the sighted participants group (M = 3.57, SD = 0.87, Mode = 3,4), visually impaired participants (M = 2.00, SD = 1.38, Mode = 1) expressed a comparable and minimal avoidance of new websites due to apprehension regarding the ability to find information of interest. Sighted participants (M = 2.19, SD = 1.25, Mode = 1) indicated a slightly greater avoidance of new websites due to these information seeking concerns though both groups indicated that this apprehension did not result in an outright avoidance of new websites. Both groups indicated a significant reliance on search engines for new content with 52% of sighted participants and 66% of visually impaired participants identifying this as a primary source of new stories, videos and images. However social networking websites were a significantly more popular source of this information for sighted individuals (33%) than for the visually impaired group (13%).

B. Participants' Website Usage Habits

Table 4 provides a comparison of website usage habits by website type. Of the nine website types provides, both visually impaired (33%) and sighted (52%) participants indicated that social networking websites were the most visited. This number was significantly higher for sighted participants than for the participants with visual impairments however. Sighted participants expressed the most difficulty using Web-logs or "Blogs" (13%) whereas social networking websites (21%) were identified as the most difficult to use by those with visual impairments. The most popular response however for both groups when referencing online difficulties was that no difficulties existed relative to most websites (sighted = 67%, visually impaired = 42%).

C. Use of Online Social Networks

Table 5 provides a comparison of the use of social networking websites between the visually impaired and sighted participants. While some commonalities were exposed between the two groups overall, the frequency with which online social networks were used, as well as the perceptions regarding the usability of these websites, differed substantially.

A majority of participants in both groups indicated that they held an online social networking account though the social network participation rate for sighted participants (95%) was significantly higher than for the visually impaired group (67%). In an open ended question, both groups indicated a preference for Facebook.com with 81% of sighted participants and 67% of individuals with visual impairments indicating that it was their most used social networking website. 77% of these sighted participants indicated a "moderate" to "extreme" frequency of social network usage compared to 42% of visually impaired participants.

Perceptions regarding the usability of social networking websites differed substantially between the two groups with sighted participants indicating both a greater enjoyment of these websites as well as significantly greater ease of use. 36% of individuals with visual impairments indicated that they would like to use social networking websites but found them too difficult to use compared to only 10% of sighted participants.

72% of sighted participants indicated some degree of "enjoyment" in their use of online social networks as compared to 41% of participants with visual impairments. An additional 45% of the visually impaired users group indicated some degree of dislike of social networking websites compared to 15% of the sighted group.

TABLE II. CODING OF RESPONSES REGARDING INFORMATION SEEKING AND ONLINE EXPLORATORY BEHAVIOR

	Sighted	**Visually Impaired**
Statement	**%**	**%**
The approximate number of websites I visit on a daily basis		
0	5	0
1 to 5	43	28
6 to 10	47	28
11 or more	**5**	**44**
The number of NEW websites I visit on a daily basis		
0	33	20
1 to 5	62	68
6 to 10	5	12
11 or more	0	0
I avoid visiting new websites		
True	5	4
False	95	96
I usually find new online content like stories, videos and images from		
A link from another website	10	21
Search engine like Google, Yahoo and Bing	52	66
Social networking websites	**33**	**13**
Word of mouth	5	0
I don't know where I find new website content	0	0

TABLE III. Mean Scores Regarding Subjects' Online Exploratory Behavior, Information Seeking And Web Usage

Statement	**Mean**	**SD**	**N**	***t-test***	**df**	**p**
I frequently visit new websites						
Visually Impaired	3.76	1.20	25	1.37	44	.176
Sighted	3.29	1.10	21			
I have trouble finding my way around new websites						
Visually Impaired	3.04	1.33	24	1.60	43	.116
Sighted	2.43	1.21	21			
I avoid trying new websites because I am concerned about being able to find the information that I am looking for						
Visually Impaired	2.00	1.38	22	0.47	41	.639
Sighted	2.19	1.25	21			
I OFTEN have trouble finding what I am looking for online						
Visually Impaired	2.32	1.39	22	1.32	41	.193
Sighted	1.81	1.12	21			
I SOMETIMES have trouble finding what I am looking for online						
Visually Impaired	3.41	1.37	22	1.27	41	.211
Sighted	2.90	1.26	21			
Generally it is easy for me to find what I am looking for online						
Visually Impaired	4.41	0.80	22	0.78	41	.437
Sighted	4.19	1.03	21			
I am more comfortable visiting websites that I am familiar with						
Visually Impaired	4.22	1.11	22	**2.13**	**41**	**.039**
Sighted	3.57	0.87	21			
I wish the internet was easier to use						
Visually impaired	3.59	1.44	22	**2.98**	**41**	**.004**
Sighted	2.38	1.20	21			
Most websites are easy to navigate						
Visually impaired	3.86	1.04	22	1.05	41	.298
Sighted	4.14	0.65	21			
I enjoy using social networking websites						
Visually impaired	2.86	1.70	22	**2.25**	**41**	**.030**
Sighted	3.86	1.15	21			
I would like to use social networking websites but I find them too difficult to use						
Visually impaired	2.55	1.50	22	**2.02**	**41**	**.049**
Sighted	1.76	1.00	21			

TABLE IV. RESPONSES RELATED TO WEBSITE USAGE

	Sighted	**Visually Impaired**
Statement	**%**	**%**
I normally visit this type of website the most often		
Blogs	0	4
Lifestyle Websites	0	0
Medical Websites	0	0
News Websites	24	17
Online Encyclopedias	0	4
Shopping Websites	19	25
Social Networking Websites	**52**	**33**
Sports Websites	5	0
Other or I don't know	0	17
I normally have the most difficulty using this type of website		
Blogs	**13**	**8**
Lifestyle Websites	0	4
Medical Websites	0	0
News Websites	0	13
Online Encyclopedias	0	0
Shopping Websites	10	4
Social Networking Websites	**5**	**21**
Sports Websites	5	8
I don't have difficulty using most websites	67	42

TABLE V. RESPONSES RELATED TO SOCIAL NETWORKING WEBSITES

	Sighted	**Visually Impaired**
Statement	**%**	**%**
Do you have an account on a social networking website?		
Yes	95	67
No	**5**	**33**
Which social networking website do you visit most often?		
Facebook	81	67
Instagram	0	4
LinkedIn	0	8
Pinterest	5	0
Twitter	**0**	**25**
Other	19	0
None	**5**	**29**

V. CONCLUSION AND FUTURE WORK

While the results of this pilot study indicate that Web accessibility generally may be improving, *equivalent usability* is still elusive. Findings at this stage of the research effort suggest that differences do exist between the online behavior of sighted users and users with visual impairments. These differences suggest that the presence of a visual impairment may have a significant impact on information seeking and online exploratory behavior. Most notably, the results of this study indicate that additional research is needed to explore the usability difficulties that users with visual impairments encounter on social networking websites specifically. Visually impaired participants indicated significant difficulties using websites of this type and were most severely challenged by social networking websites within the context of the options provided; findings mirrored by other studies [13]. This social networking usability issue is especially problematic as the use of these services has been shown to improve educational outcomes, promote self-esteem, enhance social inclusion and increase interpersonal relationships among other benefits [14-16].

While sighted users may capitalize on the aforementioned benefits with relative ease, the study suggests that individuals with visual impairments face substantial challenges in this regard. This is troubling in that studies show that users with visual impairments may benefit disproportionately from the benefits of these social networks [17]. Fears regarding diminished self-value, social stigmatization or negative self-image related to the disability may be mitigated by the type of electronic social interaction that these online social networks facilitate [18,19]. Few studies however have addressed the participation, or lack thereof, of individuals with visual impairments in these online social networks and where this issue has been examined accessibility has been determined to be relatively poor [13, 20].

Moving forward a formal study of social network accessibility will be conducted applying the lessons learned from the initial investigative effort documented within this report. This study will utilize a combination of interview and observation techniques with the goal of collecting data with a greater level of detail. Information will also be collected regarding software (screen reader, Web browser, operating system) and device type; factors intentionally omitted from this initial investigation. This is significant given the increasing use of mobile screen reading technology by users with visual impairments [13] as mobile browsing increases in popularity [21].

REFERENCES

[1] A. Fernandes, F. Martins, H. Paredes and I. Pereira, "A Different Approach to Real Web Accessibility" in Proc. International Conference on Universal Access in Human-Computer Interaction, pp. 723-727, 2000.

[2] C. S. Fichten, J. V. Asuncion, M. Barile, V. Ferraro and J. Wolforth, "Accessibility of e-learning and computer and information technologies for students with visual impairments in postsecondary education," Journal of Visual Impairment and Blindness, vol. 103, pp. 543-557, Sep. 2009.

[3] K. Pernice and J. Nielsen, Beyond ALT Text: Making the Web Easy to User for Users with Disabilities. Fremont, CA: Nielsen Norman Group, 2001.

[4] H. Petrie, F. Hamilton and N. King, "Tension, what tension?: Website accessibility and visual design," In Proc. from the International Cross-Disciplinary workshop on Web Accessibility, pp. 13-18, 2004.

[5] S. Harper, C. Goble and R. Stevens, "A Pilot Study to Examine the Mobility Problems of Visually Impaired Users Traveling the Web," ACM SIGCAPH Computers and the Physically Handicapped, vol. 68, pp. 10-19, Sep. 2000.

[6] "Visual Impairment and Blindness – Fact Sheet," http://www.who.int/mediacentre/factsheets/fs282/en/, 2013.

[7] G. Brajnik, "Automatic web usability evaluation: what needs to be done?," In Proc. 6th Conference on Human Factors and the Web, 2000.

[8] S. Krug, Don't Make Me Think: A Common Sense Approach to Web Usability. 2nd ed. Berkeley, CA: New Riders Press, 2006.

[9] World Wide Web Consortium W3C, "Web Accessibility Initiative (WAI)," http://www.w3.org/WAI/intro/accessibility.php, 2010.

[10] International Organisation for Standardisation, ISO9241 Ergonomic, Part 11: Guidance on usability. Geneva, Switzerland, 1998.

[11] J. Lazar, J. H. Feng and H. Hochheiser, Research Methods in Human Computer Interaction. West Sussex, UK: Wiley Publishing,, 2010.

[12] K. L. Wuensch, "What is a Likert Scale and How Do You Pronounce Likert?," Retrieved Feb 12, 2013 from http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm, 1998.

[13] "Web Accessibility in Mind. Screen Reader User Survey 4," http://webaim.org/projects/screenreadersurvey4/#mobil, 2013.

[14] P. Collin, K. Rahilly, I. Richardson and A. Third. The Benefits of Social Networking Services. Abbotsford, Australia : Cooperative Research Centre for Young People, Technology and Well Being, 2010.

[15] K. N. Hampton, L. S. Goulet, L. Rainie and K. Purcell, Social networking sites and our lives. Washington, DC : Pew Research Center, 2011.

[16] D. M. Boyd, Why Youth (Heart) Social Network Sites: The Role of Networking Publics in Teenage Social Life. Cambridge, MA : MIT Press, 2007.

[17] S. M. Kelly and T. J. Smith, "The digital social interactions of students with visual impairments: findings from two national surveys," Journal of Visual Impairment & Blindness, vol. 102, pp. 528-539, Sep 208.

[18] S. Kef, J. J. Hox and H. T. Habekothe, "Social networks of visually impaired and blind adolescents. Structure and effect on well-being," Social Networks, vol. 22, pp. 73-91, May 2000.

[19] F. Fovet, "Impact of the use of Facebook Amongst Students of High School Age with Social, Emotional and Behavioral Difficulties (SEBD)," In Proc. from the 39th IEEE Frontiers in Education Conference, pp. 1-6, 2009.

[20] Accessibility of Social Networking Services. Madrid, Spain: Observatory on ICT Accessibility – Discapnet, 2010.

[21] "Statcounter global stats: Mobile vs. desktop from Jan. 2009 to Apr. 2013," http://gs.statcounter.com/#mobile_vs_desktop-ww-monthly-200901-201304, 2013.

Inverted Indexing In Big Data Using Hadoop Multiple Node Cluster

Kaushik Velusamy
Dept. of CSE Amrita University
Coimbatore, India

Nivetha Vijayaraju
Dept. of CSE Amrita University
Coimbatore, India

Deepthi Venkitaramanan
Dept. of CSE Amrita University
Coimbatore, India

Greeshma Suresh
Dept. of CSE Amrita University
Coimbatore, India

Divya Madhu
Dept. of IT Amrita University
Coimbatore, India

***Abstract*—Inverted Indexing is an efficient, standard data structure, most suited for search operation over an exhaustive set of data. The huge set of data is mostly unstructured and does not fit into traditional database categories. Large scale processing of such data needs a distributed framework such as Hadoop where computational resources could easily be shared and accessed. An implementation of a search engine in Hadoop over millions of Wikipedia documents using an inverted index data structure would be carried out for making search operation more accomplished. Inverted index data structure is used for mapping a word in a file or set of files to their corresponding locations. A hash table is used in this data structure which stores each word as index and their corresponding locations as its values thereby providing easy lookup and retrieval of data making it suitable for search operations.**

Keywords—Hadoop; Big data; inverted indexing; data structure

I. Introduction

Wikipedia is an online encyclopaedia which contains over four million articles. In general, searching over such text based documents involves document parsing, index, tokenisation, language recognition, format analysis, section recognition. Hence a search engine for such large data which is done in a single node with a single forward index built over all the documents will consume more time for searching. Moreover the memory and processing requirements for this operation may not be sufficient if it is carried out over a single node. Hence, load balancing by distribution of documents over multiple data becomes necessary.

Google processes 20PB of data every day using a programming model called MapReduce. Hadoop, a distributed framework that processes big data is an implementation of MapReduce. Hence it is more apt for this operation as processing is carried out over millions of text documents.

Inverted index is used in almost all web and text retrieval engines today to execute a text query. On a user query, these search engines uses this inverted index to return the documents matching the user query by giving the hyperlink of the corresponding documents. As these indices contain the vocabulary of words in dictionary order only a small amount of documents containing the indices need to be processed.

Here, the design of a search engine for Wikipedia data set using compressed inverted index data structure over Hadoop distributed framework is proposed. This data set containing more than four million files needs an efficient search engine for quick access of data. No compromise must be made on the search results as well as the response time. Care should be taken not to omit documents that contain words synonymous user query. Since accuracy and speed is of primary importance in search, our methods could be favoured in such cases.

II. Literature Survey

[2] For large-scale data-intensive applications like data mining and web indexing MapReduce has become an important distributed processing model. Hadoop–an open-source implementation of MapReduce is widely used for short jobs requiring low response time. Both the homogeneity and data locality assumptions are not satisfied in virtualized data centres. This paper [2] shows that ignoring the data locality issue in heterogeneous environments can noticeably reduce the MapReduce performance.

The authors also address the problem of how to place data across nodes in a way that each node has a balanced data processing load. Given a data intensive application running on a Hadoop MapReduce cluster, their data placement scheme adaptively balances the amount of data stored in each node to achieve improved data-processing performance. Experimental results on two real data-intensive applications show that their data placement strategy can always improve the MapReduce performance by rebalancing data across nodes before performing a data-intensive application in a heterogeneous Hadoop cluster. The new mechanism distributes fragments of an input file to heterogeneous nodes based on their computing capacities. Their approach improves performance of Hadoop heterogeneous clusters.

According to [1], a virtualized setup of a Hadoop cluster that provides greater computing capacity with lesser resources is presented, as virtualized cluster requires only fewer physical machines with master node of the cluster set up on a physical machine, and slave nodes on virtual machines (VMs).

The Hadoop virtualized clusters are configured to use capacity scheduler instead of the default FIFO scheduler. The capacity scheduler schedules tasks based on the availability of RAM and virtual memory (VMEM) in slave nodes before allocating any job. Instead of queuing up the jobs, the tasks are efficiently allocated on the VMs based on the memory available. Various configuration parameters of Hadoop are analysed and the virtualized cluster is fine-tuned to for best performance and maximum scalability. The results show that the approach gives a significant reduction in execution times, which in turn shows that the use of virtualization helps in better utilization of the resources of the physical machines used. Given the relatively under power of the machines used in the real cluster the results were fairly relevant. The addition of more machines in the cluster leads to an even greater reduction in runtime.

According to [8], Hadoop, the emerging technology made it feasible to combine it with virtualisation to process immense data set. A method to deploy cloud stack with Map Reduce and Hadoop in virtualised environment was presented in this paper. A brief idea on setting up a Hadoop experimental environment to capture the current status and the trends of optimising Hadoop in virtualised environment was mentioned. The advantages and the disadvantages of the virtualised environment was discussed, ending with the benefits of one's compromise over the other. Making use of the virtualised environment in Hadoop fully utilizes the computing resources, make it more reliable and save power and so on. On the other side, we have to face the lower performance of virtual machine. Then some methods to optimize Hadoop in virtual machines were discussed.

III. Problem Statement

The result of any user's search query must be fast, should not miss any relevant data related to the query. A search engine designed by using distributed framework like Hadoop and inverted index data structure is fast and returns all the relevant results. In order to do this and to analyse the feasibility of deployment of a search engine for Wikipedia various requirements and parameters to be considered must be well understood and analysed.

IV. Parameters For Performance Metrics

The performance of a search operation through an inverted index built over millions of Wikipedia documents distributed over a multiple node Hadoop cluster in a virtual node could be effectively measured using various parameters such as response time ,throughput, speed up, latency hiding, computation time, communication time and thereby computation and communication ratio. In terms of the search operation in this distributed and parallel platform, response time indicates the time taken for the first of the relevant wiki documents to appear when a query is made. Through put in other words can be defined as the number of transactions per second or the maximum number of search queries that can be made per second, speed up factor refers to the time that could be saved due to a fraction of process that could be parallelized .As the documents are distributed across multiple documents, the percentage of search operation that can be parallelized and thereby the speedup achieved could be measured. [6]

$$\text{Speed-up factor}=\frac{\text{Ts}}{\text{Tp}}$$

Where Ts -Time taken for serial execution of the process and Tp - time taken after parallelization. As more time is consumed in start-up of a communication between nodes, making use of this time effectively for completing as much computations as possible would improve performance. This can be achieved via non-blocking send routines thereby helping in achieving latency hiding. Sometimes, even blocking send routines allow computations to take place until the expected messages reach the destination aiding in improving latency hiding. Total processing time includes computations and the communications carried out.

$$T_{\text{process}} = T_{\text{computation}} + T_{\text{communication}}$$

The computation time for the search operation can be calculated by counting the number of computations per process. Computation involves locating the node that has the relevant documents. [9]Communication time depends on the size of the data transferred, start-up time for each message and number of messages in a process and the mode of data transfer. Communication in multiple cluster node involves requesting a node for certain documents based on the query and the nodes responding with the requested documents.

$T_{\text{communication}} = T_{\text{start up}} + w{*}T_{\text{data}}$

Where

$T_{\text{start-up}}$ – Time needed to send a blank message
T_{data} - Time to send/receive a single data word
W - No. of data words

$$\text{Speed-up factor} =\frac{\text{Ts}}{\text{Tp}}=\frac{\text{Ts}}{\text{T computation} + \text{T communication}}$$

The computation communication ratio throws light on the how much time communication takes as a result of increased amount of data.

V. Inverted Indexing

Indexing refers to creating a link or a reference to a set of records so as to enable better identification or location. Forward indexing and inverted indexing are two main types of indexing. When an element say 97 is accessed through its index say Arr [3] in Fig 1, then it is forward indexing. When the same element is searched based on its occurrence or the number of occurrences, then it is inverted indexing.

23	45	64	97	53	72	93
Arr[0]	Arr[1]	Arr[2]	Arr[3]	Arr[4]	Arr[5]	Arr[6]

Fig. 1. Illustration of Forward and Inverted Indexing

An inverted index for a document or set of documents contains a hash table with each word as its index and a posting list as value of each index. A postings list consists of a document id, position of word in that document and frequency of occurrence of each word in that document.

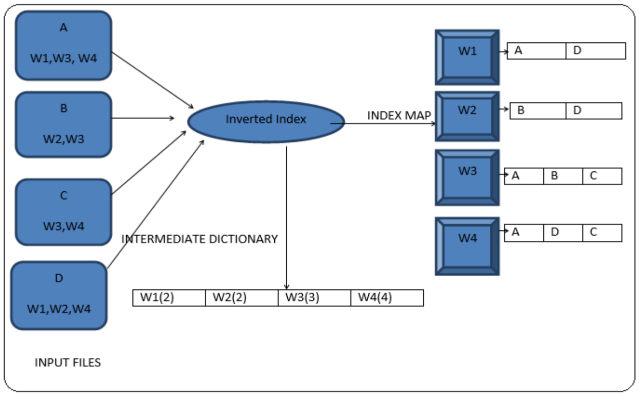


Fig. 2. Inverted Indexing-Working

If there are n documents to be indexed then a unique document id is set for each document from 0 to n-1. The postings list for a term is sorted based on various criteria. Though it is easy to sort it based on document id, for search operations other parameters are considered for sorting. Sorting done based on frequency of a term in a document is more apt for a search operation. At the end of sort processing this data structure returns the top k documents in the postings list where k is the maximum returning capacity of a search engine in a single stretch. [7].

A. Algorithm

```
Inverted_Index (int docID[n], string doc[n])
M ← new HashMap
Count ← 0
For all document with docID m from 0 to n-1
        For all term tm and position pos in doc
         With   docID m do
                M {tm, previous pos, previous m} ← M
                {tm, pos, m} +1
                Count (tm, m) ++

        For each tm in M with docID m
                Sort (count (tm, m))
```

As explained in Fig 2, input to the indexing algorithm is the set of document IDs and the contents of all the documents. Each new term in the document is formed as an index in the hash table. For each occurrence in a document its document ID is added to the postings list of that term along with its position. After each occurrence of a term in a document its corresponding frequency variable count is incremented. Postings list of each term is finally sorted based on the frequency of words in each document.

```
Algorithm Search (HashMap M, string word)
       return M[word]
```

In the search part of an inverted index, the word which is queried by the user is passed as input along with the hash map which has the set of all positions of the each word in the document. Hash map takes the word as its index and returns the value stored in that index.

VI. Data Set - Wiki dumps

All the contents of Wikipedia are available in downloadable format as wiki dumps. This can be taken by users for archival/backup purposes, offline storage, educational purpose, for republishing, etc. There are over four million files in Wikipedia, compressed form as wiki dumps of size 9.5 Giga bytes approximately. When extracted from the compressed form, it comes to around 44 Giga bytes. Database backup dumps have a complete copy of all Wikipedia documents as wikitext and the set of all its metadata in XML. Static HTML dumps has copies of all pages of Wikipedia wikis in HTML form.

Contents of dumps include page-to-page link, media metadata, title, information about each page, log data, Misc bits, etc. These are in the wrapper format described at schema Export Format which is compressed in bzip2 and .7z format. They are provided as dumps consisting of entire tables using mysqldump. Internal file system limit must be taken into account before extracting these files from compressed format.

VII. Hadoop

Map Reduce method has emerged as a scalable model that is capable of processing pet a bytes of data. Fundamental concept of MapReduce: Rather than working on one, huge block of data with a single machine, Big Data is broken up into files that further are broken into blocks by Hadoop and parallel processing and analysis is carried out. [5]

The Hadoop is a map reduce framework that provides HDFS (Hadoop Distributed File Systems) infrastructure. HDFS was designed to operate and scale on commodity hardware. Breakdown in hardware is largely compensated by replication of blocks of data in multiple nodes.

A. Hadoop Distributed Filesystem (Hdfs) Overview

HDFS (Hadoop Distributed File System) is a distributed user level file system which stores, processes, retrieves and manages data in a Hadoop cluster. HDFS infrastructure that Hadoop provides, include a dedicated master node called Name Node which contains a job tracker, stores meta-data, controls the overall distributed process execution by checking out whether all name nodes are functioning properly through periodic heart beats. It also contains many other nodes called Data Node which contains a task tracker, stores applications data. The Ethernet network connects all nodes. HDFS is implemented in Java and it is platform independent. Files in HDFS are split into blocks and each block is stored as an independent file in the local file system of Data Nodes. Each block of a HDFS file is replicated at least three times in multiple Data Nodes. Through replication of application data, provides data durability.[9]

The Name Node manages the namespace and physical location of each file. File system operations are done by HDFS client by contacting the Name node. Name Node checks a client's access permission and gets the list of Data Nodes hosting replicas of blocks. Then, requirements are sent to the "closest" Data Node by requesting a particular block. Then, a socket connection is created between the client and the Data Node. The data is transferred to the client application. When a client application writes a HDFS file, it first splits the file into HDFS blocks and the Name Node gets the list of Data Nodes which are replicas of each block and writing data is done by multithreading. [3]

If the client application is running on a Data Node, the first replica of the file is written into the local file system of the running Data Node. If the client application isn't running on a Data Node, a socket connection is created between the client and the first Data Node. The client splits the block into smaller packets and starts a pipeline: the client sends a packet to the first Data Node; the first Data Node on getting this packet, writes this to the local file system, and also sends it to the next Data Node. A Data Node can receive the data from a previous node and at the same time forward the data to the next node. When all nodes in this pipeline write the block into local file system successfully, the block write is finished and then Data Nodes update the block physical information to the Name Node. The architecture of multiple cluster implementations has been explained in Fig 3.

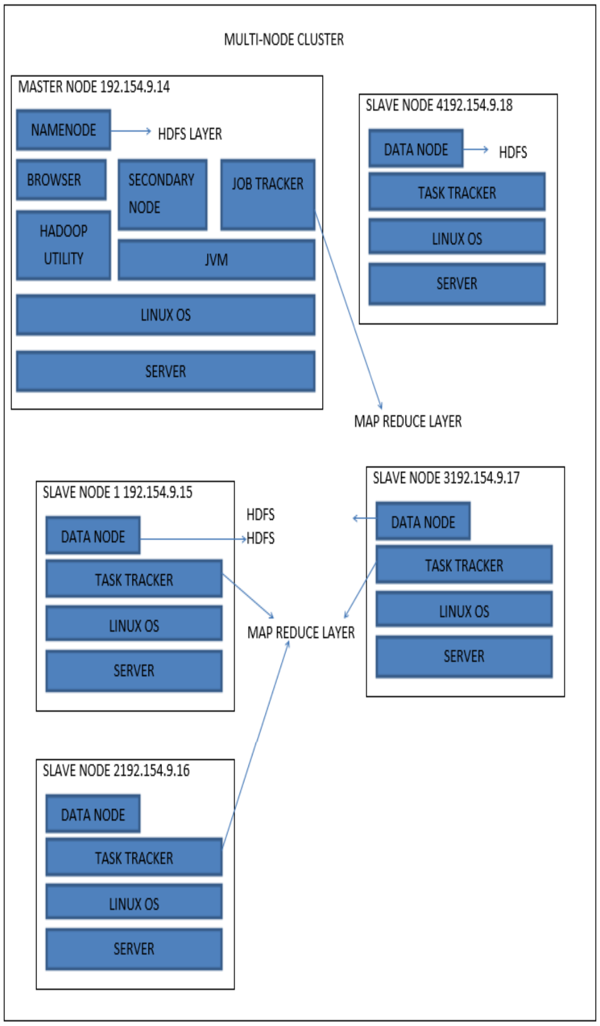


Fig. 3. Hadoop Multiple node Cluster Architecture

B. *Working process of Hadoop Architecture*

Hadoop is designed to run on a large number of machines that don't share any memory or disks. When a data is loaded into Hadoop, the software splits that data into pieces and spreads it across different servers. Hadoop keeps track of where the data resides. And because there are replica of single data, data stored on a server that goes offline or dies can be automatically replicated from a known good copy.

In a Hadoop cluster, every one of those servers has two or four or eight CPUs. Each server operates on its own little piece of the data. Results are then delivered back through reduce operations. MapReduce maps the operation out to all of those servers and then reduces the results back into a single result set. Since Hadoop spreads out data, it is possible to deal with lots of data. Since all the processors work in parallel and harness together, complicated computational questions can be performed. Node failures are automatically handled by the framework for both map and reduce functions.

VIII. Assumptions And Goals

Applications that run on HDFS have large data sets. A typical file in HDFS is Gigabytes to Terabytes in size. Therefore, HDFS must provide high bandwidth and scalability to hundreds of nodes. HDFS applications need a write-once-read-many access model for files. If a file is created and written, it is assumed that it will not be changed in future. This is to simplify data coherency and to get high throughput data access.

IX. Proposed Solution

A Hadoop cluster is established by passing Wikipedia files as input data and inverted indexing is done by taking advantage of Map Reduce.

In the map phase, the Wikipedia files are divided equally among mappers and passed as inputs. Each Wikipedia file is given a unique document ID. Each mapper indexes each term in its file into the hash map with the corresponding document ID and position in that document as a posting list. When it finds that term for the first time it creates that term as the index and writes the corresponding postings list of that term. When the term is found again, the corresponding posting list for that position is appended with the previous list to index holding that term.

A. *Map function pseudo-code*

```
Algorithm Map (int docID[x], string doc[x])
        M ← new HashMap
        Count ←0
        For all document of docID m from 0 to x-1
            For all term tm and position pos in doc
            with docID m do
             M {tm, previous pos, previous m}
        ← M{ tm, pos, m }+1
         Count (tm, m) ++
        emit (M, count (tm, m))
```

In the above algorithm X is the maximum number of documents processed within a mapper. The input file is read word by word and indexed accordingly with its document ID and corresponding position in a hash map. The variable count keeps track of the frequency of a term within each document in that mapper. At the end each mapper returns its hash map with the count value of each term in a document.

In reduce phase each reducer takes in its responsibility a term or set of terms. These terms are given an index position in a global hash map where all the terms are stored as index. When a reducer encounters its term from a mapper it appends the posting list of that mapper to its value in this hash map. After appending the entire list of that term from all the mappers, reducer sorts posting list based on count value of each document. The more the value, the preference is higher. In the same way, all the terms in this whole document are indexed in the hash map in this reduce phase.

B. Reduce function pseudo-code

```
Algorithm  Reduce (term tm, List of hash maps of each mapper[], count{tm, docID})
  G ← new HashMap //G is common HashMap for all reducers
     for each hash map H from all mappers
          for each term tm in document with docID m and position pos in H
          //n is the total number of documents
          G{ tm , previous pos, previous m} ←
                  H{ tm, pos , m }+1
          Sort( count (tm , m ))
          //values in list is sorted based on the count value of each term in a document
       emit(G)
```

In the above pseudo code each reducer takes as its input all the hash maps of various mappers and the count values of each term in a document. Reducer checks each hash map with its allotted term and if it matches with any mapper's index it appends that value in global hash map. When all the values are appended for a term it is finally sorted based on its count value in each document.

C. Retrieval

The terms in global hash map is divided among the mappers along with their corresponding posting list. When the user queries a term, the name node sends this query to the corresponding data node. Value of the term is passed to the reducer as a complete list. Reducer returns the first k values of that term to the user where k is the maximum number of pages returned for a user query.

X. Future Works

First a distributed, multiple node Hadoop cluster has been built and the millions of wiki documents has been distributed over these nodes. A compressed inverted index containing indices for words in dictionary order is to be built over these documents. After building inverted index, distributed performance evaluation for searching documents based on keyword is intended to be made. Further data analysis and text mining could be made based on index support. The results of text mining and data analysis would help in suggesting related pages based on data such as other documents where the synonyms of the query are predominantly found. Indexing can be further partitioned in to local index partitioning and global index partitioning. In term based partitioning or global index partitioning, each node in the multiple cluster stores posting list only for a subset of the term in the collection. Local index partitioning is each server building a separate index for the files that it contains. When this is done, each server indexes only the document that it contains, reducing the number of documents to thousands. This is very much lesser compared to the actual number of indices that had to be built if indexing is to be done for over a million documents. So, instead of building a single index over 4 million Wikipedia documents, local index would be built over documents on each node and an index would be built on these indices thereby quickening search and compressing indices. Further, indices built over articles (a, the, an) and other such common words would be deleted for adding accuracy.

XI. Conclusion

In this paper, a compressed inverted index data structure that could help in crawling for words in dictionary order such that all the indices built for millions of documents need not be processed has been proposed. In addition, basic factors for designing indices such as merge factors, storage technique, index size, look up speed, maintenance, fault tolerance etc. will also be taken into account. Building a local index for files within those system will prove to be a great way to solve problems that could arise in parallelism such as when a file is added, whether index updating should happen before the search operation that is currently going on and vice versa as only a portion of the entire set of documents need to be updated making the 'index merging' process very simple. In addition to storing a token word, its document id and the position in which it appears, we have suggested to add token word document id and its frequency to rank up the relevant documents. Our work has motivated several interesting open questions such as which type of inverted index data structure would be most useful for text mining. Other ways to optimise performance in search is being investigated and added over to the suggested methods.

References

[1] Raj, A. Kaur, K. ; Dutta, U. ; Sandeep, V.V. ; Rao, S. "Enhancement of Hadoop Clusters with Virtualization Using the Capacity Scheduler", Third International Conference on Services in Emerging Markets (ICSEM),Mysore, India, Dec 2012. Page(s): 50 - 57. Print ISBN: 978-1-4673-5729-6. INSPEC Accession Number: 13343537. D.O.I: 10.1109/ICSEM.2012.15. Link: http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=6468179

[2] Jiong Xie; Shu Yin ; Xiaojun Ruan ; Zhiyang Ding ; Yun Tian ; Majors, J. ; Manzanares, A. ; Xiao Qin. "Improving MapReduce performance through data placement in heterogeneous Hadoop clusters". IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), Atlanta, GA, April, 2010. Page(s): 1 - 9. Print ISBN: 978-1-4244-6533-0. INSPEC Accession Number: 11309800. D.O.I : 10.1109/IPDPSW.2010.5470880. Link: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5470880

[3] Kala Karun, A ; Chitharanjan, K ; "A review on hadoop — HDFS infrastructure extensions ", IEEE Conference on Information & Communication Technologies (ICT), JeJu Island, April 2013. Page(s): 132 - 137. Print ISBN: 978-1-4673-5759-3. INSPEC Accession Number: 13653440. D.O.I: 10.1109/CICT.2013.6558077. Link: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6558077

[4] Richard Mccreadie ; Craig Macdonald ; Iadh Ounis; "MapReduce indexing strategies: Studying scalability and efficiency". International Journal of Information Processing and Management. Volume 48 Issue 5, September, 2012. Pages: 873-888. Publisher Pergamon Press, Inc. Tarrytown, NY, USA. ISSN: 0306-4573 doi>10.1016/j.ipm.2010.12.003. Link: http://dl.acm.org/citation.cfm?id=2337723

[5] Apache Hadoop, Hadoop, HDFS, Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Pig, Zookeeper are trademarks of the Apache Software Foundation. http://www.hadoop.apache.org/ Last Published: 10/16/2013 06:37:41. Copyright © 2012. The Apache Software Foundation. 2nd October 2013.

[6] Barry Wilkinson; Michael Allen; "Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers" (2nd Edition). Publication Date: March 14, 2004, ISBN-10: 0131405632, ISBN-13: 978-0131405639 , Edition: 2. Link : http://www.amazon.com/Parallel-Programming-Techniques-Applications-Workstations/dp/0131405632

[7] Gal Lavee ; Ronny Lempel ; Edo Liberty ; Oren Somekh ; " Inverted index compression via online document routing" Published in: WWW '11 Proceedings of the 20th international conference on World Wide Web. Pages 487-496. ISBN: 978-1-4503-0632-4 doi:10.1145/1963405.1963475. Publisher ACM New York, NY, USA ©2011. Link: http://dl.acm.org/citation.cfm?id=1963475

[8] Guanghui Xu; Feng Xu; Hongxu Ma; "Deploying and researching Hadoop in virtual machines". Published in: IEEE International Conference on Automation and Logistics (ICAL), Zhengzhou, Aug. 2012. Page(s): 395 - 399. ISSN: 2161-8151. E-ISBN: 978-1-4673-0363-7. Print ISBN: 978-1-4673-0362-0. INSPEC Accession Number: 13000378. D.O.I:10.1109/ICAL.2012.6308241. Link: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6308241

[9] Shvachko, K.; Hairong Kuang ; Radia, S. ; Chansler, R. " The Hadoop Distributed File System". Published in: IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, May 2010. Page(s): 1 - 10. E-ISBN: 978-1-4244-7153-9. Print ISBN: 978-1-4244-7152-2. INSPEC Accession Number: 11536653. D.O.I: 10.1109/MSST.2010.5496972. Link: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5496972

[10] Ishii, M.; Jungkyu Han; Makino, H; "Design and performance evaluation for Hadoop clusters on virtualized environment" Published in: International Conference on Information Networking (ICOIN), Bangkok, Jan. 2013. Page(s): 244 - 249. ISSN: 1976-7684. E-ISBN: 978-1-4673-5741-8. Print ISBN: 978-1-4673-5740-1. INSPEC Accession Number: 13431469. Digital Object Identifier: 10.1109/ICOIN.2013.6496384. Link: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6496384

Mobile Robots in Teaching Programming for IT Engineers and its Effects

Attila Pásztor, Róbert Pap-Szigeti, Erika Török
Kecskemét College, Faculty of Mechanical Engineering and Automation
Kecskemét, Hungary

***Abstract*—in this paper the new methods and devices introduced into the learning process of programming for IT engineers at our college is described. Based on our previous research results we supposed that project methods and some new devices can reduce programming problems during the first term. These problems are rooted in the difficulties of abstract thinking and they can cause the decrease of programming self-concept and other learning motives.**

We redesigned the traditional learning environment. As a constructive approach project method was used. Our students worked in groups of two or three; small problems were solved after every lesson. In the problem solving process students use programmable robots (e.g. Surveyor, LEGO NXT and RCX). They had to plan their program, solve some technical problems and test their solution.

The usability of mobile robots in the learning process and the short-term efficiency of our teaching method were checked with a control group after a semester (n = 149). We examined the effects on our students' programming skills and on their motives, mainly on their attitudes and programming self-concept. After a two-year-long period we could measure some positive long-term effects.

Keywords—programmable mobile robots; project method; positive effects

I. Introduction

Programming is not a compulsory subject in IT courses for students at Hungarian high schools. The National Curriculum aims at developing the skills of writing algorithm and developing algorithmic thinking but the skills of programming are taught only in classes that prepare students for IT graduation at advanced level. Because of this, many IT students start to acquire the elements of programming and program planning only during their college studies. Some previous research results [1] proved that those students who had already learned programming at high school were much more successful in the programming courses at our college. This advantage does not depend on the weekly number of high school lessons.

In contrast, beginners usually cannot pass their first exams. Lecturers often notice a decreasing increasing interest in programming. We supposed that abstract thinking means a great problem for beginners.

The Hungarian empirical research results [1] are supported by some results from other countries. In their comprehensive study on the Greek secondary school system Sartatzemi at al. [2] paid attention to the problems of teaching programming. They emphasize, there are some essential difficulties for those who have just started to learn programming:

- The professional programming languages are too complicated for beginners, in spite of the fact that these languages provide a wide range of solutions. Students usually have to focus rather on the language than on the problem itself. Accordingly, the implementation of a simple algorithm demands high-level thinking abilities.
- Professional programming environment is usually more complex than it is necessary for beginners. The environments do not help a beginner with the identification of syntax errors. The error messages in professional environments are made for professional programmers, not for beginners. The complexity of the environment can be shocking for students.
- During the first semester students cannot solve interesting problems. In order to enable them, they have to learn not only the programming language but the methods of developing larger programs as well. It is not possible during one semester. The grounding often seems too hard and boring for beginners and can decrease their motivation.

To sum it up, students have to focus not only on algorithm. They meet the principles of programming, the structure and syntax of the language, machine control problems etc. In addition, they have to learn the methods of program planning, developing and debugging.

The results of Sartatzemi et al. confirmed that new devices and methods are necessary in order to make the learning process more effective. Researchers usually propose different approaches so that beginners could cope with programming difficulties and with the complexity of programming languages. Some of them suggest that the object-oriented paradigm is more usable in teaching programming than the functional paradigm [3]. However, this change does not give a solution for the above-mentioned problems. Other researchers prefer a possible "learning programming language" [4] with an optimal environment and strongly limited set of statements.

We wanted to introduce new devices and methods into the learning process of programming. We looked for a method to improve the participation of students and increase communication among them. At the same time, we wanted to make devices more tangible. We aimed at making the learning process more concrete, practical and interesting for our students. The programmable mobile robot Mindstorms RCX (made by LEGO) appeared suitable for the realization of our aims. These devices and their programming environment allow students to learn in a natural, experimental way. Abstract thinking can be preceded by the manipulative and concrete usage of skills [5]. It can facilitate the development of skills and can deepen the level of understanding [6]. At the same time, the students' motives can increase due to the success of these learning situations. The co-operation among students can strengthen students' social and communicative skills [7]. These experience-based learning situations can lead to experiencing the growth of knowledge and can result in a higher level of students' activity. In a well-organized learning situation students can feel the flow. This is a mental state in which the students are fully immersed in concentration and the enjoyment of the activity [8]. These very motivated periods may accelerate skills development and may increase the efficiency of learning.

Our students' learning performance and the level of the adaptability of their knowledge are influenced by many factors. The effects of some factors have been revealed by researchers. The previously acquired levels of knowledge and skills have naturally a significant effect on the knowledge to be mastered. However, the individual differences in prior knowledge are not merely sufficient to explain the differences in further learning performances [9]. Additional factors have an important role in the learning process. The individual level of learning motivation (with many sub-factors) as well as family background or the system of social relations may have an effect on the students' learning success.

II. Real and Virtual Robots in Education

A. Overview

One of the most interesting and most difficult problems in the field of artificial intelligence is to create and apply intelligent robots. Real robots have to work in a noisy, nondeterministic, continuous space and time it makes its necessary to solve a number of additional difficulties. Thanks to the burgeoning of low-cost high-performance computers, we are able to simulate robots in a virtual space. Working with robotic simulators programmers can focus on the algorithm, neglecting many of the real world's aspects.

In education, the question arises whether the use of real robots or the use of robotic simulators is more efficient in the development of students' programming skills. Using simulators the teacher can create and change the teaching environment. The complexity, inventiveness and realism of the environment can be adapted to the students' skills and age. However, students cannot “touch” the robots simulated on the screen. Because of this, the manipulative skill may be incomplete; this can cause difficulties in the process of interiorization [5].

Until the last decade robot simulators could be usually found in industrial applications. Additionally, some robot-specific simulators were used. In the last decade, possibilities of robot simulators moved towards general usability. By a special plug-in of MathLab, we can simulate punctiform robots or robots with real expansion [10].

An advanced simulation environment is the Webots mobile robotics simulator which is a commercial product developed by Cyberbotics[11]. It can simulate rolling, walking or flying robots. Additionally, this simulator can control some types of real robots (e.g. Pioneer, LEGO Mindstorms, Aibo) [12]. Repast (developed by the University of Chicago) is an open source, multi-agent simulation package based on Java [13] . The basic concepts of Repast were borrowed from the simulation environment of Swarm agent [14]. These simulators are used mainly in research.

In the past half a century some famous robot simulators were used in education too, e.g. Papert's turtle [15], Karel the robot [16] or the Spider World, used by Dalbey and Linn [17]. The Lego Mindstorms Simulator (LMS) developed by the University of Paderborn was also very popular in education. Some empirical experiences of teaching with robot simulators are shown in the next section.

B. Research result

Sartatzemi et al. [2] used Mindstorms RCX mobile robots and ROBOLAB as a programming environment. In a ten-hour course (two hours/day), 14 students solved simple tasks in an icon-oriented environment. The teacher presented ROBOLAB structures in the first part of the lessons then the students solved tasks on worksheets. Researchers concluded that Mindstorms robots and the new programming environment are an efficient and practical way for high school students to learn programming. Their empirical data and the assessment showed some important conclusions. Students can easily acquire knowledge about procedures and the controlling of robots but this knowledge is often incomplete and inaccurate. The application of a real system is useful to analyze and solve a problem. Additionally, the students can check and debug their solutions in an experiential and clear way. It seems that students can understand the basic principles of programming in this environment. However, researchers observed some difficulties. The internal difficulties of the structures of the programs are similar to other environments; they can lead to misunderstanding. Because of this, the development of larger programs seems difficult for students. Furthermore, it was difficult for many students to connect the behavior of a robot to the logic of the program.

In their experiment, Wu et al. [18] compared the effectiveness of teaching with real and simulated robots. One of their groups (75 students) used LEGO RCX or LMR robots, the other group (76 students) used LEGO Mindstorms Simulators. Both groups consisted of beginners in programming. Similar pre-knowledge was supposed, so researchers did not use any pre-test. It was a short-term experiment (seven weeks; two hours/week); because of this researchers assessed the short time effects only. Pre-written templates in leJos (Java) language with simple program structures, basic variables and functions for controlling motors, lighting and crash sensors were used for the tasks. As their empirical result showed, there was not any significant difference between the two groups in understanding the pre-written programs and programming. However, those students who used real robots, showed a more positive attitude towards learning. Students in this group could easily imagine the behavior of robots but the usage of real robots demands extra time.

Kamada et al. [19] assumed that computer-controlled machines are used in almost all areas of life. Because of this, it is advisable to learn about the mechanism and the controlling opportunities of robots at the same time. Researchers assumed that the simultaneous learning of hardware and software can lead to an easier recognition and correction of errors of computer-controlled devices. Students built their own Myurobo robots, after that they tried to control them using Dolittle programming language. The simple and cheap robots could be used not only at school but at home as well. During the process of building the robots students acquired knowledge in the field of mechanics. The structure of the robots was similar to a "glass box", so the wiring of mechanical and electrical parts could be seen. According to the teachers' opinion, this way, the mechanical and electrical structure is easier to understand for the students. The experiment was organized as a pilot-project for high school students. The researchers did not aim to assess the effectiveness of learning; they were interested in whether these devices can be applied in education. The project lasted only 10 hours: two hours to learn the programming language, four hours to build the robots, four hours to solve a programming problem. The teachers wanted to find a programming language which can be learnt easily by the students. The low price (€20) and the "glass box" style of the device were attractive for the participants. An additional benefit was the simple control language. However, the environment had some disadvantages: the difficulty of serial-to-USB conversion and the fact, that there was no real-time control from the computer. The transformed control language also showed some differences from the original Dolittle language. Based on the feedback from the participants, the researchers considered that a revised, object-oriented Dolittle language can be applicable in high school education.

Kurebayashi et al. [20] prepared a proposal for a new curriculum for primary and secondary school students. They suggested introducing a new practice-oriented subject with tri-axial robots. With these robots, the elements of mechanics, electronics and information technology may be taught in one context. They examined whether the complex way of teaching with embedded systems is more effective than the traditional methods of teaching. The new curriculum was tested on a sample with 123 high school students. Students built robots and prepared their programs. After that a competition was organized for the robots. The effectiveness of the curriculum was measured by questionnaires for students and teachers. Based on the educators' feedback, the new curriculum resulted in positive effects and high efficiency. As the students looked back, building and programming a robot was a hard but enjoyable task. Many of them planned to continue learning about robot programming; this new curriculum sparked their interest in complex, systematic learning. To sum it up, the curriculum and its content may help with the teaching of complex embedded systems.

Fagin and Merkle [21] investigated the advantages of using robots in teaching information technology. They organized a control group experiment with more than 800 students and observed them for a year. It was expected that robot assisted learning would encourage students to choose computer engineering, computer science or any related field during their college studies. Additionally, researchers supposed that the robot can be a motivational device for students. Furthermore, they expected that programming skills would develop more significantly in the experimental group than in the control group. In the experimental group, the students used LEGO Mindstorms robots and Ada/Mindstorms as a programming environment. An additional aim was the acquisition of elements of the Ada language. In comparison to the above mentioned research results this experiment showed negative effects on the programming skills. The performance of students with the robots was significantly lower than in the control group taught with traditional devices and method. There are several possible reasons for this. After uploading, students had to compile and debug their programs on the robots; for this, more time was necessary compared to the traditional method using only computers. Another reason for errors may have been that teachers were well prepared for the lessons, but they had also used the robots for the first time, so they did not have sufficient experience in organizing robot assisted lessons. As the researchers summarized, despite their potential positive effects, the robots are not panaceas in education.

III. Short-Term Effects of Using Project Method and Model Robots at Our College

A. New course: new method and new devices

Because of problems described in the introduction, a new course was developed at our college. Our students learnt this course in a non-traditional way. LEGO NXT, LEGO RCX and Surveyor as programmable model robots were used to teach the elements of programming for IT-engineering students [22]. This new course can be taken by students who have successfully passed their "Programming 1." exam in C/C++ language. That is why NQC and NXC programming languages were chosen for this course. Syntax, statements, functions etc. of these languages are very similar to those used in standard C language. We did not put emphasis on the knowledge of the internal structure of robots. Our aim was to deepen our students' programming skills and algorithmic thinking, as well as to improve their attitude towards programming with these tangible devices.

NQC and NXC languages also contain loops, conditional statements, functions, tasks and included files similarly to standard C. From an educational point of view, one of the most important features is an easy way to run our program: we can upload it to the robot via Bluetooth, and check it immediately and visually.

In this course we rarely used traditional teaching methods e.g. teacher's presentation, but we often used methods giving an opportunity for constructive learning. The most preferred one was project method. Similarly to the business sector, in a project process the analysis of the problem, planning the steps towards their own solution and the implementation are carried out in groups [23]. The rigid and commanding knowledge transfer function of teachers has changed. Primarily, their roles are to raise the problem, to provide sufficient resources for work and to co-ordinate students' work. Simultaneously, the importance of their preparatory role has grown. Group members plan the process, divide the tasks among themselves, communicate to each other and, at the end of the work, they jointly summarize and present their results.

During this process, students could acquire theoretical and practical knowledge. This knowledge may be applicable in our students’ future IT-engineering job and in their software developing projects as well. Teamwork can have positive effects on their communication skills because their thoughts, their ideas have to be expressed understandably but in a professional way [24]. Most courses of higher education rarely give opportunities for professional communication among students because of the high number of students.

In details, in our ”Model robot programming” courses we aimed at following the principles, methods and processes of the constructive approach of teaching. Only 20% of time was used for teacher’s presentation and explanation. During this period, the teacher introduced the subject, the necessary functions and the elements of the language via examples. In the remaining time, groups of two or three participants solved programming problems. The co-operation among the teammates was an important factor because they had to recognize the problem, to find a possible solution and to divide the job. As a teacher, we did not play a traditional knowledge distributor role in this phase. Instead of this, we had to support, motivate and co-ordinate our groups’ work. We could help with identifying the main points of the problem, with accessing useful resources and sample libraries, with the accomplishment of the independent research etc. In the most significant period of the learning process the groups had to construct their robots and build them from LEGO Mindstorms or Surveyor parts. In this period they had to make and check their algorithm, write and debug their program. Additionally, a documentation of their solution with their plan, photos, video clips had to be prepared. At the end of the project each group presented their solution to the other groups and answered their questions. During the evaluation of the project, in addition to the teacher's reflections, self-evaluation, the other teams’ and the teammates’ evaluation also play an important role.

Our courses provided opportunities for collaborative knowledge building [25]. In a process like this, the understanding and interpretation of problems can be strengthened. The individual's activities for personal understanding are associated with social knowledge building [26].

B. Short-term effects

With the new course introduced in the previous chapter, we wished to decrease the problems mentioned at the introduction. Based on the college course system, it was not possible to conduct an experiment for more than half a year. That is why we decided to monitor our students' results later in order to demonstrate the effectiveness of the development of motivation.

We presupposed that the usage of tangible devices may accomplish the activation and improvement of learning motives and the acquisition of basic elements of programming simultaneously [27].

H1: Real tools make learning more enjoyable.

The feeling of knowledge growth and joyful learning may lead to the flow state. It can work as a very strong learning motive. In addition, the gradually more complicated tasks may ensure a lasting strength of challenge. In this situation, the mastery motive can be activated and may play a fundamental role in skill acquisition.

H2: Programming self-concept can be improved with the use of robots.

The experiences obtained in robot programming and the achievements in problem solving tasks have an effect on students’ self-confidence.

H3: The tasks solved by students at the concrete operational level have an impact on the development of abstract programming skills.

Acquiring programming at the abstract operation level often proves to be too difficult for starting programmers. We supposed that learning with tangible devices can enhance the acquisition of the abstract knowledge elements.

1) Methods

To verify the hypotheses, we organized a study with experimental and control groups. All students in the study took our course “Programming 1.”. During the semester of our experiment, members of the experimental group (n1 = 73) used LEGO NXT robots with the methods introduced in chapter 3.1. Members of the control group were taught by traditional teaching methods.

We used a test with 15 items to assess our students’ programming skills and knowledge (Cronbach-α = 0.86). Most Mitems required a short answer. In these items students had to understand short pieces of a program, after that they had to complete or debug them. We used the same test means for the pre-test and post-test.

In order to assess our students’ programming self-concept and attitudes towards programming, we used a questionnaire containing 17 questions. To the majority of questions students could choose their answers from a five-level Likert-style response list. Six questions used for assessing the programming self-concept, were arranged into one factor (KMO = 0.87). We aggregated these variables into one new variable without weighting. This new variable was rescaled on a percent-point scale. The questionnaire contained some additional questions about students’ social background.

Some more questions were asked in the post-test questionnaire. These questions concerned the hardness and fun of the work during the experimental semester.

The result of our two sub-samples in course “Programming 1.” was very similar ($\chi 2 = 3.86$; $p = 0.38$). The pre-test difference between the experimental group and control group was not significant in their programming pre-knowledge and in their programming self-concept (Table 1).

There was a small, non-significant difference between the two sub-samples in the number of programming courses at high school ($2 = 5.42$; $p = 0.27$). Nearly half of the whole sample (46% of students) had not learned programming at high school.

TABLE I. PROGRAMMING PRE-KNOWLEDGE AND PROGRAMMING SELF-CONCEPT IN THE SUB-SAMPLES IN THE PRE-TEST

	Experimental mean (st. dev.)	*Control mean (st. dev.)*	*t (p)*
Programming pre-knowledge (%p)	44.6 (19.1)	41.3 (19.5)	1.36 (0.17)
Programming self-concept (%p)	47.2 (19.7)	46.3 (21.5)	1.57 (0.14)

Remark: we used Levene-F to compare standard deviations. The difference between standard deviations was not significant.

Based on similar results of our experimental and control groups, we supposed the differences at the end of the experimental semester were due to educational effects. The next chapter presents these changes

2) Development of the experimental and control group

We could not observe any significant development of programming skill either in the experimental group (xpre = 44.6 %p; xpost = 47.9 %p; t = -1.23; p = 0.23) or in the control group (xpre = 41.3 %p; xpost = 43.6 %p; t = -1.01; p = 0.34). The Pearson-correlation between the pre-test and post-test results are similar in the two sub-samples (rexp = 0.63; rctrl = 0.62). These results showed that the learning process with tangible devices had not directly affected our students' knowledge.

However, we could measure an important and significant difference between the experimental and control groups in the field of learning motives. During the semester members of the experimental group were absent on significantly less occasions than the members of the control group. The learning process was much more enjoyable for the experimental group (on a five-grade scale: xexp = 3.47; xctrl = 2.96; t = 3.87; p < 0.01), and they felt the course less difficult (xexp = 3.07; xctrl = 3.35; t = 1.96; p = 0.03). Simultaneously, the students' attitude towards their teacher did not change significantly (xpre-exp = 4.03; xpost-exp = 4.06; xpre-ctrl = 4.03; xpost-ctrl = 4.12), so we can suppose that the changes in the students' motives are not consequences of their teacher's personality.

The average programming self-concept of the control group remained unchanged during the experimental semester (xpre = 46.3 %p; xpost = 44.1 %p; t = -0.45; p = 0.66). However, the results showed significant changes in the experimental group members' programming self-concept (xpre = 47.2 %p; xpost = 52.2 %p; t = -2.60; p = 0.01). Differences between our sub-samples were also observed in the distributions of this variable.

These results showed that despite the short-period, significant changes in students' programming self-concept can be achieved using new devices and teaching methods. This is very important for the students' future learning performance because of the strong effect that well-developed self-concept has on learning achievement [28]. With monitoring students further we want to verify if the well-developed self-concept results in any additional programming effectiveness.

C. Assessment of the durability of effects

1) Questions of our research

As presented in chapter III.B, a short-term post-test showed positive effects on students' self-concept in the experimental group. However, we can check the durability of these effects and the usability of acquired knowledge only after a long-term period.

An important problem when organizing the long-term post-test was that most of our students in the sample had completed their college studies. That is why we could only involve our former students with available contact details in the assessment. The sample of the long-term post-test was limited by this fact. An additional problem was that it is almost impossible to organize a control group as it is very hard to create a sample whose members studied at the same time, whose previous measurement results and contact details are also available. So the assessment is based on the responses of students who studied our course previously.

Our analysis is primarily intended to clarify if the attitudes are long-lasting since the "Model robots programming" course was taken towards the topic and self-concept related to mobile robots programming. We also wanted to explore whether the beneficial short-term changes can be transferred to other areas of programming.

2) Methods

In the study 33 people took part. Previously, all of them had been involved in the course and the experiment introduced in chapter III.B. The total sample's average age is 30.9 years, standard deviation is 6.1 years. 36% of the sample was full-time, the others were correspondence students. Obviously, the full-time students' sub-sample (x = 26.2) was significantly younger than the correspondence students' (x = 33.6). The sample was considered to be a normal distribution of age. The sex ratio was not significantly different from what we can observe at the whole faculty, so we did not analyze the data in sub-samples of women and men. 15.2% of the sample was female.

We compiled a new questionnaire to assess long-term affects. The questions were related to completed studies as well as to other studies since then. We asked some questions about our former students' current job and its relationship to IT. The questionnaire consisted of 18 questions. These assessed the actual attitudes towards programming as well as to self-concept related to programming and mobile robot programming. Respondents could choose their answers from five-grade Likert-style lists.

An additional question was used for assessing our respondents' programming self-concept based on social comparison. They had to imagine a fictive situation where they had to fill in a 50-point programming test. Every respondent had to assess how many points he/she could collect if the average performance of his/her team mates was 35 points. So they had to give a norm-oriented assessment. This question could measure the respondents' self-concept [9] .

The questionnaire was sent to all former students of our course. They could answer the questions electronically. The questionnaire was sent back by 62% of those students whose contact details were available.

3) Results

The average length of pre-college IT courses was 3.5 years in the sample but 20% of respondents learned IT for only one academic year. The mean was significantly higher in the sub-sample of full-time students (x = 4.8 years) than in the sub-sample of correspondents (x = 2.6 years). The primary cause of this difference was the different age of the two sub-samples. We could observe a strong, significant Spearman-correlation between the age and the number of pre-college years with pre-college IT learning ($r = -0.60$; $p < 0.01$).

In the fields of programming, the mean of pre-college years is 1.6. There is a moderate but significant Spearman-correlation between the number of pre-college years with IT learning and the number of pre-college years with learning programming ($r = 0.55$; $p < 0.01$). The correlation between the age and the number of pre-college years with learning programming is also negative but lower, so the older sub-sample spent relatively more time with programming. It can be caused by the changes in the curriculum or by the changes in the fields of interest.

Students had learned our course for 3-5 years before our long-term post-test, so we did not analyze whether the answers depended on this variable. Because of the short period after graduation, only a very small proportion (just two people) gained further qualifications. 42.6% of the respondents deal with programming in their current work.

At the end of the ”Model robot programming” course, the average mark was 4.6 (on a five-grade scale in Hungarian schools). This mean was calculated based on the memories of our respondents. It is similar to the average of the official results of all students who had passed this course (x = 4.56; n = 127). However, this mean is significantly higher than usual at the end of ”Programming 1.” courses (usually it is around III.A), despite the fact that the programming skills to be acquired are very similar in both courses. This may partly be caused by the fact that the students, who registered for this mobile robot programming course, had better pre-knowledge and motivation within the population. But in our sample the mean of ”Programming 1.” course mark was only 3.27, that is why we suppose that the difference is due to our experiment.

During further analysis we had to emphasize that quite a long time had passed since learning our course. So the long-term post-test could only analyze the durability of the long-term effects of attitudes and motives.

The ”Model robot programming” course was considered easier (f = 60.6%) or much easier (f = 36.4%) by the respondents than other programming subjects. Only one respondent answered that this subject was more difficult for him and no one answered that it was much more difficult. There is no difference in this question between the full-time and the correspondence students ($F = 1.97$; $p = 0.17$; $t < 0.01$; $p > 0.99$). The feeling of difficulty was almost independent of age ($r = 0.02$; $p = 0.94$).

The durability of the respondents’ positive attitude towards mobile robot programming was indicated by the great proportion of those (f = 90.9%) who found the subject much funnier and more enjoyable than other subjects. Only one ex-student remarked that it was as funny as any other programming subject. The difference between full-time students and correspondence students was not significant in this variable (Welch-d = 1.77; $p = 0.10$) and it is independent of age ($r = 0.12$; $p = 0.63$).

This positive attitude may be caused by many factors. One of them is, according to the respondents' opinion, that this course provided by far more possibilities for student activity than other course. This positive attitude was similar in the sub-samples of full-time and correspondence student ($F = 0.03$; $p = 0.88$; $t < 0.01$; $p > 0.99$) despite the fact that the number of contact lessons is much less for correspondence students.

Neither in the opinion regarding difficulty, nor in the attitudes towards mobile robot programming were any significant differences between the sub-samples of those who work in IT sector and those of working in other fields („easier”: $F = 2.32$; $p = 0.14$; $t = 0.21$; $p = 0.83$; „more enjoyable”: $F = 1.52$; $t = 0.23$; $t = 0.59$; $p = 0.56$; „more possibility of activities”: $F = 8.51$; $p = 0.01$; $d = 1.79$; $p = 0.08$). Based on these results, we supposed that the effect of “past became beautiful” was not significant in the case of those who work in IT now.

An indicator of programming self-concept may be a norm-oriented comparison of the students’ own supposed result compared to his/her team’s or classmates’ supposed result in an imagined test (see Methods in this chapter). It is based on self-confidence and depends on self-evaluation compared to the peers’ results. We measured this factor on a percent-point scale with a range of 0-100. The mean of the whole sample was 35.2 %p. It is a significantly better result for programming self-concept than what had been measured in earlier studies. In this study the distribution of this variable was very asymmetric, 65.6% of the sample gave a higher value than the reference value of earlier studies. However, there were some extremely low values so the distribution significantly differed from the normal distribution ($Z = 1.65$; $p = 0{,}01$). Assumed forgetting can be the reason for it but it also indicates that with these respondents positive self-image did not last long.

There was a significant difference in this variable between the sub-samples of full-time and correspondence students (xfull-time = 39.9; xcorr = 32.7; $F = 11.40$; $p < 0.01$; $d = 2.19$; $p = 0.04$). The individual differences were much bigger in the sub-sample of correspondence students. This result supported our earlier experiments: one of the reasons for the choice of correspondence courses – together with family and social backgrounds – is the lower learning self-concept.

Those who work in IT sector had a small advantage in this variable (xIT = 37.6; xothers = 33.3) but this difference is not significant ($F = 0.51$; $p = 0.48$; $t = 1.07$; $p = 0.29$). This small difference could be a result of further workplace successes.

The self-concept related to programming and to mobile robot programming were assessed by six-six Likert-style questions. Both groups of these variables were arranged into one factor (KMOprog = 0.84; KMOmobile = 0.72). Based on this we aggregated these variables into two new variables and transformed them into percent-point scale. Both of the new variables showed a normal distribution (programming self-concept: $Z = 0.95$; $p = 0.33$; self-concept related to mobile robot programming: $Z = 0.45$; $p = 0.98$).

There was a great and significant difference between these two variables in the whole sample (xprog = 56.7 %p; xmobile = 88.5 %p; t = 6.77; p < 0.01). The Pearson-correlation between two variables is not significant (r = 0.11). This result showed that these factors are independent of each other despite the fact that mobile robot programming is a sub-field of programming. The mean of programming self-concept was similar to the results of earlier studies. However, the self-concept related to mobile robot programming was correlated to the above mentioned factor of norm-oriented comparison (r = 0.44; p = 0.02), increasing the validity of our result.

The self-concept related to programming and to mobile robot programming was similar in the two sub-samples of full-time and correspondence students (programming self-concept: xfull-tim = 55.9 %p; xcorr = 57.2 %p; self-concept related to mobile robot programming: xfull-time = 86.7 %p; xcorr = 89.5 %p).

IV. CONCLUSIONS

Our results indicate that the learning process with programmable mobile robots and with new teaching methods could improve the attitude towards mobile robot programming and self-concept, however, we could not observe any significant transfer effects to other fields of programming. This fact was underpinned by the negative and significant Spearman-correlation between the programming self-concept and the marks at the end of the “Model robot programming” course (r = -0.47; p < 0.01). The successful transfer of the mobile robot programming self-concept to other programming areas would need further positive results.

ACKNOWLEDGMENT

This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund within the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 ‘National Excellence Program’

REFERENCES

[1] R. Kiss, A.Pásztor, ”Using Programmable Robots in the Education of Programming”, Unpublished conference proceedings. Szakmai Nap, Kecskemét. 2006.

[2] [2] M. Sartatzemi, V. Dagdilelis, K. Kagani, ”Teaching Programming with Robots: A Case Study on Greek Secondary Education”, Lecture Notes in Computer Science, 37(46), pp. 502-512, 2005.

[3] M. R. Lattanzi, S. M. Henry, ”Teaching the Object-oriented Paradigm and Software Reuse: Notes from an empirical study”, Computer Science Education, 7(1), pp. 99–108, 1996.

[4] S.Brilliant, T.R.Wiseman, ”The First Programming Paradigm and Language Dilemma”, Proc. SIGCSE ’96 Symposium on Computer Science Education, pp. 338–342, 1996.

[5] J. Piaget, The Essential Piaget. ed by Gruber, H. E. and Vonèche, J. J. Basic Books, New York, 2007.

[6] J. Nagy, XXI. century and education, Budapest, Osiris Press, 2000.

[7] R.Pap-Szigeti,”Cooperative Strategies in Teaching of Web-Programming” Practice and Theory Systems Education, pp 51-64, 2007.

[8] M. Csíkszentmihályi, ”Flow. The psychology of optimal experience”, Budapest, Akadémiai Press, 2001.

[9] B. Csapó, ”The Surface Layers of School Knowledge: What Do the Ratings Reflect?”, In Csapó B. (ed.), The School Knowledge,. Budapest, Osiris Press,pp. 39-81, 1998.

[10] C. Moler, The Origins of MATLAB, 2004.

[11] O. Michel, ”Webots: a powerful realistic mobile robots simulator”, IProceeding of the Second International Workshop on RoboCup. Springer-Verlag, 1998.

[12] O. Michel, ”Professional Mobile Robot Simulation”, International Journal of Advanced Robotic Systems, 1 (1), pp 39-42, 2004.

[13] N. Collier, T. Howe, M. North. Onward and upward,”The transition to Repast 2.0.”, In Proceedings of the First Annual North American Association for Computational Social and Organizational Science Conference, 2003.

[14] N. Minar, R. Burkhart, C. Langton, M. Askenazi. ”The Swarm simulation system: a toolkit for building multi-agent simulations”, Technical report, Santa Fe Institute, Working Paper 96-06-042, 1996.

[15] C. J. Solomon, S. Papert, ”A case study of a young child doing turtle graphics in LOGO” AI Memo 375. Massachusetts Inst. of Tech., Cambridge, Artificial Intelligence Lab. 1976.

[16] R.E. Pattis, ”Karel the robot: a gentle introduction to the art of programming”, Wiley & Sons, Hoboken, NJ. 1981/1995..

[17] J. Dalbey, M. Linn, ”Spider World: A robot language for learning to program. Assessing the cognitive consequences of computer environments for learning (ACCCEL)”, Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, 1984.

[18] C. Wu, I. Tseng, S. Huang, ”Visualisation of program behaviors: physical robots versus robot simulators”, In: Mittermeir, R. T. & Syslo, M. M. (eds.): Informatics Education – Supporting Computional Thinking, Poceedings of the Third Conference on Informatics in Secondary School, Poland, July 1-4. pp. 53-62, 2008.

[19] T. Kamada, H. Aoki, S. Kurebayashi, Y. Yamamoto, ”Development of an educational system to control robots for all students”, In: Mittermeir, R. T. & Syslo, M. M. (eds.): Informatics Education – Supporting Computional Thinking. Poceedings of the Third Conference on Informatics in Secondary Schools – Evolution and Perspectives, ISSEP 2008. Torun, pp. 63-74, 2008.

[20] S. Kurebayashi, H. Aoki, T. Kamada, S. Kanemune, Y. Kuno, ”Proposal for teaching manufactoring and control programming using autonomous mobile robots with an arm”, In: Mittermeir, R. T. & Syslo, M. M. (eds.): Informatics Education – Supporting Computional Thinking. Poceedings of the Third Conference on Informatics in Secondary Schools – Evolution and Perspectives, ISSEP2008. Torun, Poland pp. 75-86, 2008.

[21] B. Fagin, L.Merkle, ”Measuring the Effectiveness of Robots in Teaching Computer Science”, Proceedings of the Thirty-Fourth SIGCSE Technical Symposium on Computer Science Education, 2003.

[22] A. Pásztor, Pap-Szigeti, ”Congruence Examination of NXT Robots in the Education of Programming at KF GAMF College”, Practyce and Theory in Systems of Education, Volume 3. Number 3-4 2008. 33-40.o. HU ISSN 1788-2591, 2008.

[23] I.Falus, Didaktika. Nemzeti Tankönyvkiadó, Budapest. 1998.

[24] S. Kagan,” Cooperative learning”, Ökonet Press, Budapest. 2001.

[25] M. Scardamalia, C. Bereiter, “Knowledge Building“, In Encyclopedia of Education. Macmillan Reference, New York, pp.1370-1373, 2003.

[26] G. Stahl, “Group Cognition: Computer Support for Collaborative Knowledge Building“, MIT Press, Cambridge, 2006.

[27] A. Pásztor, R. Pap-Szigeti, E. Lakatos Török, “Effects of Using Model Robots in the Education of Programming“, INFORMATICS IN EDUCATION An International Journal, 2010, Vol. 9, No. 1, ISSN1648-5831 pp. 133-140. 2010.

[28] A. Helmke, M. A. Aken, “The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study“, Educational Psychology, 87, pp. 624-637, 1995

Building future generation service-oriented information broker networks

A technical and legal joint perspective on the di.me case study

Sophie Wrobel
CAS Software
Karlsruhe, Germany

Mohamed Bourimi
MT AG
Ratingen, Germany

Eleni Kosta
TILT
Tilburg University
Tilburg, Netherlands

Rafael Giménez
BDCT Barcelona Digital
Barcelona, Spain

Simon Scerri
DERI NUIG
Galway, Ireland

***Abstract*—Future generation networks target collecting intelligence from multiple sources based on end-users' data and their social interaction in order to draw useful conclusions on enabling users to execute their rights to online privacy. These networks form a rising class of service-oriented broker platforms. Designers and providers of such network platforms during the design and development of their systems focus primarily on technical specifications and issues. However, given the importance and richness of user information collected, they should already at the design phase take into account legal and ethical requirements. Failure to do so, may result in privacy violations, which may, in turn, affect the success of the network due to increasing awareness with respect to users' privacy and security concerns, and may incur future costs. In this paper, we show how the di.me system balanced technical and legal requirementsthroughboth its design and implementation, while building a decentralized social networking platform. We report on our advances and experiences through a prototypical technology realizing such a platform, analyze the legal implications within the EU legal framework, and provide recommendations and conclusions for user-friendly service-oriented broker platforms.**

Keywords—di.me; online privacy; social media; software design; legal and ethical issues; broker platform; context-aware web services; user data

I. Introduction

Human beings in the modern, data-driven era are increasingly dependent on technology and systems to make information available for different purposes, with wide-ranging implications on society. Such technology needs to support transparent, conscious decision-making processes in order to earn (end-) users' trust and assist knowledge workers in gathering multiple perspectives and qualitative insights to form useful knowledge [1]. Popular online social networks (OSNs) such as Facebook and LinkedIn encounter difficulty in this area today in several respects and have been often criticized for establishing complicating user interfaces in order to discourage users from making informed choices about the handling of their personal information, which this paper addresses.

A. *Challenges facing modern online social networks*

To some extent the interests of software providers align with these of their users. The user wants to utilize software and the provider needs some amount of personal information to provide it. But besides the amount of data that is necessary and the restricted use of these data for a legitimate and obvious purpose, the commercial interest of providers contradicts the interests of individuals. From the provider's perspective, user information represents a valuable asset. Hence, providers and their commercial customers have a strong interest in collecting and processing more information about their users, e.g., in order to improve their protfolios, or to offer customer-oriented services. This counts especially for OSNs and web services in general, which have been a trend in the recent years. Many social networks and services are free of charge. Their business is at least co-financed by innovative exploitation and commercialization of the users' personal data [2]. As a result, design choices in OSNs reflect the provider's financially-driven goal of maximizing personal data exploitation.

In order to obtain and commercialize personal data, social network interface design has evolved to encourage data entry. Research has been done on everything from the optimal warning message color, to presentation layout, or auto-complete suggestions based on the information available about a user's friends [3]. As a result, the user is often encouraged to incrementally provide more personal information, often without fully understanding the consequences this may have on their digital identity due to a lack of digital literacy: even in the 14-49 age group, digital proficiency lies below 60% in most major European countries [4]. Concretely, digital literacy

The di.me project is supported in part by the European Commission under the Seventh Framework Program FP7/2007-2013.

involves skill and understanding of social networking, transliteracy, maintaining privacy, maintaining identity, creating content, organizing and sharing content, reusing and repurposing content, filtering and selecting content, and self-broadcasting [5].

B. *The need to redesign for more privacy*

OSN platforms and services play an increasingly important role in all private and business activities. Two of the key challenges facing OSN users with limited digital literacy are the implications of data transfer, and the rights they have on their personal information. These challenges extend beyond the realm of OSNs: they are equally applicable to the Internet of Things (IoT), or any other electronic data broker transmitting information between two online services or parties within a concisely defined context.

At the heart of these challenges lie data protection issues. European citizens have a right to protect their personal data, which can only be collected and processed for specified purposes and usually on a consensual basis. Moreover, they have the right to request information about all collected data about them, and the right to ask their rectification or deletion [6]. GéraldSantucci, Head of the Knowledge Sharing Unit at the European Commission’s DG CONNECT, writes, “How can we have the Internet of Things (or the ‘Internet of Everything’) while preserving our fundamental right to privacy? Several answers exist, but we have seen that they can actually be clustered around two: the first one is technology itself - embedding privacy and security in the very design of new systems and components; the second one is adequate rules and regulations. A combination of technology and regulation can also be a wise approach [7].” So, how can technology and regulation together effectively support such negotiation?

As Lessig argues, if law can regulate software, and software can regulate individual behavior, then software provides lawmakers with an effective way to shape the way their subjects behave [8]. Following that paradigm, the software provider has a responsibility to ensure that data protection requirements and other privacy obligations imposed by legislative institutions are technically supportedbyconsidering them early in the design of the respective systems. The service provider has a responsibility to ensure that collected data is handled securely and appropriately, and network providers have a responsibility to ensure that communication channels are protected, while users are responsible for their conduct. Such responsibilities should be taken into account already at the design phase of systems and applications, in respect of the ‘privacy by design’ principle, which enlists technology to protect individual privacies by default on a continual basis [9]. By introducing law and ethics as core values during the software requirements and design phase, the resulting implementation could provide a solution that does not conflict with the right to privacy, or with exploitation interests (to some extent for current requirements)[1].

This paper considers how introducing these core values during the requirements and design phase of di.me resulted in a privacy-oriented service-oriented architecture (SOA), which has the potential to intelligently assist, without restricting, a safe and deliberate participation of less digitally literate individuals in popular OSNs. It describes the di.me context and architecture, and analyzes how di.me reacts in select use cases against critical data handling concerns that are commonly expressed against popular OSNs.

This paper focuses on the legal requirements relating to data protection affecting software design. The service provider and network provider layers lie out of the focus of this paper, although they are also affected by data protection regulation [10].

The remainder of this article is structured as follows: while the current section motivated the problem statement, the next section addresses di.me as a case study in this respect. Section III compares our contribution to related work. Finally, Section IV concludes our contribution and outlines potential future directions.

II. The Di.Me Case Study

di.me is a distributed OSN, which additionally serves as a personal information broker platform. It operates as a digital identity management tool, allowing users to maintain an overview of their data across various supported online services, such as LinkedIn or Twitter.di.me operates as a privacy-enhancing technology (PET) platform, by intelligently warning users when their online interactions involving data may lead to undesirable consequences. It also operates as a data exchange broker, by allowing users to share personal data with other online services in a secure and safe manner. By considering legal and ethical values during the requirements phase of di.me, as well as the subsequent system and component design and implementation, the result is a privacy-oriented information broker platform, which negotiates between di.me users and other OSNs to enable free-choice and context-specific data transactions [11].

A. *Situational description of di.me*

To describe how di.me operates and the issues it solves, consider a series of illustrative scenarios revolving around a typical modern individual, Alice. These scenarios will be treated from both a technical and legal perspective in the following discussion.

1) *Multiple digital identities*

Alice acts differently under different situations. For simplicity, consider two roles which Alice fulfils: (1) *business*: on a business trip, she meets a new potential customer, Bob. They exchange business contact information, and Alice invites Bob for a dinner conversation. During dinner, they make a verbal commitment on a business partnership. The following day, Alice sends Bob a sales contract. (2) *friend*: taking advantage of the travel opportunity, Alice does some sightseeing. She meets a friendly lady, Carol, at the beach, and excitedly posts about it on Twitter. They befriend each other on Facebook, where Alice posts pictures of their beach trip, and promise to stay in touch.

[1] Future or change requirements, e.g., after lab and user trials could result in changes that need redesign and retrofitting of the current implementation. The following Sections describe therefore our contributions for the current design and implementation of the di.me userware.

In di.me, these multiple digital identities are embodied in the form of profiles. A profile is a set of information about a user that she provides to other users or services. A di.me user then gives other persons and groups access to her profile by sharing a profile card with them. While a profile card does not contain any information itself, it is a context-specific access token allowing a particular recipient to retrieve the associated profile information [11].

2) Intelligent context recognition

Personal devices can be used to determine where a user is and what she does. Suppose Alice carries her mobile phone with her constantly, and does most of her work on her company laptop. When she is connected to di.me from her company laptop, there is a very good chance that she is working. Personal devices are just one contributing data type in Alice's context (e.g. geo-locational, attentional, nearby peers, environment conditions, IP address, etc.), from which her situation can be deduced – for example, whether she is actually working, or whether she is hanging out with her friends after work.

In di.me, the user's context can be deduced by considering the live contextual information stemming from her devices (e.g. desktop, mobile device). Each device has always one dynamic live context. Snapshots of this live context are saved as static situations [11].

3) Information flow management

Alice is an active digital community member: she shops online, pays through online banking, and even has digital health care records and smart utility metering. But these conveniences aren't always as convenient as she would like: Every time she visits a new online shop, she needs to fill in all registration information all over again. And with each online shop or online social network having its own terms and conditions in what the end user often perceives ascryptic legal language, she can't be bothered to read through them every time. On the other hand, she often finds herself wishing that some registrations could take place automatically – for example, automatic registration at all baby bonus programs at her favorite stores when her child's birth shows up on her digital health record. However, after a few purchases, she starts to receive invitations and advertisements on baby products from companies that she has never heard of, and has no idea who might have given them her contact information.

di.me wishes to tackle many problems from the privacy and legal common point of view. One of these problems is concernedwith data transfers without the knowledge of the users. It acts as an information broker by allowing Alice to share her information with the parties she wants to share it with, while warning her if she inadvertently tries to share her information with parties that she may not want to share her information with [11].

4) Broker platforms in the digital landscape

Today, there are many platform solutions specializing in the sphere of contextualized information. Some focus on connecting information from entities, characterized as "big data", and others focus on connecting people, often called "social media". But these two trends are closely connected to tosome extent: they both deal with sharing contextualized information, which gives rise to service-oriented digital intelligence – a space in which broker platforms assist users to achieve meaningful information connectivity that is not addressed by popular market solutions: a mediating platform that can connect between data-driven platforms with people. As illustrated in Fig. 1 [12], this is a field that is largely unexplored by mainstream commercial offerings, but also a field which will flourish as a natural next step in internet connectivity.

Broker platforms in a SOA approach, like di.me, allow people-centric platforms to communicate with technology-centric platforms [13], while restricting data processing for a particular purpose in a defined context.Data brokerage in a service-oriented internet needs to consider not just technical but also legal implications, and define and negotiate responsibilities appropriately across multiple involved parties, including the user, the service provider, the network provider, and the software provider. While di.me itself does not facilitate negotiation, it does facilitate controlled data transfer in a user-centric way.

B. Technical description of di.me

The implementation of the di.me platform prototype technologically enables personal data usage in a controlled, trustworthy, and intelligent way [14]. It specifies a platform incorporating user-control deeply in design: a personal server (PS) that enables a di.me node in a decentralized network to connect to other users' PSs or external services, like various social networking platforms as mentioned above, and this by using distinct identities [15]. This node integrates all personal data in a personal information sphere, including user interests, contact information, files or resources, and social network services. Intelligent features and PETs further guide user interactions with the digital sphere, illustrated by context-aware access control [16], trust and privacy advice, or organizing their personal information sphere [14]. Besides integrating existing networks and services, the platform provides its own OSN functionalities, which are not available in known and popular OSN, in particular network anonymity [17][18][19].

1) Semantic model: information classification in di.me

The di.me Ontology Framework,based on the Personal Information Model (PIM) Ontology[2], is a differentiating concept allowing di.me to react to users with multiple digital identities, multiple use contexts, and differing objectives when sharing information. Each person in the di.me network owns a PS and an associated Research Definition Framework (RDF) store that contains the PIM representation. Amongst other information, the PIM includes references to persons, groups, service accounts (DAO), devices (DDO), resources (NIE), profiles (NCO) and live posts (DLPO). The PIM is extended byprivacy preferences (PPO instances), which enables the representation of databoxes, profiles and whitelists/blacklists, privacy and trust levels(NAO), andcontext information (DCON instances), which includethe unique live context representations of situations [20][21].

[2] Ontology descriptions are available underhttp://www.semanticdesktop.org/ontologies/ with exception of the PPO, which can be found under http://vocab.deri.ie/ppo

Data-driven Intranet *Connecting Knowledge* • Internet of Things • Semantic Search • Open Data Alliances	**Service-oriented Intranet** *Connecting Intelligence* **<Future: di.me and other broker platforms>**
Static Intranet *Collecting Information* • Traditional Websites	**People-centric Intranet** *Connecting People* • Social Networks • Communities • Talent Markets

← Information Connectivity →

← Social Connectivity →

Fig. 1. Four quadrants of internet platforms for technology-mitigated information connectivity [11].

This extended ontology set, depicted in Fig. 2, combines information from personal and contextual spheres, which together with the trust and recommendation engines allow di.me to identify context recognition as well as to derive privacy recommendations.

2) System context: deriving contexts but protecting identity

di.me's global architecture follows a decentralized approach emphasizing near real-time asynchronous network interoperability, data-centrality, and user control. The PS, hosted in the Personal Server Layer (see Fig. 3), is the central element in the system architecture, being responsible for collecting, safeguarding and managing the entire user's data.

Client applications triggered from user's personal devices provide light-weight user interfaces to access the PS. Communications between the personal devices and the PS pass through a proxy layer to minimize traceability. The personal server is responsible for holding the user's information and providing computational capabilities, and can be securely deployed on the user's personal devices, on trusted commercial hosting services or in a hosting service provided by the di.me system. These concepts are well-aligned with those being pushed today by relevant initiatives within the distributed social networks scenario [22].

The wide range of devices allow for di.me to derive contextual information: Usage of a certain device in connection with particular users or a particular location can imply a particular context. For example, Alice sharing a document from her laptop connected from her office IP address implies that she is probably in her 'business' profile in an 'at work' situation. In order to protect Alice's identities from being traced back to her, her requests are routed through the di.me proxy layer.

3) System architecture: powering smart recommendations

The PS itself comprises of multiple components which work together to provide intelligent analysis of identity and context information provided by the clients. Its high-level PS internal architecture, shown in Fig. 4, is related to that of dynamic webapplications [14], and also favored by the separation of the addressed concerns inherent to the multi-layersystems.

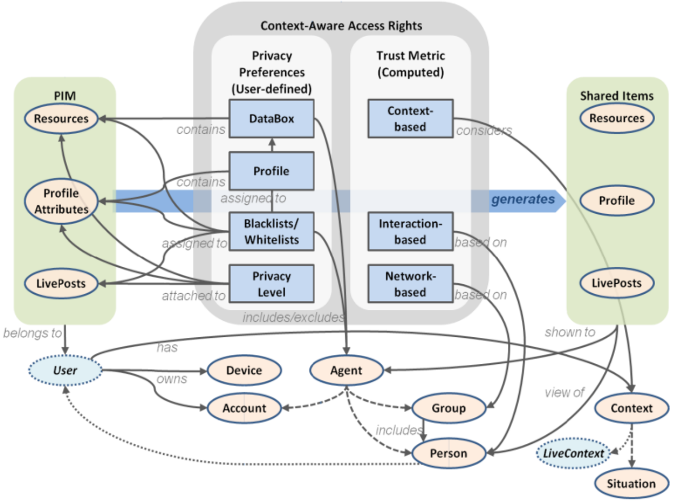


Fig. 2. High-level di.me semantic model description.

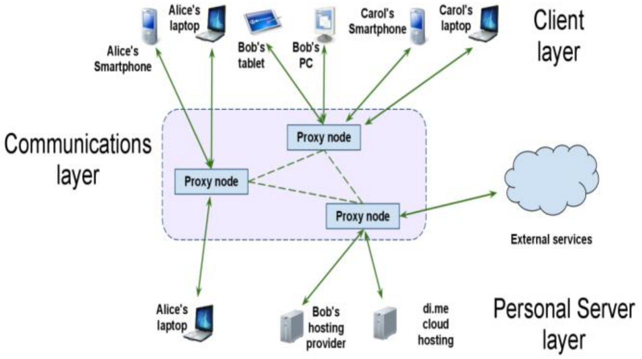


Fig. 3. Global di.me system architecture schema in a large-scale deployment.

Within this approach, the persistence layer isolation also benefits the decoupling from the underlying database technology and enables a multi-engine deployment. This feature is especially useful for the di.me system, intended to store heterogeneous data with fairly different access requirements such as the user's personal data, context data or service crawling schedule information.

The semantic and storage modules are used to store semantic as well as environment data. Semantic data includes information required for semantic deduction, such as 'the beach' – which could be a potential location nearby. Environment data includes information required for system operation, but without a semantic value, such as the strengths of the nearby wifi access points. In addition, the semantic module crawls connected web services to retrieve associated data at a pre-defined refresh interval. For example, Alice can connect to twitter, and the crawler would retrieve tweets, profiles, and friends and followers once an hour. The contextprocessor module derives contextual information from environment data.

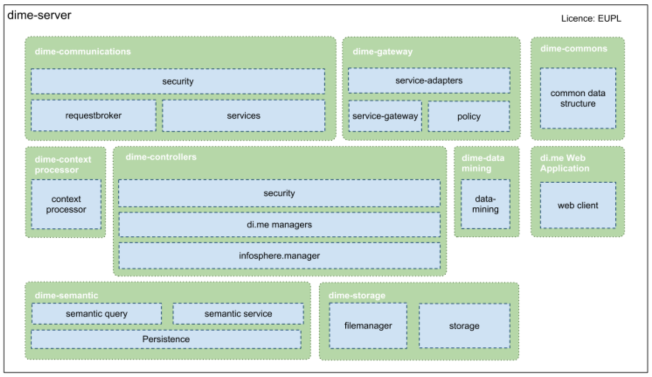


Fig. 4. Multi-layered architectural model for the personal server.

For example, based on how strong nearby wifi signals are and what the nearby wifi network access pointnames are, the contextprocessor can determine when Alice is in the vincity of her office network. The datamining module derives an adaptive privacy score to persons and to data indicating how trustworthy that particular resource is and checks whether data updates trigger a warning. The privacy score is calculated according to the di.me trust model which accepts inputs from the di.me semantic store and outputs a probability score, which adapts over time with respect to the user's interaction patterns. The gateway module manages and transforms communication entering or leaving the personal server with relevant policy rules. For example, a 'no twitter at work' policy would prevent her from posting to twitter if she was in an 'at work' situation [21].

When Alice posts "Sitting on the beach with @carol!" on Twitter through di.me, di.me's semantic analysis of the message recognizes an activity (sitting), a location check-in (beach), and a person (Carol). Triggers set by the controller module in the datamining module are fired, as the combination of personal information relating to third parties (Carol's identity) and their common location (beach) is a potential privacy issue, and this causes di.me to present Alice with a warning message informing her about that risk, and asking if she is sure she wants to post [14]. Unlike popular applications where users are expected to have these digital literacy skills, di.me allows non-literate users to participate in online social networks while informing them of risks only during relevant situations and thus minimizing the likelihood that the warnings get ignored. The final layer is the authentication and authorization layer, which ensures that all transactions are only honoured when the credentials are valid.

C. Legal perspectives

In order to ensure the protection of individuals, the European legislation on data protection applies when the processing of personal data takes place. The data can be processed only under the grounds mentioned in the Data Protection Directive [22] and their processing has to respect the basic data protection principles. The obligations stemming from the data protection legislation have to be taken into account already from the designing phase of systems and applications ("privacy and security by design") [23].The European Data Protection Directive is currently under review. In January 2012, the European Commission presented its proposals for the reform of the data protection legal framework of the European Union, proposing the replacement of the Data Protection Directive with a Regulation, which was the outcome of consultation and debates of three intense years [24]. The proposed Regulation dedicates an article to the principles of data protection by design and by default[3]. According to this principle, both at the time of the determination of the means for processing and at the time of the processing itself, a controller must implement appropriate technical and organisational measures and procedures in such a way that the processing will meet the requirements of the Regulation and ensure the protection of the rights of data subjects.

Before moving into the detailed analysis of di.me from the legal perspective, a short introduction must be made to the terminology that is relevant for data processing operations. The term 'personal data'[4] is defined as 'any information relating to an identified or identifiable natural person ('data subject')'; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental economic, cultural or social identity. As regards the phrase 'identified or identifiable person', the possibility of matching data processed by a computer to a specific person will depend on a number of factors, such as who is doing the matching and what their technical capabilities are, what type of data is involved, whether other data are available to aid the matching etc.

'Data processing' is defined as "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction"[5]. It follows that the definition of processing is very broad, so that it is difficult to conceive any operation performed on personal data, which would not be covered by it. It is important to note that even the mere storage of personal data constitutes 'data processing', so that simply storing data on a server or other medium is deemed to be processing, even if nothing else is being done with the data.

The relative data protection legislation defines three distinctive categories of parties:

- *'Data subject':* the individual to whom personal data refer to.
- *'Data controller':*an entity which alone or jointly with others "determines the purposes and means of the processing of personal data"[6]
- *'Data processor':* a third party who simply processes personal data on behalf of the data controller without controlling the contents or use of the data.[7]

[3] Article 23 of the draft Regulation

[4] Art. 2(a) Data protection directive

[5] Article 2 of directive 95/46/EC [18], hereafter called Data Protection Directive.

[6] Article 2 (d) Data Protection Directive

The classification of an entity as ‘data controller’ or ‘data processor’ is of great importance, for several issues, such as who shall carry the obligations appointed to the ‘data controller’ by the Data Protection Directive and who is to define the details of the data processing. As a rule of thumb it can be said that the data controller is liable for violations of the Data Protection legislation, while the role of the data processor is reduced.

Under the regime established by the Data Protection Directive, a key concept is that of ‘data subject’s consent’. If the data controller obtains the data subject’s consent then he/she is broadly free to process the personal data. The Directive defines ‘data subjects’ consent’ as being freely given, specific and informed[8]. It supplements this in the substantive provisions when referring to consent as being ‘unambiguously’ given[9]. Indeed, the definition of ‘consent’ in the Data Protection Directive is quite restrictive, requiring that the data subject be clearly informed in advance of what he is consenting to and that any processing of the data going beyond what is disclosed to him will be deemed not to have been consented to, meaning that it will be invalid. Particular risks arise in the online environment since there is an increased danger that the data subject might not have been fully informed or might not understand exactly what he is consenting to.

The related EU FP6 funded PRIME project relates to a privacy and identity management system that was demonstrated through collaborative E-Learning and Location-Based Services (LBSs). This differs from a broker platform in that its scope is more heavily directed towards inter-service connectivity, and LBSs are just a subset of potential di.me contextual entities. PRIME developed a set of requirements for Identity Management Systems (IdMSs) translating the obligations of the data protection legislation into requirements for IdMSs [25][26]. The PRIME requirements list has been used for di.me, which considered them in its design and development process.

To illustratehow di.me addresses issues surrounding the ethics of data transmission and user privacy today, consider several relevant di.me API and behavior around some critical ethical concerns around data handling within the scope of the previously described scenarios:

1) Linkability: Transfer of data to other contacts

di.me respects and safeguards user privacy by using strong, secure pseudomization techniques[15][16][17][18][19]. Because di.me acts as the intermediary and not the end service[10], it is not possible to make the legal analysis very concrete on this aspect. However, it is an important aspect of ensuring that users can exercise their right to digital privacies.

Pseudonymity: di.me uses the *idemix*[11] library to create secure credentials for information exchange between profiles. *idemix* allows the desired pseudonymous credential exchange, while still offering the possibility to de-anonymize user pseudonyms when needed – such as in the case of abuse or for accounting purposes, as required by law enforcement or for financial transactions [27]. This enables di.me to operate by transmitting personal data only via secure credentials, and on a completely pseudonymous basis, which is critical in ensuring that multiple identities managed from one central point can be unlinkable in all information flows within the di.me environment and this at least at the technological level [17][18][19].

Data exchange profiles: Each user can adopt and manage multiple public and private digital identities [15], which can be completely unlinkable if he strongly adopts *idemix* as an anonymous credential system at the level of personal attributes, with special attention to shared attributes across different identities. This separation of profiles allows a clean separation of business and private data, and which private data is shared with which business.

Trust metric [14][16][28][29]: An adaptive user trust index allows warning messages to be displayed, preventing a user from unknowingly sending a confidential file to the wrong audience. This concept is also applicablebeyond the di.me prototype no interactions in and between social circles, such as friends or business contacts. The di.me trust metric is calculated based on several inputs, illustrated in Fig. 5, including:

- *pre-defined trust dimensions:* When Alice uploads a photo in di.me to share with Carol, the photo is automatically given a privacy value of high. She can change this if appropriate.
- *recognition of user context:* When Alice shares her photo, di.me recognizes that this is a potential risk situation.
- *previous interaction:* The trust model uses available information from the semantic engine about the sort of information Alice has shared with Carol in the past, the situation, purpose, and context under which Alice is in now, and the current privacy value of both the photo and of Carol in order to calculate a probability value for the risk involved.

2) When a risk is identified, this generates an advisory, which is presented in the user interface, and Alice sees an advisory asking whether she is aware of the privacy risk involved in sharing her photo.Tracking context in information sharing

When you share information in di.me, di.me reveals personal information relevant to the share (See Table I).Note that each information share is associated with a *saidSender*. This *ServiceAccount* is a representation of a unique combination of a particular profile card and a particular web service account. The profile card is an access ticket to a set of information, available at downloadUrl upon presentation of appropriate access credentials, with respect to a particular context.

[7] Article 2 (e) Data Protection Directive

[8] Article 2 (h) Data Protection Directive

[9] Article 7 (1) and 26 (1) (a) Data Protection Directive

[10] End seviceper definition from legal point of view

[11] https://prime.inf.tu-dresden.de/idemix/

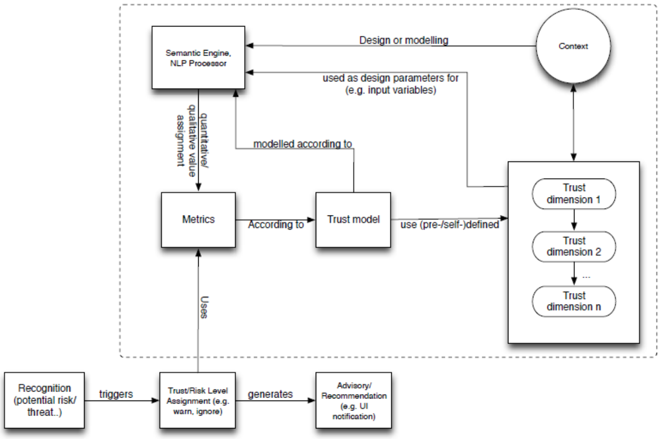

Semantic Engine, NLP Processor
Design or modelling
Context
used as design parameters for (e.g. input variables)
quantitative/ qualitative value assignment
modelled according to
Metrics
According to
Trust model
use (pre-/self-)defined
Trust dimension 1
Trust dimension 2
...
Trust dimension n
Uses
Recognition (potential risk/ threat..)
triggers
Trust/Risk Level Assignment (e.g. warn, ignore)
generates
Advisory/ Recommendation (e.g. UI notification)

Fig. 5. The system context of the di.me trust model.

When Alice shares her sales contract to Bob, this sharing is done through her business profile card. This contextual information is stored in di.me and is visible together with all other information related to the file or person involved in the sharing. Further, based on this interaction, di.me associates both the sales contract file and Bob with Alice's ‘business’ role. If Alice tries to share the sales contract with Carol, who is associated with her ‘friend’ role, di.me warns Alice: Carol is associated with Alice's ‘friend’ context, but not her ‘business’ context, and this would give Carol insight toAlice’s ‘business’ profile, which she does not yet have. Alice can then decide if the action was inadvertent, or whether the action was intentional and approve it. This functionality of di.me aims at protecting the privacy of users and raise awareness with regard to the sharing of their personal information. Although in principle users have the right to share their personal information whenever and with whichever entity they wish to, very often they do not realize that they are actually sharing personal information. di.me does not create profiles based on users personal data and context for any other purpose but to enable users to control the sharing of their personal information in an easy and comprehensive way. In this way it enables users to control better the information about them that they are sharing.

3) Designating purpose in information sharing

The second component of the ServiceAccount used for sharing information is a web service account. This is different for sharing between di.me users and sharing with other social networks. For example, when Alice shares the sales contract file to Bob directly using di.me, di.me creates a unique adapter for each contact’s profile that she shares to. More concretely, if she knew Bob as both a business partner and a friend, she could share the sales contract to Bob the business partner, to Bob the friend, or to both: di.me creates one web service account for each of the relationships she has with Bob, and this allows di.me to build an overview of the purpose associated to the data sharing by tying the purpose to the profile-specific web service recipient for information sharing.Alice is asked to choose which profile of Bob she wishes to send the file to and in this way she is offered the possibility to keep personal and professional information separate. This functionality actually assists Alice in determining the purposes for which she wishes to use specific information.

TABLE I. API DEFINITION FOR SHARING

Type	**POST /api/dime/rest/<user>/resource/@me**	
	Field name	***Description***
timestamp	created	When the user shared the information
string	downloadUrl	URL to access the shared information
GUID	Guid	List of service-specific configuration settings
string	imageUrl	URL to obtain a thumbnail of the shared information
timestamp	lastModified	When the information was last modified
string	Name	Name of the resource being shared
GUID	saidSender	Service account ID of the sender
List <GUID>	groups	List of groups to share with
List <GUID>	persons	List of persons to share with
integer	privacyLevel	Privacy level of the files being shared
integer	fileSize	Size of the file to be shared
string	mimeType	MIME type of the content being shared
string	Type	What is being shared. Valid values: resource
URI	userId	The user who is sharing (@me for current user)

On the other hand, when Alice shares her beach pictures to Carol by posting them to Facebook, this sharing is done under a generic information processing purpose as defined in Facebook’s user agreements. While this does not allow a specific data purposing to be shared over Facebook, it allows technical support for popular platforms which do not allow for specific data purposing in their API. To mitigate the damage that this could do, di.me’s architecture includes a multi-dimensional policy matrix which allows service providers, corporations or users to enforce desired technical guidelines, such as ensuring that collected data is consistent with minimum data collection requirements, providing a default trust metric value for information for a particular data source, or ensuring that predefined combinations of outgoing data are blocked. These functionalities of di.me enable service providers and corporations to comply with the data protection legislation. Allowing for the collection of only adequate, relevant and non excessive data in relation to the purposes for which the data are collected or further processed is a fundamental data protection principle, commonly known as ‘data minimisation principle’. By warning Alice when the data she is sharing is not consistent with the context she is sharing in, di.me assists Alice in protecting her data.

4) Erasure of data

Very few popular OSNs support data deletion, although they just may support hiding old data from the user’s visible experience[12]. As such, information shared via external services

[12] Providers must retain data for a specified duration as specified by data retention laws.

may not allow data deletion, and di.me cannot change this. For sharing inside of the di.me network, however, there is a mechanism to revoke access to data: the data is available only via the shared *downloadUrl*. If the di.me personal server hosting the *downloadUrl* happens to be offline, then the data displayed is shown from cached values that are updated at the next successful regular crawler synchronization, at which point the old values are updated with the current information available at the *downloadUrl* – which could include that the data has been deleted. The result is that when Alice removes the 'phone' attribute from her 'friend' profile, Carol will not be able to see Alice's phone number anymore. di.me enables users to erase their data, without requiring any activity from the party that holds their information. In this way di.me provides an advanced functionality allowing the users to exercise their right to erasure of their data.

5) Control over data

di.me crawls data from all connected services on a regular basis and stores this data in its semantic store in order to provide the user with context-specific trust and privacy warnings. This data is crawled at regular intervals and refreshed, replacing old data from connected services. It does not broker data to third parties without explicit consent; each user can only share his own personal data with other services. However, di.me can be operated in single-user and multi-user modes. In the single-user mode, a single user runs the server and controls the data on the server for private use. The more controversial scenario is the multi-user mode, in which multiple users share a single di.me server instance. Each user still only has access to his own data, but the data is stored on one communal infrastructure. di.me allows users to have full control over their data when it is operated in single-user mode. When di.me is operated in multi-user mode, profiles are still maintained separately: there are no common profiles even if two data owners share a mutual contact. This allows di.me to ensure data set access in the same manner as when operating in a single-user mode. In this way, di.me enhances the transparency of the transactions and allows user to remain aware of any data sharing that relates to them.

6) Data monitoring

di.me crawls connected external services on a regular basis and alerts the user of substantial changes. For example, if it detects that Alice has befriended Bob on Facebook, and that there are so many similarities in Bob's data on Facebook and her di.me contact Bob, di.me makes a recommendation that you merge Bob's profiles to be associated as the same person. This construct would mean that Alice knows Bob in two contexts: as a 'friend', but also in her 'business' profile. These recommendations allow Alice to structure her contacts in a more organised way, and facilitate the sharing of her information in a more efficient way depending on which of Bob's profiles she wishes to send the information to, as described above.

7) Exercise data subject access right

di.me provides an overview of services that users are connected to. Each service is described in di.me as a *ServiceAdapter*, which is described in Tables II and III. The most important descriptors here are the *SAdapter.Description* field – which provides the user with a description of what the service is intended to do, and what connecting the service willbring as a benefit – and the *SAdapterSetting* definitions. One critical instance could be a Privacy Statement document included as a mandatory link, to which the user must agree, and is enforced when a service connection is built. When the service adapter connects to a service which displays the terms and conditions during the authorization protocol (as OAuthservices do), di.me does not need to include this as a mandatory link, but for services without such protocols (as services using basic HTTP authentication), this inclusion is critical. The privacy statement document ensures that the user consents to the collection and processing of the clearly defined purposes of di.me, namely to:

- Exchange and share profiles, messages, and data
- Provide the user with full control over who gets access to which information
- Allow the user acces via internet or the Android applicaton'di.me mobile'
- Manage data from different user devices
- Enable to connect to information from other social networks (e.g. messages, liveposts, profiles, or contacts) and to update this information regularly
- Provide recommendations on data privacy and trust
- Analyse the situation of the user (e.g. to show which contacts are located nearby)

TABLE II. API ANNOTATION FOR SADAPTER (SERVICE ADAPTER)

Type	SAdapter	
	Field name	***Description***
URI	authUrl	URL at which credential exchange takes place
boolean	isConfigurable	Whether the service can be configured or not
List<SAdapterSetting>	settings	List of service-specific configuration settings
string	description	Description of the service
URI	userId	User associated with the service

TABLE III. API ANNOTATION FOR SADAPTERSETTING

Type	SAdapterSetting	
	Field name	***Description***
string	name	Description of what the setting is
enum	fieldtype	Possible values: boolean, string, password, account, link
boolean	mandatory	Whether the setting is required or not
<mixed>	Value	User-provided setting value

A similar mandatory Data Subject Access Link could provide information about where data requests can be sent. This is included by default for each service in the suggested di.me configuration files and labeled 'You can request your data from <link>'. This link allows the user to exercise his right to request and retrieve information about his personal data, when they have been transferred through the di.me system. In this way di.me facilitates the exercise of a cornerstone data protection right of the users. All of these important links are then displayed in an easily accessible form in di.me in the service overview screen.

di.me's own data can be exported through calling the /api/dime/rest/<user>/dump API call, which provides a copy all the data that di.me stores on <user>. This authenticated call is only accessible for the user himself.

III. Related Work

Due to the multi-disciplinary nature of this contribution, note that this article is a summary of three years of research and design activities within the di.me consortium, which is constituted by nine partners from different countries across Europe. The project considered requirements categories in order to balance research and development outcomes in a multilateral manner [11]. The cited literature in previous sections reflects these outcomes throughout the project duration: trust, privacy and security were considered throughout the project, and in this order[13] for trust metrics and advisories [14][16][28][29], anonymity and secure communication [17][18][19], while considering unlinkability in the case of multiple identity support in a decentralized OSN [15]. The focus of this article, however, remains on how these numerous contributions are aligned with legal and ethical issues.

Building on results of projects such as PRIME[26], PrimeLife [14] and PICOS [15], incorporating leading privacy-oriented design methodology models such as privacy-by-design [9], and considering ethical perspectives expressed by contemporary media theorists [5][8][13],di.me demonstrates that a strategic privacy-oriented approach to social networking is feasible. di.metakes the data protection principles that are included in the European Data Protection Directive into account, and ensures the rights of the users.However, the pure technical consideration of technologies such as PETs is not enough to assess the consideration of all requirements from the legal and ethical points of view. There are many trade-offs (e.g. between privacy and context awareness) that could result in violations. For instance, since di.me supports multipleidentities, it was crucial to integrate unlinkability support in it. From a software engineering perspective, linkability as non-functional requirements (NFRs) may conflict with other competing NFRs such as providing context and collaboration awareness[16] at the user interface level, or negatively affecting user experience in terms of performance penalties by using anonymity networks.

Furthermore, there were many parties involved within the consortium and all requirements had to be considered from the legal point of view. For this, requirements negotation, elicitation, alignement, and priorization support necessarily occurred at process level. In order to address such complex cross-functional integration issues [30], the AFFINE methodology[17] [31] was followed within some workpackages in order to facilitate multi-lateral requirements cross-functional integration.Indeed, a complex analysis of all requirements by involving different partners with different goals and assessing thereby the correctness of design and implementation of agreed requirements can not be just solved by using various PETs (as demonstrated in [15] and solve in [18] and [19]). For instance, AFFINE enforces the earlier consideration of multilateral security requirements along with other (N)FRs also by involving all stakeholders, negotiating and aligning their potentially conflicting interests in the design [18] and development process, which meets our argumentation for privacy-by-design according to [7] and [8] in previous sections.[19]

IV. Conclusion

Introducing law and ethics as core values during the requirements and design phase of di.me resulted in a distributed OSN implementation that does not conflict with the EU right to privacy, and is also not contrary to exploitation interests of potential network operators. The resulting di.me prototype demonstrates that technology and regulation can work together effectively to support data access negotiation, and offers a protection mechanism for the less digitally-literate by presenting them with warning messages only in relevant scenarios, which make conscious and informed decisions concerning the potential repercussions of their interactions in and around OSNs. Although the prototype itself does not have the critical mass of users to become a replacement for current popular OSNs, it presents a concept that those OSNs could adopt, should they be required to.

Policy makers shape technology, and technology, in particular software, shapes user behavior. With American technology companies operating the vast majority of popular OSNs, the way European users of OSNs behave is slowly being shaped by this technological choice. But European policy makers can shape technology, and so requiring technology to

[13] The reader may excuse the emerging impression that the authors are citing their own work more than necessary. For accuracy, we cite these contributions since they represent sub-contributions in the involved research areas of security and privacy, data mining and linked data, usability engineering, etc.

[14] http://primelife.ercim.eu/

[15] http://www.picos-project.eu/

[16] Social, group, and workspace awareness answering 'who' is collaborating with 'whom', 'where', 'when', and 'why'.

[17] Agile Framework For Integrating Nonfunctional requirements Engineering is a Scrum-based method and the suggestion for supporting technology in form of a SOA/AOP layer towards earlier consideration of NFRs such as Privacy, Security, Trust and competing (N)FRs while building socio-technical systems such as di.me. AFFINE envisages involving experts or at least responsible(s) from each NFR category of relevance, e.g., legal and ethical concerns in order to ensure the right consideration from the beginning in the design and implementation of the respective system also at architectural level. TheAFFINE methodology is being now embraced by the company MT AG for the Integration Services business line.

[18] The solution's design process considers an attacker model and threat analysis.

[19] Santen began motivating his work by citing from Viega and McGraw (2001), who stated, "Bolting security onto an existing system is simply a bad idea. Security is not a feature you can add to a system at any time". He further argues, "the discipline of "Security Engineering" is far from mature today, and that, in practice, it still is not an integral part of the engineering processes for IT systems and software is based on the fact that security awareness results from reports on attacks – and not from the latest security feature that would make an application even more secure than it already was before."

implement technical support enabling protection of personal privacies would allow Europeans to continue valuing their right to privacy, even in the digital world, while allowing innovation in data brokerage and consensual, ethical commercialization of personal data.

Rising service-oriented broker platforms should consider law and ethics as core values during design phase, and in particular the concept of privacies, and existing OSNs should be required adopt these values if they wish to continue operating in the European market. Technically, such an adaptation could build upon the concepts of users having multiple context-specific digital identities, each of which serves for a particular purpose, and managing contextual information release. This could create a data exchange framework that respects law, ethics, and privacy without sacrificing commercial interest in data exploitation.

Di.me allows users to share personal information to other users and to other networks while providing the user with additional protection of their data, in particular by warning the user about the consequences of their actions if they have potentially unintended consequences. This protection is secure and allows the user to maintain control of his data, as the personal data is stored on the user's personal server – which could even be the user's laptop – and thus within the user's control. With a sizable percentage of the European population not being digitally literate, this approach could be important in enabling citizens to make informed decisions on exercising their right to protection and privacy of their personal data online.

Currently, the Directive is under review and may be replaced by a Regulation. One of the proposed changes is the strengthening of the principles of privacy-by-design by default and the promotion of data protection certification schemes. Moreover, standardisation initiatives will need to be promoted. Standardisation initiatives to ensure that social networking platform implementations are consistent with the revised data protection directive may be an interesting topic to investigate. di.me's APIs could contribute a basis for a privacy-oriented standardization intiative for cross-platform information brokerage of personal data. Further, the standardization mechanism could include a best-practice model for privacy-oriented design in social networking, to which di.me's approach could also serve as a foundational basis.

ACKNOWLEDGMENT

Even though just a few authors from the digital.me consortium contributed to this paper, it was surely not possible to compile such an article without the background efforts of all members ofthe di.me project. Special thanks are especially due to Fabian Hermann, Simon Thiel, Marcel Heupel, Ismael Rivera, BenediktWestermann, Lars Fischer as well as Philipp Schwarte.

REFERENCES

[1] J. Shim, M. Warkentin, J. Courtney, D. Power, R. Sharda, and C. Carlsson, "Past, present and future of decision support technology." Decision Support Systems, Vol. 33, Iss. 2, June 2002, pp. 111-126.

[2] L. Determann, "Social Media Privacy: A Dozen Myths and Facts," 2012 Stan. Tech. L. Rev. 7, pp. 1-14. [online] http://stlr.stanford.edu/pdf/determann-socialmediaprivacy.pdf

[3] "Conversion Rate Optimization." Blog run by Unbounce Marketing Solutions Inc. [online] http://unbounce.com/conversion-rate-optimization/

[4] I. Borges and D. Sinclair, "Media literacy, digital exclusion and older people." Brussels: AGE Platform Europe, December 2008. [online] http://www.age-platform.eu/images/stories/EN/pdf_AGE-media-A4-final-2.pdf

[5] S. Wheeler, "Digital literacies for engagement in emerging online cultures." Communication and Learning in the Digital Age, Barcelona: eLCRPS, Issue 5, pp. 14-25, November 2012.

[6] Articles 7 and 8, Charter of Fundamental Rights of the European Union. 2010/C 83/02.

[7] G. Santucci, "Privacy in the Digital Economy." The Privacy Surgeon, September 2013, p. 11 [online] http://www.privacysurgeon.org/blog/wp-content/uploads/2013/09/Privacy-in-the-Digital-Economy-final.pdf

[8] L. Lessig, Code and Other Laws of Cyberspace. New York: Basic Books, 1999.

[9] A. Cavoukian, "Privacy by Design … take the challenge." Toronto: Information and Privacy Commissioner of Ontario, 2009.

[10] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector.

[11] S. Thiel et al, "A requirements-driven approach towards decentralized social networks." Future Information Technology, Application, and Service Lecture Notes in Electrical Engineering. Vol. 164, Part 6, 2012. pp. 709-718.

[12] J. Stroh, Untitled post. Visual Metaphors community, 22 April 2013 [online] https://plus.google.com/100641053530204604051/posts/HDAVJBYBoSp

[13] O. Berg, "The Digital Workplace concretized". The Content Economy, 28 September 2012 [online] http://www.thecontenteconomy.com/2012/09/the-digital-workplace-concretized.html

[14] M. Bourimi, I. Rivera, M. Heupel, K. Cortis, S. Scerri, and S. Thiel, Simon, "Integrating multi-source user data to enhance privacy in social interaction," Proceedings of the 13th International Conference on Interacción Persona-Ordenado (INTERACCION 2012), art.51., New York: ACM, 2012, pp. 51-58.

[15] S. Thiel, F. Hermann, M. Heupel, and M. Bourimi, "Unlinkability Support in a Decentralised, Multiple-identity Social Network." To appear in the Proceedings of the Open Identity Summit 2013. Kloster Banz, Germany.

[16] M. Heupel et al, "Context-aware, trust-based access control for the digital.me userware," Proceedings of the 5th International Conference on New Technologies, Mobility and Security (NTMS) 2012.

[17] M. Bourimi, et al, "Towards transparent anonymity for user-controlled servers supporting collaborative scenarios," 9th International Conference on Information Technology: New Generations (ITNG), 2012. Pp. 102-108, April 2012.

[18] L. Fischer, M. Heupel, M. Bourimi, D. Kesdogan and R. Gimenez, "Enhancing Privacy in Collaborative Scenarios Utilising a Flexible Proxy Layer," Proceedings of the International Conference on Future Generation Communications 2012. London, UK.

[19] P. Schwarte et al, "Multilaterally secure communication anonymity in decentralized social networking," to appear in IEEE Xplore as part of the Proceedings of the 10th International Conference on Information Technology: New Generations (ITNG 2013).

[20] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "Techniques for the Identification of Semantically-Equivalent Online Identities." 8194, LNCS, 2013.

[21] B. Gorriz and S. Thiel, "Package Structure," 17 October 2013. [online] https://github.com/dime-project/meta/wiki/Package-Structure

[22] European Parliament and the Council of the European Union, Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31 (23.11.1995).

[23] J. Dumortier, and C. Goemans. ‘Privacy protection and identity management’, in B. Blažičand W. Schneider (Eds.) Security and Privacy in Advanced Networking Technologies, Ios Press, 2004, p. 193.

[24] European Commission, Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) COM (2012) 11 final – 2012/0011 (COD), 25 January 2012, commonly known as ‘draft Data Protection Regulation’.

[25] E. Kosta et al, "Requirements for Privacy Enhancing Tools." PRIME Consortium, 20 March 2008 [online] https://www.prime-project.eu/prime_products/reports/reqs/pub_del_D1.1.d_final.pdf

[26] J. Camenisch, R. Leenes, and D. Sommer (eds), Digital Privacy: Privacy and Identity Management for Europe (PRIME). Springer, 2011.

[27] J. Camenisch and E. Van Herreweghen, "Design and implementation of the idemix anonymous credential system," Proceedings of the 9th ACM conference on Computer and communications security (CCS 2002), New York: ACM, 2002, pp. 21-30.

[28] M. Heupel, M. Bourimi and D. Kesdogan, "Trust and Privacy in the di.me Userware," to appear in Kurosu, M. (ed) Human-Computer Interaction, Part III, HCII 2013. LNCS, vol. 8006. Heidelberg: Springer, 2013, pp. 29-38.

[29] M. Heupel, S. Scerri, M. Bourimi and D. Kesdogan, "Privacy-preserving concepts for supporting recommendations in decentralized OSNs," Proceedings of the 4th International Workshop on Modeling Social Media in conjunction with ACM Hypertext. Paris, 2013.

[30] A. Botzenhardt, H. Meth and A. Mädche, "Cross-functional Integration of Product Management and Product Design in Application Software Development: Exploration of Success Factors," Proceedings of the International Conference on Information Systems (ICIS) 2011. Paper 10.

[31] M. Bourimi, T. Barth, J. M. Haake, B. Ueberschär and D. Kesdogan, "AFFINE for Enforcing Earlier Consideration of NFRs and Human Factors when Building Socio-Technical Systems Following Agile Methodologies," Proceedings of the 3td Conference on Human-Centred Software Engineering (HCSE) 2010.

pGBbBShift

Perceptual Generalized Bitplane-by-Bitplane Shift

Jaime Moreno
Superior School of Mechanical and Electrical Engineers,
National Polytechnic Institute of Mexico,
IPN Avenue, Lindavista, Mexico City, 07738, Mexico.

***Abstract*—The paper we present pGBbBShift. This algorithm permits to code any Region of Interest (ROI) in a perceptual way, i.e. the presented algorithm introduces some characteristics of the HumanVisual System. Furthermore, it introduces features of chromatic induction to the GBbBShift method when bitplanes of ROI and background areas are coded. Thus, the included features balance visual importance of some pixels regardless their numerical importance, namely we avoid to use Information Theory criteria. Visual criteria are applied using the CIWaM, which is contrast band-pass filter that predicts color perception. pGBbBShift is compared against classicalROI algorithms, such as the MaxShift method of JPEG2000 and results show that there is no perceptual difference. pGBbBShift method is an open algorithm that can be applied in any wavelet based image coder such as JPEG2000, SPIHT or SPECK. Finally, we applied pGBbBShift to Hi-SET coder and we obtain the best results when the overall visual image quality is assessed**

Keywords*—Image Coding; JPEG2000; Hi-SET; region of interest(ROI); bitplane coding; wavelet coding; maximum shift (MaxShift);bitplane-by-bitplane shift (BbBShift); generalized bitplane-by-bitplane shift (GBbBShift)

I. Introduction

A. JPEG2000ROI Coding

Region of interest (ROI) image coding is a feature that modern image coders have, which allows to encode an specific region with better quality than the rest of the image or background (BG). ROI coding is one of the requirements in the JPEG2000 image coding standard [1], [2], which defines twoROI methods[3], [4]:

1) *Based on general scaling [3]*
2) *Maximum shift(MaxShift)[4]*

The general ROI scaling-based method scales coefficients in such a way that the bits associated with the ROI are shifted to higher bitplanes than the bitplanes associated with the background, as shown in Figure 1(b). It implies that during a embedded coding process, any background bitplane of the image is located after the most significant ROI bitplanes into the bitstream. But, in some cases, depending on the scaling value, φ, some bits of ROI are simultaneously encoded with BG. Therefore, this method allows to decode and refine the ROI before the rest of the image. No matterφ, it is posible to reconstruct with the entire bitstream a highest fidelity versionof the whole image. Nevertheless, If the bitstream is terminated abruptly, theROI will havea higher fidelity than BG.

The scaling-based method is implemented in five steps:.

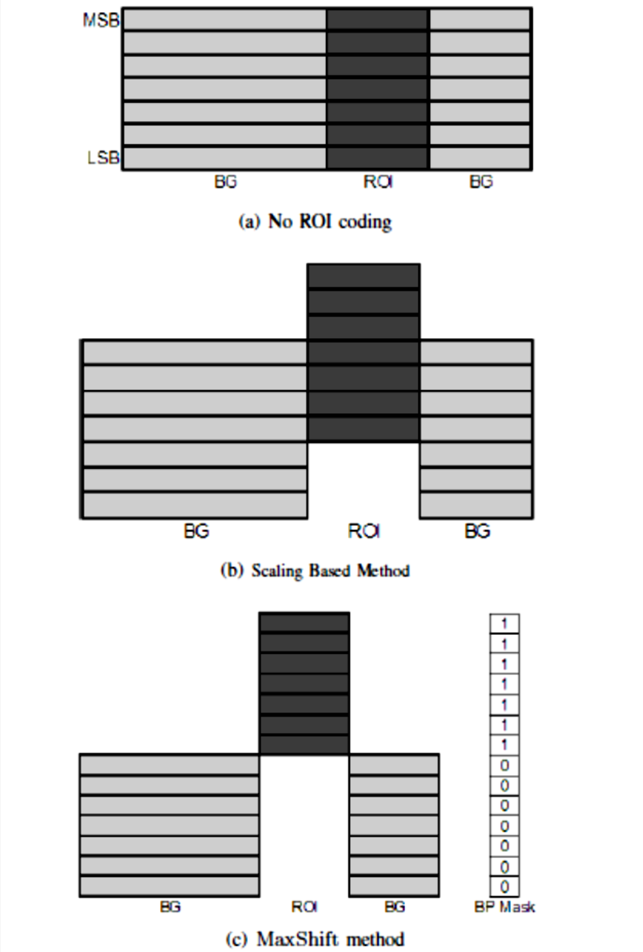
1) A wavelet transform of the original images is performed.

2) AROI mask is defined, indicating the set of coefficients that are necessary for reaching a lossless.

3) Wavelet coefficients are quantized and stored in a sign magnitude representation, using the most significant part of the precision. It will allow to downscale BG coefficients.

4) A specified scaling value, φ', downscales the coefficientsinside the BG.

5) The most significant bitplanes are progressively entropyencoded.



(a) No ROI coding

(b) Scaling Based Method

(c) MaxShift method

Fig. 1. JPEG2000 ROI Coding. (a) No ROI coding, (b) Scaling based ROI coding method ($\varphi = 3$) and (c) MaxShift method, $\varphi = 7$. Background is denoted as BG, Region of Interest as ROI and Bitplane mask as BP_{mask}. MSB is the most significant bitplane and LSB is the least significant bitplane.

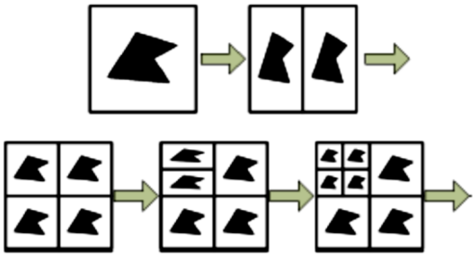


Fig. 2. ROI mask generation, wavelet domain.

The inputofROI scaling-based methodis the scalingvalue φ, while MaxShift method calculates it. Hence, the encoder defines from quantized coefficients this scaling value such that:

$$\varphi = \lceil \log_2 (\max \{\mathcal{M}_{BG}\} + 1) \rceil \quad (1)$$

where max $\{\mathcal{M}_{BG}\}$ is the maximum coefficient in the BG. Thus, when ROI is scaled up φ bitplanes, the minimum coefficient belonging to ROI will be place one bitplane up of BG (Fig. 1(c)). Namely, 2^{φ} is the smallest integer that is greater than any coefficient in the BG. MaxShift method is shown in Figure 1(c). Bitplane mask (BPmask) will be explained in section II-B.

At the decoder side, the ROI and BG coefficients aresimply identified by checking the coefficient magnitudes. Allcoefficients that are higher or equal than the φthbitplanebelong to the ROI otherwise they are a part of BG. Hence, itis not important to transmit the shape information of the ROIor ROIs to the decoder. The ROI coefficients are scaled downφ bitplanes before inverse wavelet transformation is applied.

B. *Perceptual Coding*

1) Chromatic Induction Wavelet Model: In order to generate an approximation to how every pixel is perceived from a certain distance taking into account the value of its neighboring pixels the Chromatic Induction Wavelet Model (CIWaM) is used. CIWaM attenuates the details that the human visual system is not able to perceive, enhances those that are perceptually relevant and produces an approximation of the image that the brain visual cortex perceives. CIWaM takes an input image I and decomposes it into a set of wavelet planes $\omega_{s,o}$ of different spatial scales s (i.e., spatial frequency ν) and spatial orientations o. It is described as:

$$\mathcal{I} = \sum_{s=1}^{n} \sum_{o=v,h,dgl} \omega_{s,o} + c_n \ , \quad (2)$$

where n is the number of wavelet planes, c_n is the residualplane and o is the spatial orientation either vertical, horizontalor diagonal. The perceptual image I_ρ is recovered by weightingthese $\omega_{s,o}$ wavelet coefficients using the extended ContrastSensitivity Function (e-CSF), which considers spatial surround information (denoted by r), visual frequency (ν related tospatial frequency by observation distance) and observationdistance (d). Perceptual image I_ρ can be obtained by

$$\mathcal{I}_\rho = \sum_{s=1}^{n} \sum_{o=v,h,dgl} \alpha(\nu, r)\, \omega_{s,o} + c_n \ , \quad (3)$$

where α(ν, r) is the e-CSF weighting function that tries toreproduce some perceptual properties of the HVS. The termα(ν, r) $\omega_{s,o} \equiv \omega_{s,o;\ \rho\ ,d}$ can be considered the perceptualwavelet coefficients of image I when observed at distance d.For details on the CIWaM and the α(ν, r) function, see [5].

2) Quantization: We employ the perceptual quantizer(ρSQ) either forward (F-ρSQ) and inverse (I-ρSQ. Quantization is the only causethat introduces distortion into a compression process. Eachtransform sample at the perceptual image Iρ (from Eq. 3) ismapped independently to a corresponding step size either Δsor Δn, thus Iρ is associated with a specific interval on the realline. Then, the perceptually quantized coefficients Q (F-ρSQ),from a known viewing distance d, are calculated as follows:

$$\mathcal{Q} = \sum_{s=1}^{n} \sum_{o=v,h,d} sign(\omega_{s,o}) \left\lfloor \frac{|\alpha(\nu, r) \cdot \omega_{s,o}|}{\Delta_s} \right\rfloor + \left\lfloor \frac{c_n}{\Delta_n} \right\rfloor \quad (4)$$

The perceptual inverse quantizer (I-ρSQ) or the recoveredα’(ν, r) introduces perceptual criteria to the classical InverseScalar Quantizer and is given by:

$$\hat{\mathcal{I}} = \begin{cases} \sum_{s=1}^{n} \sum_{o=v,h,d} sign(\widehat{\omega_{s,o}}) \dfrac{\Delta_s \cdot (|\widehat{\omega_s^o}| + \delta)}{\widehat{\alpha}(\nu, r)} & \\ & |\widehat{\omega_{s,o}}| > 0 \\ + (|\widehat{c_n}| + \delta) \cdot \Delta_n \ , & \\ 0, & \widehat{\omega_{s,o}} = 0 \end{cases} \quad (5)$$

II. Related Work

A. *BbBShift*

Wang and Bovik proposed the bitplane-by-bitplane shift(BbBShift) method in [6]. BbBShift shifts bitplanes on abitplane-by-bitplane strategy. Figure 3(a) shows an illustrationof the BbBShift method. BbBShift uses two parameters, $\varphi_1$and φ_2, whose sum is equal to the number of bitplanes forrepresenting any coefficient inside the image, indexing thetop bitplane as bitplane 1. Summarizing, the BbBShift methodencodes the first φ1 bitplanes with ROI coefficients, then, BGand ROI bitplanes are alternately shifted, refining graduallyboth ROI and BG of the image (Fig. 3(a)). The encodingprocess of the BbBShift method is defined as:

1) For a given bitplane bpl with at least one ROI coefficient:

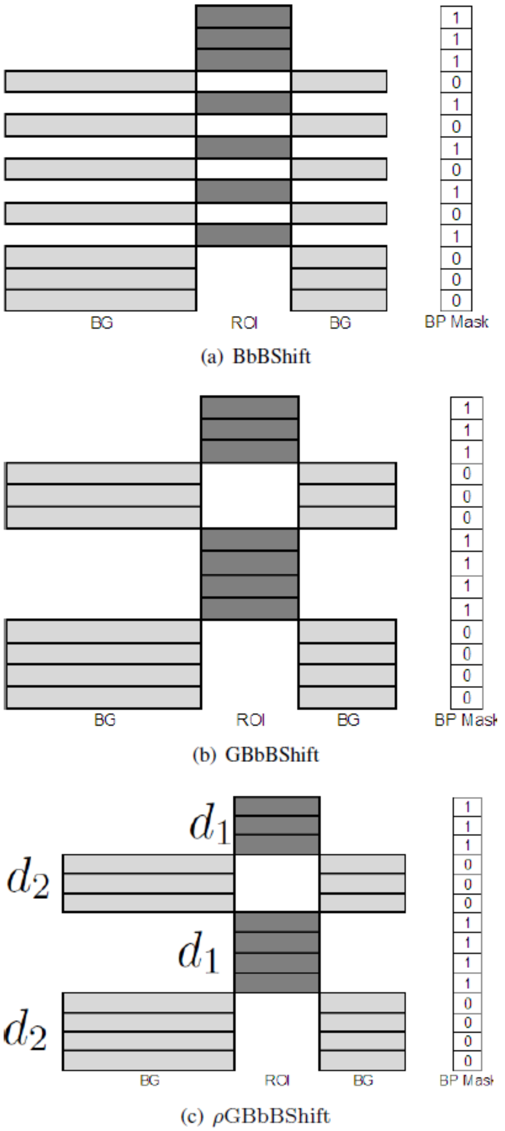
- If $bpl \le \varphi_1$, bpl is not shifted.
- If $\varphi_1 < bpl \le \varphi_1 + \varphi_2$, bpl is shifted down to $\varphi_1 + 2\,(bpl - \varphi_1)$

2) For a given bitplane bpl with at least one BG coefficient:

- If $bpl \le \varphi_2$, bpl is shifted down to $\varphi_1 + 2bpl - 1$
- If $bpl > \varphi_2$, bpl is shifted down to $\varphi_1 + \varphi_2 + bpl$

B. GBbBShift

In practice, the quality refinement pattern of the ROI andBG used by BbBShift method is similar to the general scalingbased method. Thus, when the image is encoded and thisprocess is truncated in a specific point the quality of the ROIis high while there is no information of BG.



(a) BbBShift

(b) GBbBShift

(c) ρGBbBShift

Fig. 3. ROI Coding Methods. (a) BbBShift, $\varphi_1 = 3$ and $\varphi_2 = 4$, (b) GBbBShift and (c)ρGBbBShift. Background is denoted as BG (For ρGBbBShift method is perceptually quantized by ρSQ at d_2), Region of Interest as ROI (For ρGBbBShift method is perceptually quantized at d_1 by ρSQ) and Bitplane mask as BP_{mask}.

Hence, Wang and Bovik [7] modified BbBShift method andproposed the generalized bitplane-by-bitplane shift (GBbBShift)method, which introduces the option to improve visualquality either of ROI or BG or both. Figure 3(b) shows thatwith GBbBShift method it is posible to decode some bitplanesof BG after the decoding of same ROI bitplanes. It allowsto improve the overall quality of the recovered image. Thisis posible gathering BG bitplanes.

Thus, when the encoding process achieves the lowest bitplanes of ROI, the quality of BGcould be good enough in order to portray an approximationof BG.

Therefore, the main feature of GBbBShift is to give theopportunity to arbitrary chose the order of bitplane decoding,grouping them in ROI bitplanes and BG bitplanes. This isposible using a binary bitplane mask or BPmask, whichcontains one bit per each bitplane, that is, twice the amount ofbitplanes of the original image. A ROI bitplane is representedby 1, while a BG bitplane by 0.

For example, the BPmask forMaxShift method in Figure 1(c) is 11111110000000, whilefor BbBShift in Figure 3(a) and GBbBShift in Figure 3(b) are11101010101000 and 11100011110000, respectively.

At the encoder side, the BPmask has the order of shiftingboth the ROI and BG bitplanes. Furthermore, BPmask isencoded in the bitstream, while the scaling values φ or φ_1 andφ_2 from the MaxShift and BbBShift methods, respectively,have to be transmitted.

III. ρGBbBShift Method

In order to have several kinds of options for bitplane scalingtechniques, a perceptual generalized bitplane-by-bitplaneshift(ρGBbBShift) method is proposed. The ρGBbBShiftmethod introduces to the GBbBShift method perceptual criteriawhen bitplanes of ROI and BG areas are shifted.

This additionalfeature is intended for balancing perceptual importanceof some coefficients regardless their numerical importance andfor not observing visual difference at ROI regarding MaxShiftmethod, improving perceptual quality of the entire image.

Thus, ρGBbBShift uses a binary bitplane mask or BPmaskin the same way that GBbBShift (Figure 3(c)). At the encoder,shifting scheme is as follows:

1) Calculate φ using Equation 1.

2) Verify that the length of BP_{mask} is equal to 2φ.

3)
- **For all ROI Coefficients, forward perceptual quantize them using Equation 4 (F-ρSQ) with viewing distance d_1.**
- **For all BG Coefficients, forward perceptual quantize them using Equation 4 (F-ρSQ) with viewing distance d_2, being $d_2 \gg d_1$.**

4) Let τ and η be equal to 0.

5) For every element i of BP_{mask}, starting with the least significant bit:
- **If $BP_{mask}(i) = 1$, Shift up all ROI perceptual quantized coefficients of the $(\varphi - \eta)$-th bitplane by τ bitplanes and increment η.**
- **Else: Shift up all BG perceptual quantized coefficients of the $(\varphi - \tau)$-th bitplane by η bitplanes and increment τ.**

At the decoder, shifting scheme is as follows:

1) Let $\varphi = \frac{length\ of\ \bar{BP}_{mask}}{2}$ be calculated.
2) Let τ and η be equal to 0.
3) For every element i of BP_{mask}, starting with the least significant bit:
   - If $BP_{mask}(i) = 1$, Shift down all perceptual quantized coefficients by τ bitplanes, which pertain to the $(2\varphi - (\tau + \eta))$-th bitplane of the recovered image and increment η.
   - Else: Shift down all perceptual quantized coefficients by η bitplanes, which pertain to the $(2\varphi - (\tau + \eta))$-th bitplane of the recovered image and increment τ.
4) Let us denote as $c_{i,j}$ a given non-zero wavelet coefficient of the recovered image with 2φ bitplanes and $\overline{c}_{i,j}$ as a shifted down c obtained in the previous step, with φ bitplanes.
   - If $(c_{i,j}$ & $BP_{mask}) > 0$, inverse perceptual quantize $\overline{c}_{i,j}$ using Equation 5 (I-ρSQ) with d_1 as viewing distance.
   - If $(c_{i,j}$ & $BP_{mask}) = 0$, inverse perceptual quantize $\overline{c}_{i,j}$ using Equation 5 (I-ρSQ) with d_2 as viewing distance.

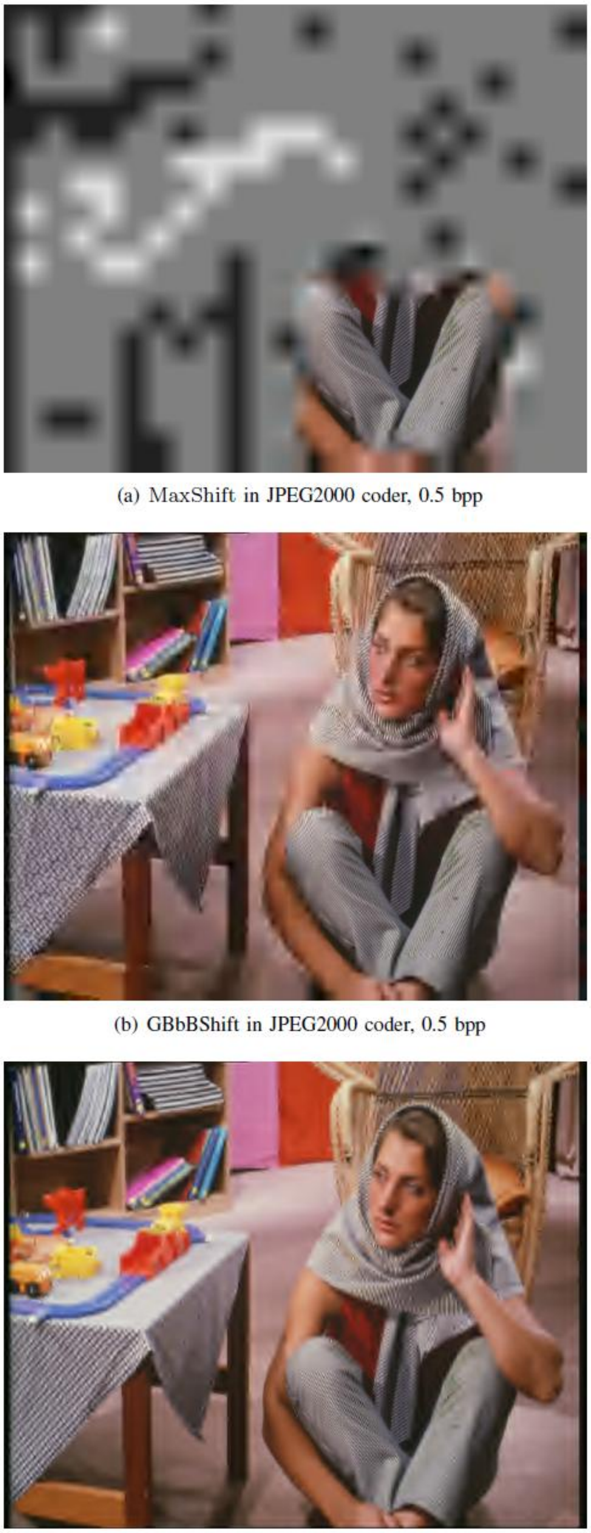
IV. EXPERIMENTAL RESULTS

The ρGBbBShift method, as the other methods presentedhere, can be applied to many image compression algorithmssuch as JPEG2000 or Hi-SET. We test our methodapplying it to Hi-SET and the results are contrasted withMaxShift method in JPEG2000 and Hi-SET. The setupparameters are φ = 8 for MaxShift and BPmask =1111000110110000, d1 = 5H and d2 = 50H, where H ispicture height (512 pixels) in a 19-inch LCD monitor, forρGBbBShift. Also, we use the JJ2000 implementation whenan image is compressed by JPEG2000 standard[8].

A. Application in well-known Test Images

Figure 4 shows a comparison among methods MaxShift andGBbBShift applied to JPEG2000, in addition to, ρGBbBShiftapplied to Hi-SET. The 24-bpp image Barbara is compressedat 0.5 bpp. It can be observed that without visual differenceat ROI, the ρGBbBShift method provide better imagequality at the BG than the general based methods defined inJPEG2000 Part II[1].

In order to better qualify the performanceof MaxShift, GBbBShift and ρGBbBShift methods, first,we compared these methods applied to the Hi-SET coderand then, we compare MaxShift and ρGBbBShift methodsapplied to the JPEG2000 standard and Hi-SET, respectively.



(a) MaxShift in JPEG2000 coder, 0.5 bpp

(b) GBbBShift in JPEG2000 coder, 0.5 bpp

(c) ρGBbBShift in Hi-SET coder, 0.5 bpp

Fig. 4. 512 × 640 pixel Image *Barbara* with 24 bits per pixel. ROI is a patch of the image located at [341 280 442 442], whose size is 1/16 of the image. Decoded images at 0.5 bpp using MaxShift method in JPEG2000 coder((a) $\varphi = 8$), GBbBShift method in JPEG2000 coder ((b)$BP_{mask} = 1111000110110000$) and ρGBbBShift method in Hi-SET coder ((c)$BP_{mask} = 1111000110110000$).

We compress two different gray-scale and color images ofLenna at different bit-rates. ROI area is a patch at the centerof these images, whose size is 1/16 of the image. We employthe perceptual quality assessment CwPSNR, which weights the mainstream PSNR by meansof a chromatic induction model.

Figs. 5(a) and 5(b) show the comparison among MaxShift(Blue Function), GBbBShift(Green Function) and ρGBbBShift(Red Function) methods applied to Hi-SET coder. 512 × 512 pixel Image Lenna for gray-scale is employ for this experiment. These Figures also show that the ρGBbBShift method gets the better results both in PSNR(objective image quality, Fig. 5(a)) and CwPSNR (subjective image quality, Fig. 5(b)) in contrast to MaxShift and GBbBShift methods. In addition, when MaxShift method applied to JPEG2000 coder and ρGBbBShift applied to Hi-SET coder are compared, ρGBbBShift obtains less objective quality (Fig. 5(c)), but better subjective quality for gray-scale images (Fig. 5(d)).

Figure 6 shows a visual example, when image Lenna is compressed at 0.34 bpp by JPEG2000 and Hi-SET. Thus, it can be observed that ρGBbBShift provides an important perceptual difference regarding the MaxShiftmethod(Fig. 6(d)). Furthermore, Figs. 6(b) and 6(c) show the examples when MaxShift and GBbBShift methods, respectively, are applied to the Hi-SET coder.

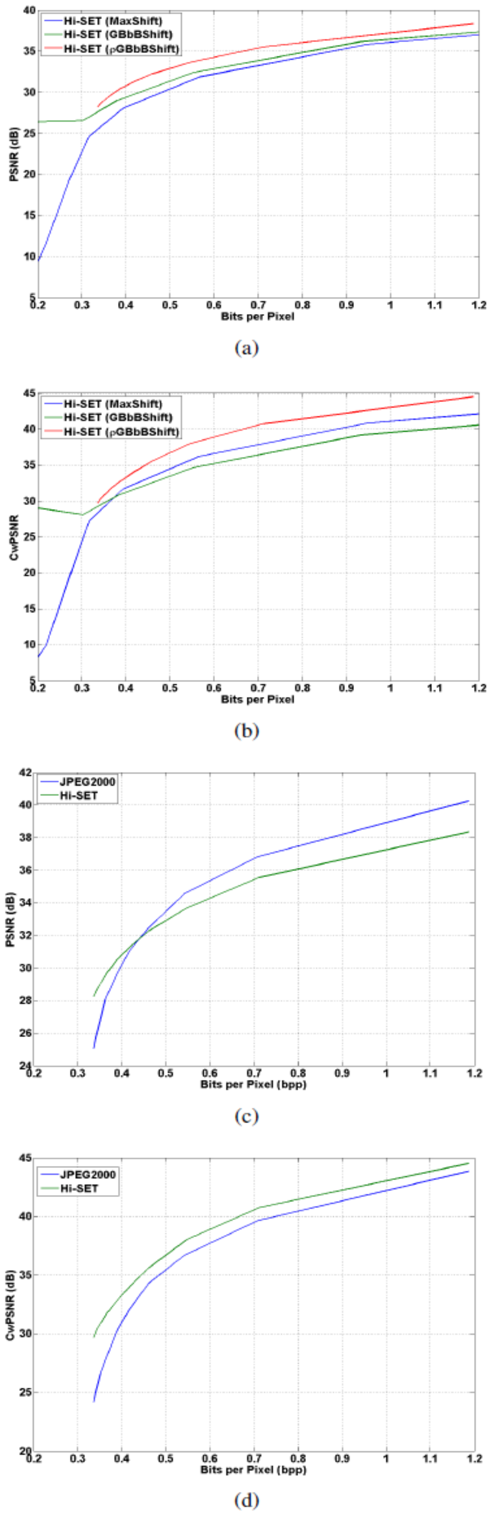
B. *Application in other image compression fields*

The usage of ROI coded images depends on an specificapplication, but in some fields such as manipulation andtransmission of images is important to enhance the imagequality of some areas and to reduce it in others[9], [10]. InTelemedicine or in Remote Sensing (RS) it is desirable tomaintain the best quality of the ROI area, preserving relevantinformation of BG, namely the most perceptual frequencies.

Figure 7 shows an example of the application of ROI inRemote Sensing. Image 2.1.05, from Volumen 2: aerials ofUSC-SIPI image database 8 bits per pixel[11], is compressedat 0.42 bpp. MaxShift method spends all the bit-ratio forcoding ROI, located at [159 260 384 460], while ρGBbBShiftbalances a perceptually lossless ROI area with an acceptablerepresentation of the BG.

Hence, the overall image qualitymeasured by PSNR in Figure 7(a) is 16.06 dB, while inFigure 7(b) is 24.28 dB. When perceptual metrics assess theimage quality of the ρGBbBShift coded image, for example,VIFP=0.4982, WSNR=24.8469 and CwPSNR=27.07, whilefor MaxShift coded image VIFP=0.2368, WSNR=11.33 andCwPSNR=16.72.

Thus, for this example, both PSNR and thesesubjective metrics reflect important perceptual differences betweenROI methods, being ρGBbBShift method better thanMaxShift method.



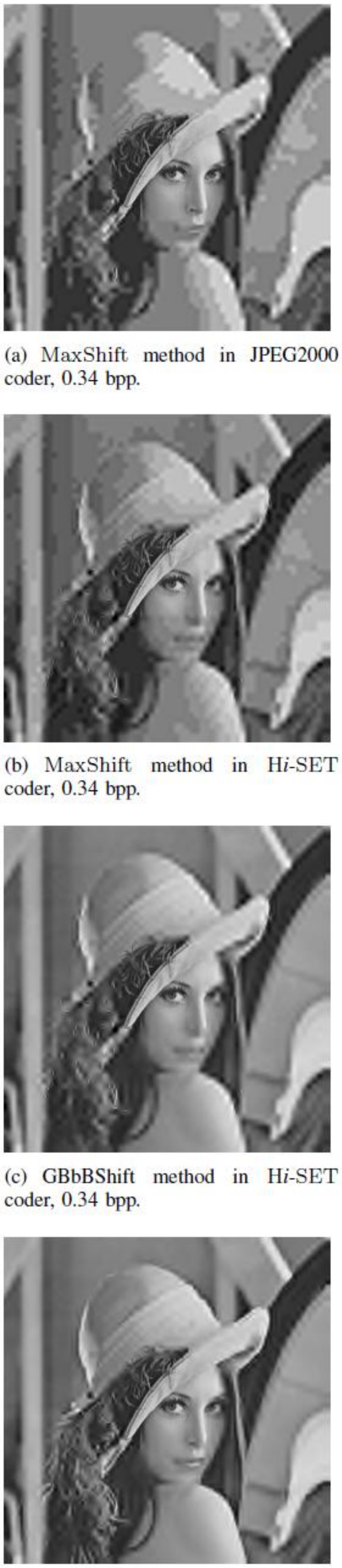
(a)

(b)

(c)

(d)

Fig. 5. (a-b) Comparison among MaxShift(Blue Function), GBbB-Shift(Green Function) and *ρ*GBbBShift(Red Function) methods applied to H*i*-SET coder. (c-d) Comparison between MaxShift method applied to JPEG2000 coder and *ρ*GBbBShift applied to H*i*-SET coder. 512 × 512 pixel Image *Lenna* with 8 bits per pixel is employed for this experiment. ROI is a patch at the center of the image, whose size is 1/16 of the image. The overall image quality of decoded images at different bits per pixel are contrasted both (a and c) objectively and (b and d)subjectively.



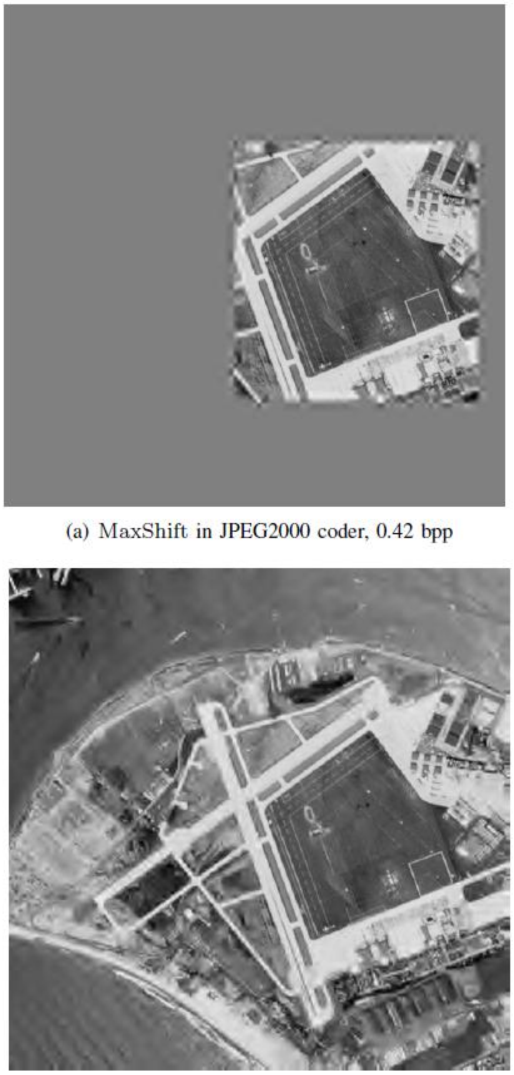
(a) MaxShift method in JPEG2000 coder, 0.34 bpp.

(b) MaxShift method in H*i*-SET coder, 0.34 bpp.

(c) GBbBShift method in H*i*-SET coder, 0.34 bpp.

(d) ρGBbBShift method in H*i*-SET coder, 0.34 bpp.

Fig. 6. 512 × 512 pixel Image *Lenna* from CMU image database with 8 bits per pixel. ROI is a patch at the center of the image, whose size is 1/16 of the image. Decoded images at 0.34 bpp using $\varphi = 8$ for MaxShift method (a) in JPEG2000 coder and (b) in H*i*-SET coder, and $BP_{mask} =$ 1111000110110000 for (c) GBbBShift and (d) ρGBbBShift methods in H*i*-SET coder.



(a) MaxShift in JPEG2000 coder, 0.42 bpp

(b) ρGBbBShift method in H*i*-SET coder, 0.42 bpp

Fig. 7. Example of a remote sensing application. 512 × 512 pixel Image *2.1.05* from *Volumen 2: aerials* of USC-SIPI image database at 8 bits per pixel. ROI is a patch with coordinates [159 260 384 460], whose size is 225 × 200 pixels. Decoded images at 0.42 bpp using MaxShift method ((a) $\varphi = 8$) in JPEG2000 coder and ρGBbBShift method ((b)$BP_{mask} =$ 1111000110110000) in H*i*-SET coder.

V. Conclusions

A perceptual implementation of the Region of Interest,ρGBbBShift(), is proposed, which is a generalized methodthat can be applied to any wavelet-based compressor. Weintroduced ρGBbBShift method to the Hi-SET coder and itvisually improves the results obtained by previous methodslike MaxShift and GBbBShift.

Our experiments show thatρGBbBShift into Hi-SET provides an important perceptualdifference regarding the MaxShift method into JPEG2000,when it is applied to conventional images like *Lenna* or*Barbara*.

Acknowledgment

This work is supported by The National Polytechnic Institute of Mexico by means of a granted fund by the

Committee of Operation and Promotion of Academic Activities (COFAA), in addition to the Project No. 20131312 granted by the Academic Secretary and National Council of Science and Technology of Mexico by means of Project No. 204151/2013.

REFERENCES

[1] M. Boliek, E. Majani, J. S. Houchin, J. Kasner, and M. Carlander, Information Technology: JPEG2000 Image Coding System (Extensions), JPEG 2000 Part II final committee draft ed., ISO/IEC JTC 1/SC 29/WG 1, Dec. 2000.

[2] M. Boliek, C. Christopoulos, and E. Majani, Information Technology: JPEG2000 Image Coding System, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.

[3] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, ser. ISBN: 0-7923-7519-X. Kluwer Academic Publishers, 2002.

[4] E. Atsumi and N. Farvardin, "Lossy/lossless region-of-interest image coding based on set partitioning in hierarchical trees," in International Conference on Image Processing, vol. 1, oct 1998, pp. 87 –91 vol.1. [5] X. Otazu, C. Párraga, and M. Vanrell, "Toward a unified chromatic induction model," Journal of Vision, vol. 10(12), no. 6, 2010.

[5] Z. Wang and A. C.Bovik, "Bitplane-by-bitplane shift (BbBShift) – a suggestion for JPEG2000 region of interest image coding," IEEE Signal Processing Letters, vol. 9, no. 5, pp. 160 – 162, May 2002.

[6] Z. Wang, S. Banerjee, B. L. Evans, and A. C. Bovik, "Generalized bitplane-by-bitplane shift method for JPEG2000 ROI coding," IEEE International Conference on Image Processing, vol. 3, pp. 81–84, September 22-25 2002.

[7] C. Research, ´ EcolePolytechniqueF´ed´erale de Lausanne, and Ericsson. (2001) JJ2000 implementation in Java, available at http://jj2000.epfl.ch/. annon Research, ´ EcolePolytechniqueF´ed´erale de Lausanne and Ericsson. [Online]. Available: http://jj2000.epfl.ch/

[8] J. Bartrina-Rapesta, F. Auli-Llinas, J. Serra-Sagrista, A. Zabala-Torres, X. Pons-Fernandez, and J. Maso-Pau, "Region of interest coding applied to map overlapping in geographic information systems," in IEEE International Geoscience and Remote Sensing Symposium, 23-28 2007, pp. 5001 –5004.

[9] J. Gonzalez-Conejero, J. Serra-Sagrista, C. Rubies-Feijoo, and L. Donoso-Bach, "Encoding of images containing no-data regions withinJPEG2000 framework," in 15th IEEE International Conference on ImageProcessing, 12-15 2008, pp. 1057 –1060.

[10] S. and Image Processing Institute of the University ofSouthern California. (1997) The USC-SIPI image database, availableat http://sipi.usc.edu/database/. Signal and Image Processing Instituteof the University of Southern California. [Online]. Available:http://sipi.usc.edu/database/

Personalized Real Time WeatherForecasting With Recommendations

A new era of weather forecasting

Abhishek Kumar Singh
Assistant Professor, CSE, BIT
Durg, C.G., India

Aditi Sharma
Assistant Professor,CSE, JIIT
Noida, India

Rahul Mishra
P.Hd. Scholar
IIIT-Delhi

Abstract—**Temperature forecasting and rain forecasting in today's environment is playing a major role in many fields like transportation, tour planning and agriculture. The purpose of this paper is to provide a real time forecasting to the user according to their current position and requirement.**

The simplest method of forecasting the weather, persistence, relies upon today's conditions to forecast the conditions tomorrow i.e. analyzing historical data for predicting future weather conditions. The weather data used for the DM research include daily temperature, daily pressure and monthly rainfall.

Keywords—Weather Forcasting; NOAA

I. Introduction

Forecasting the temperature and rain on a particular day and date is the main aim of this paper. In the paper we forecast rain and temperature for Europe; year up to 2051 and also we forecast temperature of world; year up to 2100.Our paper is aimed to provide real time weather forecast service at finest granularity level with recommendations. We grab user’s location (longitude, latitude) using GPS data service whenever user requests for our services. Our system will process the users query and will mine the data from our repository to draw appropriate results. Users will be provided with recommendations also and that is the key facility of our service. Personalized forecast is generated for each individual user based on their location.

II. Scopes

The project mainly focuses on forecasting weather conditions using historical data. This can be done by extracting knowledge from this given data by using techniques such as association, pattern recognition, nearest neighbor etc.

- Disaster Mitigation: Predicting storms, floods, droughts
- Helping those sectors which are most dependent on weather such as agriculture, aviation also depends on weather conditions.

III. Target Segment

Some target segments are following.

1) Our target users are mainly normal citizens they can use our services for their lots of benefits like:

a) Suppose a user is stuck on the way to home due to heavy rain then using our service, he will be able to know whether there is any another highway or route nearby where it’s not raining or less raining.

2) Using our service any individual can get weather information specially personalized for him irrespective of what is the time or place.

3) If our service is connected with THERMOSTATs of some house then temperature of the house can be controlled automatically using forecast information provided by us using location of house based on GPS.

4) In transportation industry our services can be used to take some important decisions like: Which route is better for transportation, where snow fall probability is quite low etc.

5) There are enormous more areas where our service will be helpful like tourism, food processing industry, Aviation Industry, Oil and natural gas exploration and production activities etc.

IV. PROCESS DETAILS:

The process we have briefed in earlier section can be depicted pictorially and which is self-explanatory.

We can divide our process in two modules namely:

1) *Weather Mining*
2) *Recommendation*

B. Weather Mining:

- Data collection: We have collected weather data from WORLD DATA CENTER for climate, Hamburg. We have decided to use NWS API for data collection in future.
- Data formatting and cleaning: We have converted our data from .NC (netcdf) format to .CSV (comma-separated values) format because WEKA supports .CSV format.
- Clustering: Using WEKA, we have performed clustering on weather data to draw inferences.
- Recommendation: We have planned to use recommendation algorithm as user to locationcollaborative algorithm similar to user to item collaborative algorithm. This algorithm uses user location (N*M) metrics.

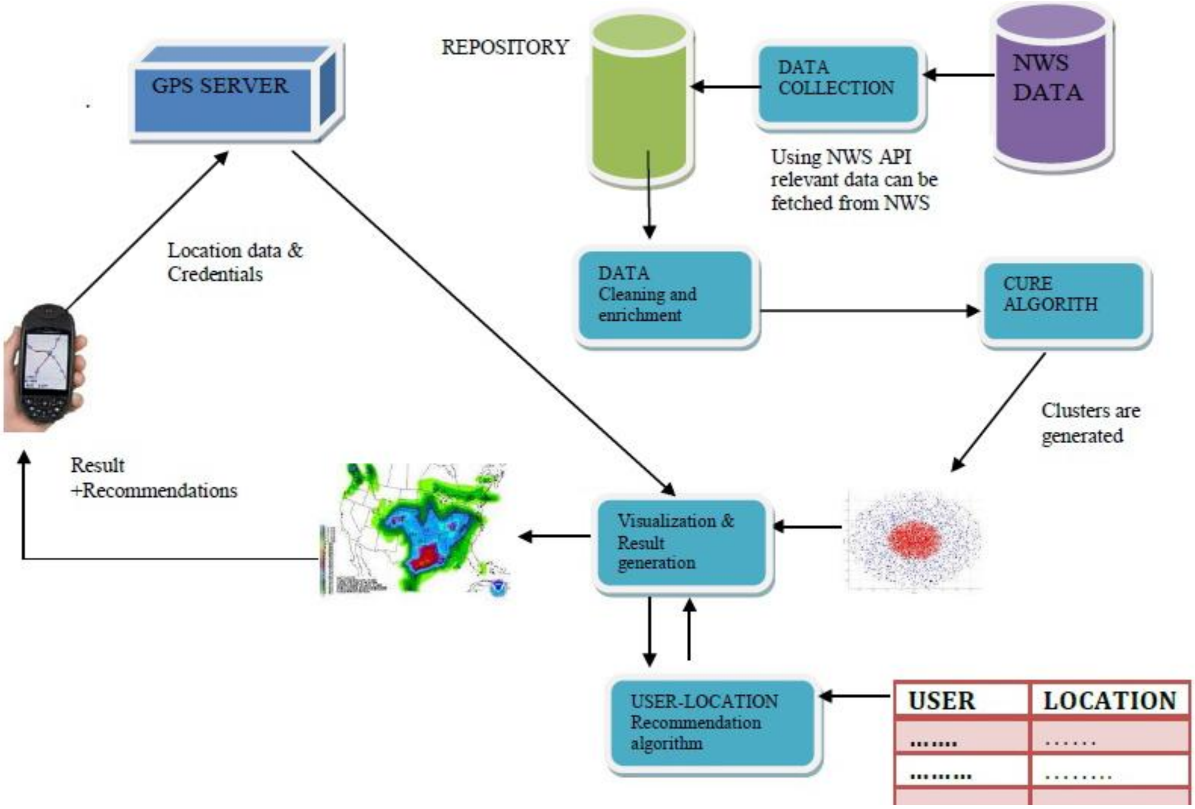

GPS SERVER
REPOSITORY
DATA COLLECTION
NWS DATA
Using NWS API relevant data can be fetched from NWS
Location data & Credentials
DATA Cleaning and enrichment
CURE ALGORITH
Clusters are generated
Result +Recommendations
Visualization & Result generation
USER-LOCATION Recommendation algorithm
USER
LOCATION
....
......
....
....

Fig. 1. Process details

- Visualization: To generate visualization for user, we have used NOAA weather and climate tool kit.

For the part of the implementation, on which your project focused most, which algorithms you implemented or used and if any modifications were needed to those algorithms or if you did some initial preprocessing, discuss here For the other phases of data mining, discuss briey. E.g., if you focused most on visualization, you can talk about: which data (Example: downloaded from some website put the URL here; did some survey, then talk about how you did the survey etc) collection approach was used in the project?

C. *Recommendation*

- Extract the location of the user
- Extract the destination of the user
- And then recommend the best path according to the conditions. (As shown in figure 1)

V. Technical Skills

We haven't implemented software for clustering and visualization our self, we have used WEKA and NOAA for this purpose. We have faced some problems to work with these tools and eventually solved them on our own.

VI. Some Snapshots:

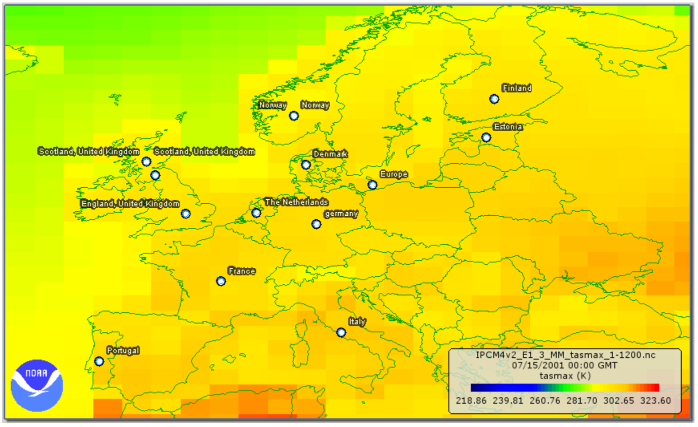


Fig. 2. Temperature Forecast for Europe in7/2001

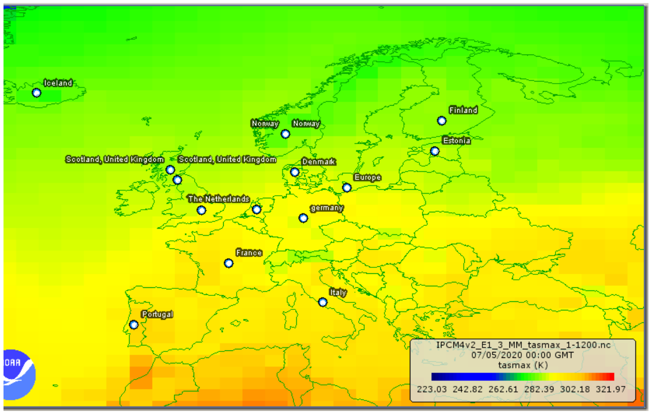


Fig. 3. Temperature Forecast for Europe in7/2020

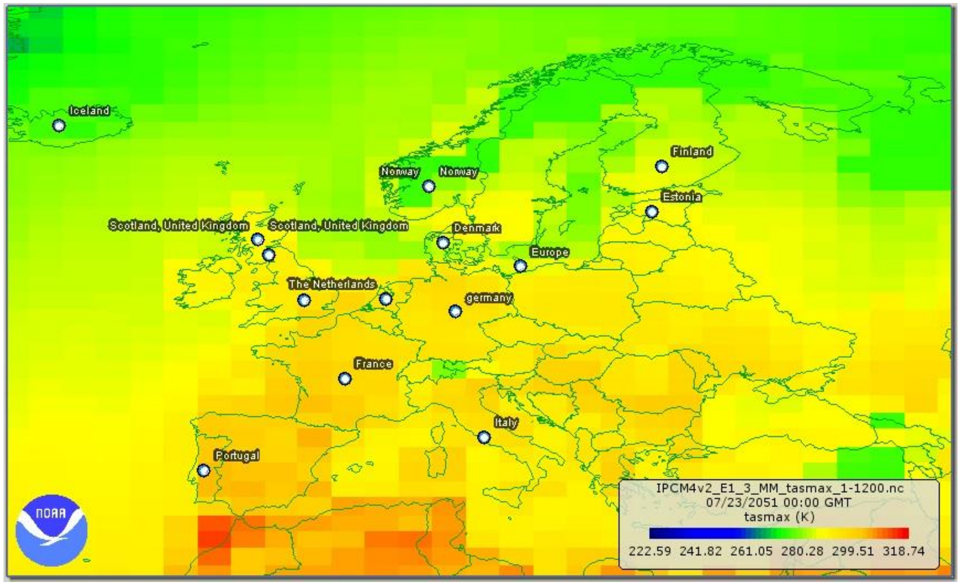


Fig. 4. Temperature Forecast for Europe in 7/2051

- WEKA was crashing initially for large data set.
  - ➢ We resolved this issue by increasing heap size of WEKA.
- Since WEKA doesn’t support netcdf format. We had to convert our data to .CSV format.
  - ➢ We have written code in C++ to convert data from .NC to .CSV.

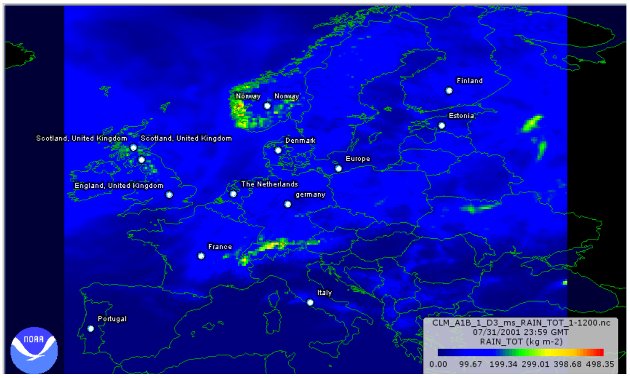
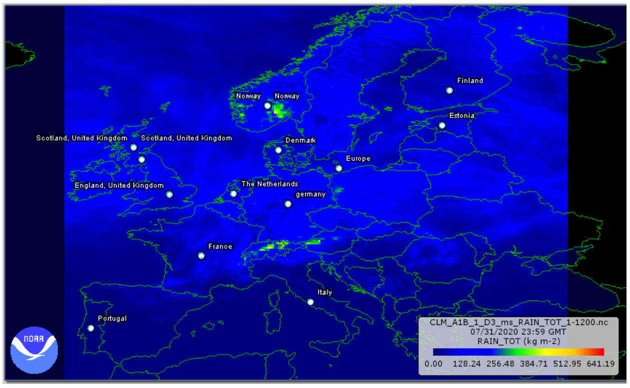


Fig. 5. Rain Forecast for Europe in 7/2001



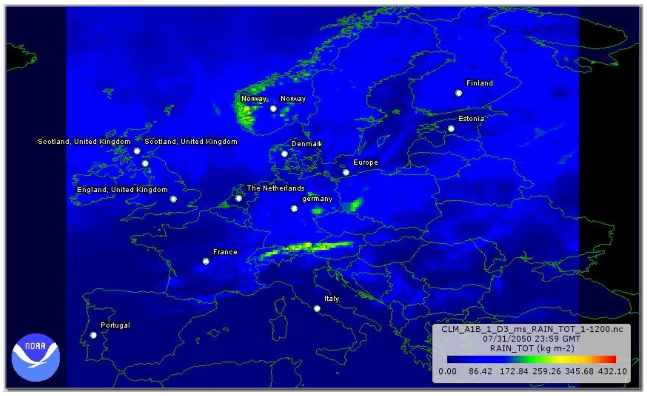


Fig. 7. Rain Forecast for Europe in 7/2050

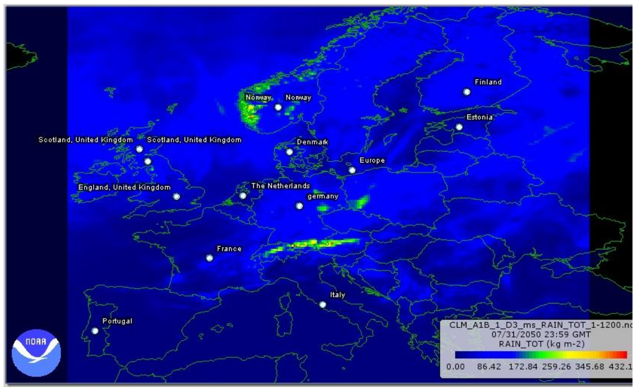


Fig. 8. Rain Forecast for Europe in 7/2100

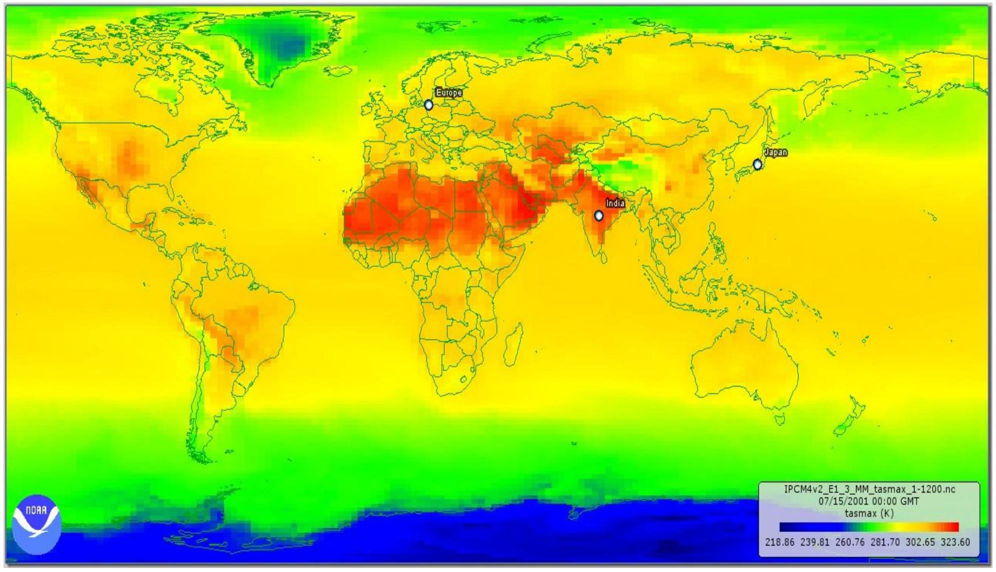


Fig. 9. Temperature Forecast for world in 7/2001

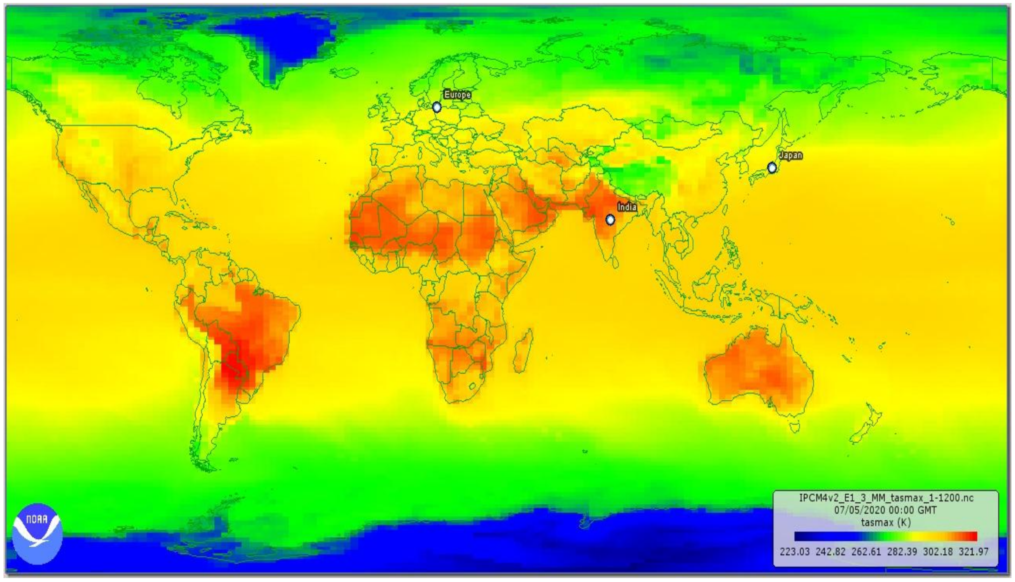


Fig. 10. Temperature Forecast for world in 7/2020

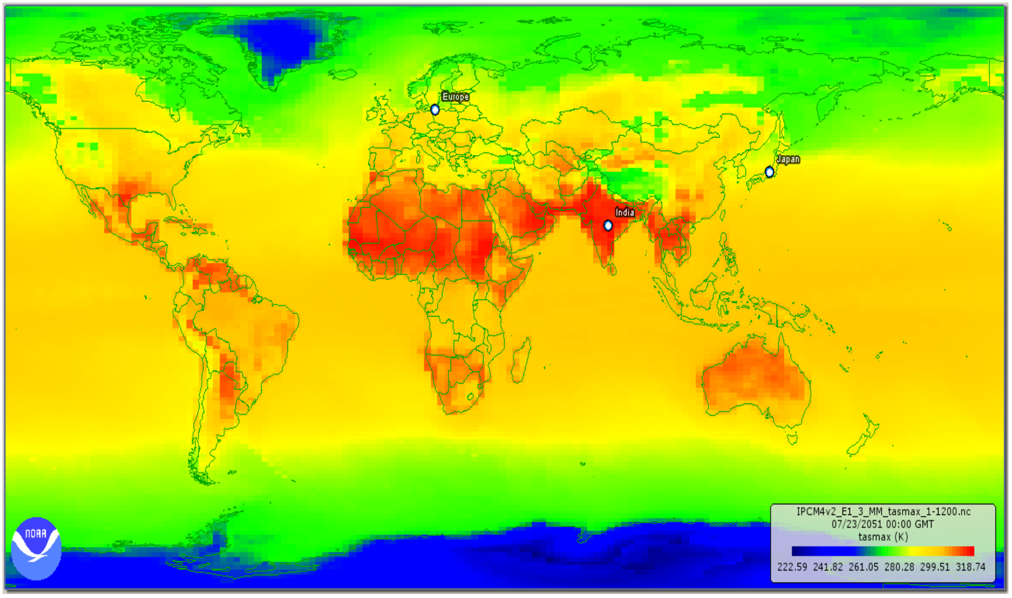


Fig. 11. Temperature Forecast for world in 7/2061

References

[1] Weka tool.

[2] NOAA tool.

[3] World Climate Data Center for Data

Software Effort Estimation Inspired by COCOMO and FP Models: A Fuzzy Logic Approach

Alaa F. Sheta and Sultan Aljahdali
Computer Science Department
College of Computers and Information Technology
Taif University
Taif, Saudi Arabia
asheta66@gmail.com, aljahdali@tu.edu.sa

***Abstract*—Budgeting, bidding and planning of software project effort, time and cost are essential elements of any software development process. Massive size and complexity of now a day produced software systems cause a substantial risk for the development process. Inadequate and inefficient information about the size and complexity results in an ambiguous estimates that cause many losses. Project managers cannot adequately provide good estimate for both the effort and time needed. Thus, no clear release day to the market can be defined. This paper presents two new models for software effort estimation using fuzzy logic. One model is developed based on the famous COnstructive COst Model (COCOMO) and utilizes the Source Line Of Code (SLOC) as input variable to estimate the Effort (E); while the second model utilize the Inputs, Outputs, Files, and User Inquiries to estimate the Function Point (FP). The proposed fuzzy models show better estimation capabilities compared to other reported models in the literature and better assist the project manager in computing the software required development effort. The validation results are carried out using Albrecht data set.**

I. Introduction

According to Dr. Patricia Sanders, Director of Test Systems Engineering and Evaluation at OUSD, in her 1998 Software Technology Conference keynote address, 40% of the DoD's software development costs are spent on reworking the software, which on the year 2000 equal to an annual loss of $18 billion. Furthermore, Sanders stated that only 16% of software development would finish on time and on budget.

The dimension and complication of computer based-systems grown noticeably during the past few decades [1]–[4] and the tendency will certainly continue in the future specially in Military Application, NASA Space Shuttle systems, Air Force and business for huge Enterprises. Some NASA and Air Force projects have estimated that the cost of software development could be up to 50% of their development cost. It was stated in [5]:

> Given that software-intensive projects are among the most expensive and risky undertakings of the 21st century, the investment in weapons from fiscal years 2003 through 2009 will exceed $1 trillion. Furthermore, many of the DoD's most important technology projects will continue to deliver less than promised unless changes are made. Improving how we acquire software-intensive systems is both long overdue and an imperative.

Although many research papers appears since 1960 providing numerous models to help in computing the effort/cost for software projects, being able to provide accurate effort/cost estimation is still a challenge for many reasons. They include: 1) the uncertainty in collected measurement, 2) the estimation methods used which might have many drawbacks and 3) the cost drivers which comes with various characteristics based on the methodology of development.

In this paper, we provide a detailed study on the algorithmic software effort estimation models. We provide our initial idea on using fuzzy models to build a Takagi Sugeno fuzzy model for the software effort. We developed two fuzzy models. They utilize both the SLOC and FP parameters. Our experimental results cover 24 software projects based Albrecht data set.

II. Literature Review

In 1994, Zadeh presented the definition of Soft Computing techniques [6]. He mentioned that soft computing is not a homogeneous body of concepts and techniques. At that time, the principal techniques which compose the domain of soft computing were fuzzy logic, neurocomputing, and probabilistic reasoning. Later on, the domain was expanded to cover techniques such as Genetic Algorithms (GAs), Swarm Intelligence (SI), Differential Evolution (DE) and many others. In the past, soft computing techniques were explored to build efficient effort estimation models structures [7], [8]. In [9], author explored the use of Neural Networks (NNs), GAs and Genetic Programming (GP) to provide a methodology for software cost estimation. Later authors in [10], provided a detailed study on using Genetic Programming (GP), Neural Network (NN) and Linear Regression (LR) in solving the software project estimation. Many data sets provided in [11], [12] were explored with promising results. In [13], authors provided a survey on the cost estimation models using artificial neural networks. Fuzzy logic and neural networks were used for software engineering project management in [14]. A fuzzy COCOMO model was developed in [7].

Recently, Soft Computing and Machine Learning Techniques were explored to handle many software engineering problems. They include the effort and cost estimation problems. In [15], author provided an innovative set of models modified from the famous COCOMO model with interesting results. Later on, many authors explored the same idea with some modification [16]–[18] and provided a comparison to the work presented in [15]. The idea of using Takagi Sugeno Fuzzy Logic was primary presented in [19] to see how a rule based system can solve the software effort estimation problem. Authors in [20] presented an extended work on the use of Soft Computing Techniques to build a suitable model structure to utilize improved estimations of software effort for NASA software projects. On doing this, Particle Swarm Optimization (PSO) was used to tune the parameters of the COCOMO model. A data set for NASA software projects [21] were used to test the developed models. Author provided a comparison between various software cost estimation models. They include COCOMO-PSO, Fuzzy Logic (FL), Halstead, Walston-Felix, Bailey-Basili and Doty models with excellent performance results.

III. COnstructive COst Model

Many software cost estimation models where proposed to help in providing a high quality estimate to assist project manager in taking best decisions for a project [22], [23]. COCOMO is one of a very famous software effort estimation models. COCOMO was introduced by Boehm in 1981 [22], [23]. This model consists of mathematical equations that identify the developed time, the effort and the maintenance effort. The model was developed based on 63 software projects. The estimation accuracy is suggestively improved when adopting models such as the Intermediate and Complex COCOMO models [23]. Equation 1 shows the basic COCOMO model:

$$E = \alpha(SLOC)^{\beta} \tag{1}$$

E presents the software effort computed in man-months. $SLOC$ stands for Source Line Of Code computed in Kilo. The values of the parameters α and β depend mainly on the class of software project. Software projects were classified based on the complexity of the project into three categories. They are: Organic, Semidetached and Embedded models [24]. Extensions of COCOMO, such as COMCOMO II, can be found [25], however, for the purpose of research reported, in this paper, the basic COMCOMO model is used. The three models are given in Table I. These models are expected to give different results according to the type of software projects.

TABLE I. Basic COCOMO Models

Model Name	Effort (E)	Time (T)
Organic Model	$E = 2.4(KLOC)^{1.05}$	$T = 2.5(E)^{0.38}$
Semi-Detached Model	$E = 3.0(KLOC)^{1.12}$	$T = 2.5(E)^{0.35}$
Embedded Model	$E = 3.6(KLOC)^{1.20}$	$T = 2.5(E)^{0.32}$

IV. Function Point Model

Software size helps in developing an initial estimate for software effort/cost estimation during software development life cycle. COCOMO model provided this estimate based on the SLOC. It was reported that SLOC produced many problems [26], [27]. For example, in modern software programming, auto-generate tools produced large number of line of codes. SLOC also changes with the developer's experience, difference in programming languages, variation in the graphical user interface (GUI) code generation, and lack of functionality. The estimation of SLOC under this condition seems uncertain to measure. This is why Albrecht proposed his idea of computing the software size based on the system functionality [28], [29].

A. Albrecht's Function Points

Albrecht's function point gained acceptance during the 1980's and 1990's because of the tempting benefits compared to the models based on the SLOC [30], [31]. Because FP is self-governing and independent of language type, platform, it can be used to identify many productivity benefits. FP is designed to estimate the time required for a software project development, and thereby the cost of the project and maintaining existing software systems.

In 1979 Albrecht [28], published his article on FP methodology while he was working at IBM. The proposed FP has no dimension. FP was computed based on the analysis of project requirements. The requirements help in identifying the number of function to be developed along with the complexity of each function. Thus, there was no need to measure the size of LOC but only concern about project functionality. Once the number of FP measured, the average number of function points per month specified and the labor cost per month is estimated; the total budget can be computed. Albrecht originally proposed four function types [28]: files, inputs, outputs and inquiries with one set of associated weights and ten General System Characteristics (GSC). In 1983, the work developed in (Albrecht and Gaffney 1983), proposed the expansion of the function type, a set of three weighting values (i.e. simple, average, complex) and fourteen General System Characteristics (GSCs) were proposed as given in Table II.

In [32], Kemerer provided a famous study reporting the results of the comparative accuracy for four software cost estimation models. They are the Function Points [28], SLIM [33], COCOMO [22], and ESTIMACS. The results were produced using data collected from 15 completed software projects. Each model was tested based on its predictive capability on computing software cost. The results showed that the models require substantial calibration. Kemer also identified the main attributes which affect software productivity. Recently, using Albrecht's Function Point analysis (FPA) method and using analogous approach, authors [34] provided a methodology they claim it is more reliable and accurate in predicting the software size at an early stage of the software life cycle. Recently, FP gain more attention as a powerful approach for estimating software effort [35]–[37].

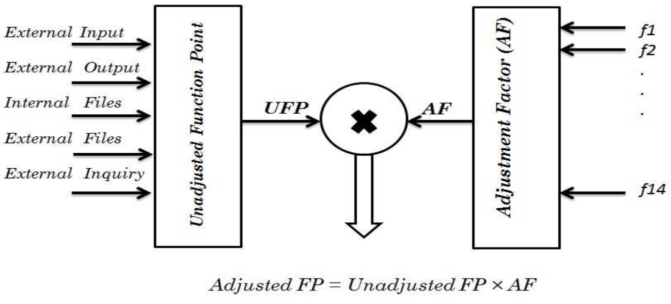


Fig. 1. Function Point Computation Model

In Albrecht FP, there are two parts in the model, which are *Unadjusted Function Point* (UFP) and *Adjusted Function Point* (AFP). The UFP consists of five components. They are:

- External Inputs (EI),

TABLE II. 1983 FUNCTION TYPES AND WEIGHTS

Function Type	Simple	Average	Complex
External Input	3	4	6
External Output	4	5	7
Internal Files	7	10	15
External Files	5	7	10
External Inquiry	3	4	6

- External Outputs (EO),
- External Inquires (EQ),
- Internal Logical Files (ILF) and
- External Interface Files (EIF).

There are also 14 GSCs factors that affect the size of the project effort, and each is ranked from "0"- no influence to "5"-essential. GSCs consists of 14 factors known as $f_1, f_2, \ldots, f_{14}$. These factors are listed in listed in Table III. The sum of all factors is then multiplied given in Equation 2 which constitute the Adjustment Factor (AF) defined in the range [0.65, -1.35].

$$AF = 0.65 + 0.01 \sum_{i=1}^{14} f_i \qquad (2)$$

TABLE III. GENERAL SYSTEM CHARACTERISTICS (GSCS)

1	Data Communications
2	Distributed Functions
3	Performance
4	Heavily Used Configuration
5	Transaction Rate
6	Online Data Entry
7	End User Efficiency
8	Online Update
9	Complex Processing
10	Reusability
11	Installation Ease
12	Operational Ease
13	Multiple Sites
14	Facilitate Change

Then, the Unadjusted FP is then multiplied by the UFP to create the Adjusted Function Point (AFP) count as given in Equation 3. The Adjusted FP value is always within 35% of the original UFP figure. A diagram which shows the process of computing FP is given in Figure 1.

$$Adjusted\ FP = Unadjusted\ FP \times AF \qquad (3)$$

V. WHAT IS FIS?

A block diagram which provide the main architecture of a fuzzy rule based system is shown in Figure 2. The proposed fuzzy logic system, used in this study, consists of number of components.

1) **Fuzzification:** In this stage, the model inputs and outputs variables are defined. These inputs and outputs are transformed to set of fuzzy domains.
2) **Inference Mechanism:** Fuzzy inference mechanism concerns on developing a relationship between the model inputs and outputs. The mapping constructs the system decision making. The process of fuzzy inference include: Membership functions, Fuzzy set operation, and If-Then rules. FIS may be summarized as two processes:
    - *Aggregation:* Compute the IF part (i.e. antecedent) of the rules. The antecedent variables reflect information about the process operating conditions.
    - *Composition:* Compute the THEN part (i.e. consequence) of the rules. The rule's consequent is normally presented as a linear regression model [38]–[40]. This model has set of parameters usually estimated using least square minimization criterion.
3) **Defuzzification:** The computed output based on the fuzzy rules are then converted to real values.

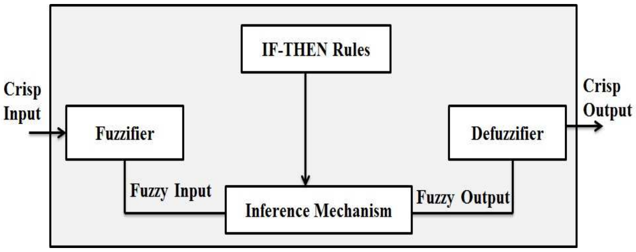


Fig. 2. The proposed fuzzy logic system

VI. PROPOSED FL MODEL

The proposed fuzzy model should be able to mathematically represents the relationship between the effort model inputs $x_1, \ldots x_n$ and the effort y; n is the number of inputs to the fuzzy model. The proposed fuzzy model is always represented by set of If-Then rules. The proposed fuzzy model equation is given as follows:

$$y = FM(x_1, \ldots, x_n) \qquad (4)$$

The fuzzy region in the product space is developed based on the membership functions and the antecedent of the rule. The *antecedent* variable gives the condition of the process status now. The rule's consequent is defined as a set of local linear regression models which relates y with $x_1, \ldots, x_4$ given as in Equation 5.

$$y = a_0 + a_1 x_1 + \cdots + a_n x_n \qquad (5)$$

A rule-based fuzzy model requires the identification of the following:

1) the $antecedent$,
2) the $consequent$ structure,
3) the type of the membership functions for different operating conditions and
4) the estimation of the consequent parameters using least square estimation.

The developed fuzzy models implemented based the Takagi-Sugeno technique [38], [39]. The proposed technique does not require any *a prior* knowledge about the operating regimes. If a sufficiently number of measurements are collected which reflects the operating ranges of interest, the developed fuzzy model will be an efficient one [38], [39].

VII. Computation Criteria

The performance of the developed two models; the SLOC and the FP models based FL shall be evaluated using number of evaluation criteria. They are:

- Variance-Accounted-For (VAF):

$$VAF = [1 - \frac{var(y-\hat{y})}{var(y)}] \times 100\% \tag{6}$$

- Euclidian distance (ED):

$$ED = \sqrt{(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

- Manhattan distance (MD):

$$MD = (\sum_{i=1}^{n}|y_i - \hat{y}_i|) \tag{8}$$

- Mean Magnitude of Relative Error (MMRE):

$$MMRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{9}$$

where y and $\hat{y}$ are the actual effort and the estimated effort based on the developed fuzzy model and N is the number of measurements used in the experiments, respectively.

VIII. Experimental Results

A. The Albrecht data set

A statistical summary of the features used in the analysis of the Albrecht data set is displayed in Table IV. The feature to be predicted (i.e. goal feature or dependent variable) is effort, measured in work-hours, while the potential independent variables (i.e. descriptor features) are adjusted function points, the number of master files, the number of inputs, the number of inquiries and the number of outputs [41].

TABLE IV. Summary Statistics for Albrecht Data Set [41]

Feature	Count	Min	Max	Mean	Median
Effort	24	0.5	105.20	21.88	11.45
FP	24	199.00	1902.00	647.62	506.00
Files	24	3.00	60.00	17.38	11.50
Inputs	24	7.00	193.00	40.25	33.50
Inquiries	24	0.00	75.00	16.88	19.3
Outputs	24	12.00	150.00	47.25	39.00

B. Fuzzy Effort Model based SLOC

We developed a fuzzy model based COCOMO for the effort taking in consideration one attribute which is the SLOC. We used the FMID Matlab Toolbox [42] to develop our experimental results. The set of rules which describe the effort as a function of SLOC is given in Table V. In Table VI, we show the values of each evaluation criteria adopted in this study. In Figure 3, we show the membership function for the SLOC based model. Three membership functions are shown which reflect the relationship between the SLOC and the Effort on three sub-models. Figure 4 show the actual and estimated effort using fuzzy logic. The values of the actual and computed effort based fuzzy model is presented in Table VII. The characteristics between the two curves look very similar with high VAF criteria.

TABLE V. Fuzzy Rules for the Effort based SLOC Model

1. **If** $SLOC$ is A_1 **then**
   $E = 3.20 \cdot 10^{-1} SLOC + 1.57 \cdot 10^{-1}$
2. **If** $SLOC$ is A_2 **then**
   $E = 2.25 \cdot 10^{0} SLOC - 1.96 \cdot 10^{2}$
3. **If** $SLOC$ is A_3 **then**
   $E = -1.47 \cdot 10^{0} SLOC + 3.24 \cdot 10^{2}$

TABLE VI. Computation Criteria for the FL based SLOC Model

VAF	ED	MD	MMRE
96.158%	2.85×10^{-13}	100.27	0.4337

C. Fuzzy Effort Model based FP

We developed a fuzzy model for the effort taking in consideration four attribute inspired by the FP model. Three memberships were used. The set of rules which describe the effort as a function of FP is given in Table V. In Figure 5, we show the membership function for the FP based model. Figure 6 show the actual and estimated effort using fuzzy logic based FP model. We received a very high VAF reflecting good performance modeling. The actual and estimated values of the effort based the FP model is given in Table IX. The developed model's performance were computed using three different criteria as reported in Table X. It can be seen that the performance of the developed fuzzy models based historical data were able to achieve significant modeling results.

IX. Conclusions and Future Work

In this paper we studied the problem of effort estimation for software projects. This is a challenging problem for software project manager. We explored the use of fuzzy logic as a soft computing technique which can simplify the modeling process of the effort. Two models inspired from the COCOMO and FP were developed based fuzzy logic. The developed fuzzy models implemented based the Takagi-Sugeno technique. The developed fuzzy models were tested using the Albrecht data set reported in [41]. The models are simple and show the

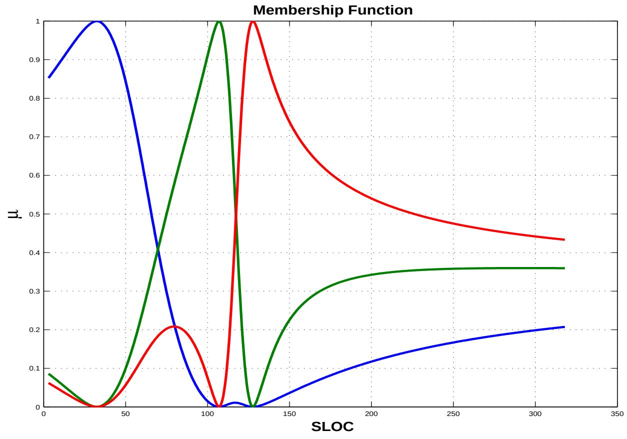


Fig. 3. Membership functions for the FL based SLOC Model

TABLE VIII. FUZZY RULES FOR THE EFFORT BASED FP MODEL

1. **If** $Inputs$ is A_{11} **and** $Outputs$ is A_{12} **and** $Files$ is A_{13} **and** $Inquiries$ is A_{14} **then**
   $FP = 5.30 \cdot 10^0 Inputs + 4.52 \cdot 10^0 Outputs - 8.10 \cdot 10^0 Files + 4.54 \cdot 10^0 Inquiries + 1.08 \cdot 10^2$
2. **If** $Inputs$ is A_{21} **and** $Outputs$ is A_{22} **and** $Files$ is A_{23} **and** $Inquiries$ is A_{24} **then**
   $FP = -3.55 \cdot 10^0 Inputs + 1.23 \cdot 10^1 Outputs + 1.12 \cdot 10^1 Files + 9.29 \cdot 10^0 Inquiries - 6.48 \cdot 10^1$
3. **If** $Inputs$ is A_{31} **and** $Outputs$ is A_{32} **and** $Files$ is A_{33} **and** $Inquiries$ is A_{34} **then**
   $FP = 7.08 \cdot 10^0 Inputs + 1.11 \cdot 10^1 Outputs + 1.45 \cdot 10^1 Files - 8.52 \cdot 10^0 Inquiries - 4.18 \cdot 10^2$

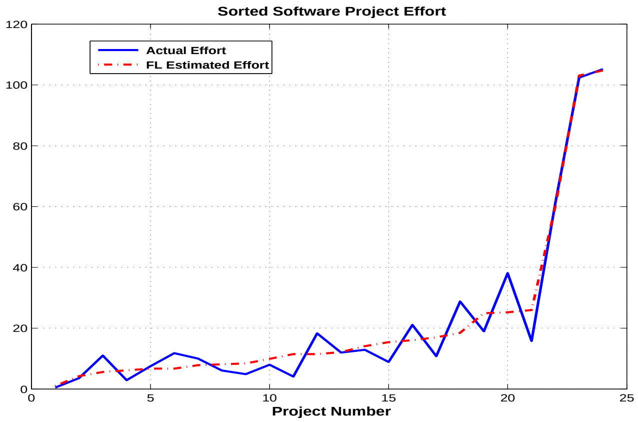


Fig. 4. Actual and estimated Effort for the FL based SLOC Model

TABLE VII. ACTUAL AND ESTIMATED EFFORT: FL-SLOC MODEL

SLOC	Effort	FL-Effort
3	0.5	1.0333
15	3.6	4.2453
20	11.0	5.6040
22	2.9	6.1542
24	7.5	6.7096
24	11.8	6.7096
28	10.0	7.8400
29	6.1	8.1276
30	4.9	8.4177
35	8.0	9.9177
40	4.1	11.4941
40	18.3	11.4941
42	12.0	12.1397
48	12.9	14.1040
52	8.9	15.4103
54	21.1	16.0532
57	10.8	16.9950
62	28.8	18.4773
93	19.0	24.9212
94	38.1	25.2247
96	15.8	25.9818
110	61.2	60.0224
130	102.4	103.0938
318	105.2	104.7296

mathematical relationship between the effort and the main model inputs. This work can be extended by exploring other forms of soft computing techniques.

ACKNOWLEDGMENTS

Dr. Sultan Aljahdali would like to acknowledge the financial support of King Abdulaziz City for Science and Technology and Taif University during his short research visit to Prof. Jayantha Herath with the Computer Science Department, St. Cloud State University, St. Cloud, MN 56301, USA on the Summer of 2013.

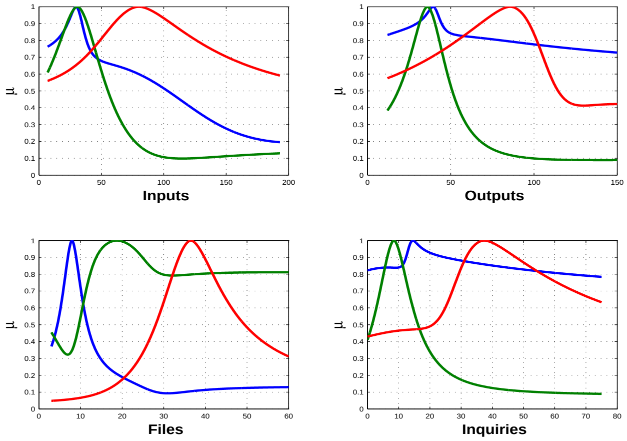


Fig. 5. Membership functions for the FL based FP Model

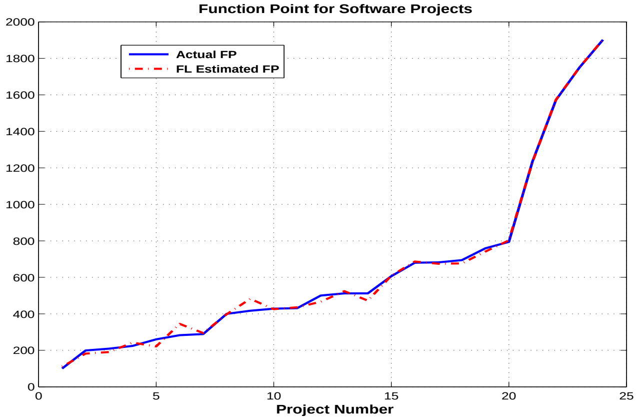


Fig. 6. Actual and estimated Effort for the FL based FP Model

REFERENCES

[1] C. F. Kemere, “An empirical validation of software cost estimation models,” *Communication ACM*, vol. 30, pp. 416–429, 1987.

[2] J. W. Park R, W. Goethert, “Software cost and schedule estimating: A process improvement initiative,” tech. rep., 1994.

[3] M. Boraso, C. Montangero, and H. Sedehi, “Software cost estimation: An experimental study of model performances,” tech. rep., 1996.

[4] K. Pillai and S. Nair, “A model for software development effort and cost estimation,” *IEEE Trans. on Software Engineering*, vol. 23, pp. 485–497, 1997.

[5] L. Pracchia, “Improving the dod: Software acquisition processes,” *The Journal of Defense Software Engineering*, vol. 4, pp. 4–7, 2004.

[6] L. A. Zadeh, “Soft computing and fuzzy logic,” *IEEE Softw.*, vol. 11, pp. 48–56, Nov. 1994.

TABLE IX. ACTUAL AND ESTIMATED EFFORT:FL-FP MODEL

Inputs	Outputs	Files	Inquiries	FP	FL-FP
34	14	5	0	100	110.67
15	15	3	6	199	181.88
7	12	8	13	209	190.7
33	17	5	8	224	241.97
12	15	15	0	260	221.13
13	19	23	0	283	345.08
17	17	5	15	289	294.78
27	20	6	24	400	397.29
28	41	11	16	417	482.22
70	27	12	0	428	425.33
10	69	9	1	431	435.66
25	28	22	4	500	466.19
41	27	5	29	512	524.29
28	38	9	24	512	472.31
42	57	5	12	606	609.17
45	64	16	14	680	686.46
43	40	35	20	682	674.52
61	68	11	0	694	676.44
40	60	12	20	759	741.23
40	60	15	20	794	802.39
48	66	50	13	1235	1231.9
69	112	39	21	1572	1573.2
25	150	60	75	1750	1749.9
193	98	36	70	1902	1903.3

TABLE X. COMPUTATION CRITERIA FOR THE FL BASED FP MODEL

VAF	ED	MD	MMRE
99.741%	2.78×10^{-12}	398.33	0.0495

[7] J. Ryder, *Fuzzy COCOMO: Software Cost Estimation*. PhD thesis, Binghamton University, 1995.

[8] A. C. Hodgkinson and P. W. Garratt, “A neuro-fuzzy cost estimator,” in *Proceedings of the Third Conference on Software Engineering and Applications*, pp. 401–406, 1999.

[9] M. A. Kelly, “A methodlolgy for software cost estimation using machine learning techniques,” Master’s thesis, Naval Postgratuate School, Monterey, California, 1993.

[10] J. J. Dolado and L. F. andez, “Genetic programming, neural network and linear regression in software project estimation,” in *Proceedings of the INSPIRE III, Process Improvement through training and education*, pp. 157–171, British Company Society, 1998.

[11] A. J. Albrecht and J. R. Gaffney, “Software function, source lines of code, and development effort prediction: A software science validation,” *IEEE Trans. Software Engineering*, vol. 9, no. 6, pp. 630–648, 1983.

[12] J. E. Matson, B. E. Barret, and J. M. Mellinchamp, “Software developemnet cost estimation using function points,” *IEEE Trans. Software Engineering*, vol. 20, no. 4, pp. 275–287, 1994.

[13] M. Shepper and C. Schofield, “Estimating software project effort using analogies,” *IEEE Tran. Software Engineering*, vol. 23, pp. 736–743, 1997.

[14] S. Kumar, B. A. Krishna, and P. Satsangi, “Fuzzy systems and neural networks in software engineering project management,” *Journal of Applied Intelligence*, vol. 4, pp. 31–52, 1994.

[15] A. F. Sheta, “Estimation of the COCOMO model parameters using genetic algorithms for NASA software projects,” *Journal of Computer Science*, vol. 2, no. 2, pp. 118–123, 2006.

[16] H. Mittal and P. Bhatia, “A comparative study of conventional effort estimation and fuzzy effort estimation based on triangular fuzzy numbers,” *International Journal of Computer Science and Security*, vol. 1, no. 4, pp. 36–47, 2007.

[17] M. Uysal, “Estimation of the effort component of the software projects using simulated annealing algorithm,” in *World Academy of Science, Engineering and Technology*, vol. 41, pp. 258–261, 2008.

[18] P. S. Sandhu, M. Prashar, P. Bassi, and A. Bisht, “A model for estimation of efforts in development of software systems,” in *World Academy of Science, Engineering and Technology*, vol. 56, pp. 148–152, 2009.

[19] A. Sheta, “Software effort estimation and stock market prediction using takagi-sugeno fuzzy models,” in *Proceedings of the 2006 IEEE Fuzzy Logic Conference, Sheraton, Vancouver Wall Centre, Vancouver, BC, Canada, July 16-21*, pp. 579–586, 2006.

[20] A. Sheta, D. Rine, and A. Ayesh, “Development of software effort and schedule estimation models using soft computing techniques,” in *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE CEC 2008) within the 2008 IEEE World Congress on Computational Intelligence (WCCI 2008), Hong Kong, 1-6 June*, pp. 1283–1289, 2008.

[21] J. W. Bailey and V. R. Basili, “A meta model for software development resource expenditure,” in *Proceedings of the International Conference on Software Engineering*, pp. 107–115, 1981.

[22] B. Boehm, *Software Engineering Economics*. Englewood Cliffs, NJ, Prentice-Hall, 1981.

[23] B. Boehm, *Cost Models for Future Software Life Cycle Process: COCOMO2*. Annals of Software Engineering, 1995.

[24] O. Benediktsson, D. Dalcher, K. Reed, and M. Woodman, “COCOMO based effort estimation for iterative and incremental software development,” *Software Quality Journal*, vol. 11, pp. 265–281, 2003.

[25] B. Boehm and et all, *Software Cost Estimation with COCOMO II*. Prentice Hall PTR, 2000.

[26] T. DeMarco, *Controlling Software Projects*. New York, USA: Yourdon Press/Prentice Hall, Englewood Cliffs, 1982.

[27] C. Jones, *Programming Productivity*. New York, USA: McGraw-Hill, 1986.

[28] A. J. Albrecht, “Measuring application development productivity,” in *Proceedings of the Joint SHARE, GUIDE, and IBM Application Developments Symposium*, pp. 83–92, 1979.

[29] A. J. Albrecht and J. E. Gaffney, “Software function, source lines of code, and development effort prediction: A software science validation,” *IEEE Transactions on Software Engineering*, vol. 9, no. 6, pp. 639–648, 1983.

[30] R. Rask, P. Laamanen, and K. Lyytinen, “A comparison of albrecht’s function point and symons’ mark ii metrics,” in *Proceedings of the thirteenth international conference on Information systems*, ICIS ’92, (Minneapolis, MN, USA), pp. 207–221, University of Minnesota, 1992.

[31] S. Furey, “Why we should use function points [software metrics],” *IEEE Softw.*, vol. 14, pp. 28, 30–, Mar. 1997.

[32] C. F. Kemerer, “An empirical validation of software cost estimation models,” *Commun. ACM*, vol. 30, pp. 416–429, May 1987.

[33] L. Putnam, “A general empirical solution to the macro software sizing and estimation problem,” *IEEE Transactionson Software Engineering*, vol. 4, no. 4, pp. 345–381, 1978.

[34] J. Wu and X. Cai, “A software size measurement model for large-scale business applications,” in *Proceedings of the 2008 International Conference on Computer Science and Software Engineering - Volume 02*, CSSE ’08, (Washington, DC, USA), pp. 39–42, IEEE Computer Society, 2008.

[35] L. Lavazza and S. Morasca, “Convertibility of function points into COSMIC function points: A study using piecewise linear regression,” *Inf. Softw. Technol.*, vol. 53, pp. 874–884, Aug. 2011.

[36] S. Choi, S. Park, and V. Sugumaran, “A rule-based approach for estimating software development cost using function point and goal and scenario based requirements,” *Expert Syst. Appl.*, vol. 39, pp. 406–418, Jan. 2012.

[37] L. Lavazza, S. Morasca, and G. Robiolo, “Towards a simplified definition of function points,” *Inf. Softw. Technol.*, vol. 55, pp. 1796–1809, Oct. 2013.

[38] R. Babuška, *Fuzzy Modeling and Identification*. PhD thesis, Delft Univesrsity of Technology, 1996.

[39] H. A. Babuška, R. Braake, A. J. Krijgsman, and H. B. Verbruggen, “Comparison of intelligent control schemes for real-time pressure control,” *Control Engineering Practice*, vol. 4, pp. 1585–1592, 1996.

[40] A. F. Sheta, *Modeling the Tennessee Eastman Chemical Reactor Using Fuzzy Logic*. New York, USA: Book Chapter. The ISE Book Series on Fuzzy System Engineering-Theory and Practice, published by Nova Science, ISBN: 3-540-25322-X, 2005.

[41] C. Schofield, *An Empirical Investigation into Software Effort Estimation by Analogy*. PhD thesis, Bournemouth University, 1998.

[42] R. Babuška, *Fuzzy Modeling and Identification Toolbox*. Delft University of Technology, The Netherland, http://lcewww.et.tudelft.nl/bubuska, 1998.

A General Framework of Generating Estimation Functions for Computing the Mutual Information of Terms

D. Cai and T.L. McCluskey
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK, HD1 3DH
Email: {d.cai; t.l.mccluskey}@hud.ac.uk

***Abstract*—Computing statistical dependence of terms in textual documents is a widely studied subject and a core problem in many areas of science. This study focuses on such a problem and explores the techniques of estimation using the expected mutual information measure. A general framework is established for tackling a variety of estimations: (i) general forms of estimation functions are introduced; (ii) a set of constraints for the estimation functions is discussed; (iii) general forms of probability distributions are defined; (iv) general forms of the measures for calculating mutual information of terms (MIT) are formalised; (v) properties of the MIT measures are studied and, (vi) relations between the MIT measures are revealed. Four estimation methods, as examples, are proposed and mathematical meanings of the individual methods are respectively interpreted. The methods may be directly applied to practical problems for computing dependence values of individual term pairs. Due to its generality, our method is applicable to various areas, involving statistical semantic analysis of textual data.**

***Keywords*—mutual information of terms (MIT); term dependence; statistical semantic analysis; probability estimation.**

I. Introduction

Analysing and computing statistical dependence (relatedness, proximity, association, similarity) of terms (features, concepts, phrases, words) in textual documents is a widely studied subject in many areas of science. The subject has achieved importance and popularity during the past four decades or so, due chiefly to its demonstrated applications in numerous seemingly diverse areas of science. One of the commonly used tools of analysis and computation is the expected mutual information measure (EMIM) drawn from information theory [?], [?].

The issue of computing the mutual information of terms is an active research topic. A variety of methods have been developed in order to assign dependence values to individual term pairs, and then some decision is made on the basis of the values. Many studies have used the measure for a variety of tasks in, for instance, feature selection [?], [?], [?], [?], document classification [?], face image clustering [?], multi-modality image registration [?], information retrieval [?], [?], [?], [?], [?]. However, it seems that mutual information methods have not achieved their potential. The main problem we face in using EMIM is obtaining actual probability distributions, as the true distributions are invariably not known, and we have to estimate them from training data. This work explores techniques of estimation.

Before introducing a series of formulae, let us first clarify the difference between a term *state value* distribution and a term *occurrence frequency* distribution. A term is usually thought of as having *states* 'present' or 'absent' in a document. Thus, for an arbitrary term t, it will be convenient to introduce a variable δ taking values from set $\Omega = \{1, 0\}$, where $\delta = 1$ expresses that t is present and $\delta = 0$ expresses that t is absent. Denote $t^\delta = t$, $\bar{t}$ when $\delta = 1, 0$, respectively. We call Ω a *state value space*, and each element in Ω a *state value*, of t. Similarly, for an arbitrary term pair (t_i, t_j), we introduce a variable pair (δ_i, δ_j) taking values from set $\Omega \times \Omega = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$. We call $\Omega \times \Omega$ a state value space, and each element in $\Omega \times \Omega$ a state value pair, of (t_i, t_j).

Let $D = \{d_1, d_2, ..., d_m\}$ be a *collection* of documents (training data), and $V = \{t_1, t_2, ..., t_n\}$ a *vocabulary* of terms used to index individual documents in D. Denote $V_d \subseteq V$ as the set of terms occurring in document $d \in D$. Thus, for a given d, the term occurrence frequency distribution, generally denoted by $p_d(t) = p(t|d)$, is over V, whereas for a given term t occurring in d, its state value distribution, denoted by $P_d(\delta) = P(t^\delta|d)$, is over Ω. Obviously, each term $t \in V_d$ is matched to a state value distribution and there are $|V_d|$ state value distributions in total for the document d.

There exists statistical dependence between two terms, t_i and t_j, if the state value of one of them provides mutual information about the probability of the state value of the other [?]. The study [?] shows that there is a relationship between the frequencies (or probabilities) of terms and the mutual information of terms. Therefore, term t_i taking some state value δ_i (say $\delta_i = 1$) should be looked upon as complex because another state value (say $\delta_i = 0$) of t_i, and state values of many other terms (i.e., all terms $t_j \in V - \{t_i\}$), may be dependent on this δ_i [?].

Mathematically, for two arbitrary distinct terms $t_i, t_j \in V$, the *expected mutual information* [?] about the probabilities of the state value pair (δ_i, δ_j) of term pair (t_i, t_j) can be expressed by EMIM:

$$I(\delta_i; \delta_j) = \sum_{\delta_i, \delta_j = 1, 0} P(\delta_i, \delta_j) \log \frac{P(\delta_i, \delta_j)}{P(\delta_i) P(\delta_j)}$$

which measures the amount of information that δ_j provides about δ_i, and vice versa.

Intuitively, a high $I(\delta_i;\delta_j)$ value indicates more of the information that one of two terms t_i and t_j carries is determined by the other and thus the terms are more dependent; a low $I(\delta_i;\delta_j)$ value on the other hand suggests that t_i and t_j are better able to provide self-information and thus are likely to be independent. However, the current study does not support this intuition and instead points out:

1) one should consider the mutual information of t_i and t_j under the individual state values (δ_i, δ_j), where $\delta_i, \delta_j = 1, 0$;

2) one cannot assert that t_i and t_j are highly dependent for their co-occurrence from a high $I(\delta_i;\delta_j)$ value.

The estimation of probability distributions, $P(\delta)$ and $P(\delta_i, \delta_j)$, required in $I(\delta_i;\delta_j)$ is crucial and remains an open issue for effectively distinguishing potentially dependent term pairs from many others and, therefore, the main concern of our current study. We attempt to establish a general framework for constructing estimation functions, with a set of constraints, in order to define $P(\delta)$ and $P(\delta_i, \delta_j)$ meeting some criteria. We next formalise measures for computing the mutual information of terms (MIT) under the individual state values and study corresponding properties of the MIT measures, which is an underlying basis for practical applications. We then propose four estimation methods, as examples, to clarify and illustrate our ideas described in the current study by interpreting their mathematical meanings and discussing corresponding properties. The four estimation methods may be applied directly to practical problems for assigning a dependence value to each term pair.

The remainder of the paper is organized as follows. Section II establishes a general framework for constructing estimation functions and defining probability distributions. Section III formalises the MIT measures and studies their properties. Section IV proposes four estimation methods and discusses corresponding properties. Section V addresses some key points of our study. Conclusions are drawn in Section VI.

II. A General Estimation Framework

In practical applications, the probability distributions of state values may be estimated from training data. This section establishes a general framework in order to define two arguments, $P(\delta)$ and $P(\delta_i, \delta_j)$, required in $I(\delta_i;\delta_j)$. The definition of the joint state value distribution, $P(\delta_i, \delta_j)$, is a more complicated task and the main concern of this section.

In the current study, the probability distributions are defined from estimation functions and, therefore, we need to first introduce the concept of estimation functions. Let $\Xi \subseteq D$ be the set of sample documents considered, and $V_\Xi \subseteq V$ the set of terms occurring in at least one of the documents in Ξ. We have the following definition.

Definition 2.1 For arbitrary terms $t, t_i, t_j \in V$, where $i \neq j$, we define two non-negative functions, denoted by $\psi_\Xi(t)$ and $\gamma_\Xi(t_i, t_j)$, with the form:

$$\begin{aligned} \psi_\Xi(t) &\begin{cases} > 0 & t \in V_\Xi \\ = 0 & t \notin V_\Xi \end{cases} \\ \gamma_\Xi(t_i, t_j) &\begin{cases} > 0 & (t_i, t_j) \in V_\Xi \times V_\Xi \\ = 0 & (t_i, t_j) \notin V_\Xi \times V_\Xi \end{cases} \end{aligned} \tag{1}$$

satisfying a set of *constraints*

$$0 \leq \gamma_\Xi(t_i, t_j) \leq \psi_\Xi(t_i), \psi_\Xi(t_j) < 1 \tag{2}$$

and call $\psi_\Xi(t)$ and $\gamma_\Xi(t_i, t_j)$ the *general forms of estimation functions*.

Definition 2.2 For arbitrary given terms $t, t_i, t_j \in V_\Xi$, where $i \neq j$, suppose $\psi_\Xi(t)$ and $\gamma_\Xi(t_i, t_j)$ are the estimation functions given in Definition 2.1. We define $P_\Xi(\delta)$:

$$\begin{aligned} P_\Xi(\delta = 1) &= \psi_\Xi(t) \\ P_\Xi(\delta = 0) &= 1 - \psi_\Xi(t) \end{aligned} \tag{3}$$

and define $P_\Xi(\delta_i, \delta_j)$:

$$\begin{aligned} P_\Xi(\delta_i = 1, \delta_j = 1) &= \gamma_\Xi(t_i, t_j) \\ P_\Xi(\delta_i = 1, \delta_j = 0) &= \psi_\Xi(t_i) - \gamma_\Xi(t_i, t_j) \\ P_\Xi(\delta_i = 0, \delta_j = 1) &= \psi_\Xi(t_j) - \gamma_\Xi(t_i, t_j) \\ P_\Xi(\delta_i = 0, \delta_j = 0) &= 1 - \psi_\Xi(t_i) - \psi_\Xi(t_j) + \gamma_\Xi(t_i, t_j) \end{aligned} \tag{4}$$

and call $P_\Xi(\delta)$ and $P_\Xi(\delta_i, \delta_j)$ the *general forms of probability distributions* of state values of term pair (t_i, t_j).

Theorem 2.1 Suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i, \delta_j)$ are given in Definition 2.2. Then $P_\Xi(\delta)$ is a probability distribution on $\Omega = \{1, 0\}$; $P_\Xi(\delta_i, \delta_j)$ is a probability distribution on $\Omega \times \Omega$; $P_\Xi(\delta_i)$ and $P_\Xi(\delta_j)$ are the marginal distributions of $P_\Xi(\delta_i, \delta_j)$.

Proof: Clearly, from the above definition and constraints given in (2), $P_\Xi(\delta)$ is a probability distribution on $\Omega = \{1, 0\}$. Also, by the constraints and four expressions in (4), we have

$$P_\Xi(\delta_i, \delta_j) \geq 0$$

for $\delta_i, \delta_j = 1, 0$ and

$$\sum_{\delta_i, \delta_j = 1,0} P_\Xi(\delta_i, \delta_j) = 1$$

Thus $P_\Xi(\delta_i, \delta_j)$ is a probability distribution on $\Omega \times \Omega$. Also, it can easily be seen:

$$\begin{aligned} P_\Xi(\delta_i = 1) &= \sum_{\delta_j = 1,0} P_\Xi(\delta_i = 1, \delta_j) = \psi_\Xi(t_i) \\ P_\Xi(\delta_i = 0) &= \sum_{\delta_j = 1,0} P_\Xi(\delta_i = 0, \delta_j) = 1 - \psi_\Xi(t_i) \end{aligned}$$

Hence, $P_\Xi(\delta_i)$ is the marginal distributions of $P_\Xi(\delta_i, \delta_j)$. A similar discussion may be given for $P_\Xi(\delta_j)$. □

Let us next examine the absolute continuity of $P_\Xi(\delta_i, \delta_j)$ with respect to $P_\Xi(\delta_i)P_\Xi(\delta_j)$, or in symbols, $P_\Xi(\delta_i, \delta_j) \ll P_\Xi(\delta_i)P_\Xi(\delta_j)$. The following theorem serves this purpose.

Theorem 2.2 Suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i, \delta_j)$ are given in Definition 2.2. Then, $P_\Xi(\delta_i, \delta_j) \ll P_\Xi(\delta_i)P_\Xi(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

Proof: The proof is trivial: It can be easily seen, by expressions (1) and (3), that it always has $0 < P_{\Xi}(\delta_i), P_{\Xi}(\delta_j) < 1$ for $\delta_i, \delta_j = 0, 1$ if $t_i, t_j \in V_{\Xi}$. □

It should be emphasized that in order to speak of the mutual information of terms, we must verify the two arguments of $I(\delta_i, \delta_i)$ meeting the following three *criteria* simultaneously:

1) $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are probability distributions,
2) $P_{\Xi}(\delta_i)$ and $P_{\Xi}(\delta_j)$ are the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$,
3) $P_{\Xi}(\delta_i, \delta_j)$ is absolutely continuous with respect to $P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$.

Meeting these three criteria is the major premise when applying $I(\delta_i; \delta_j)$ to effectively capture the mutual information inherent among terms. We will give an example to clarify our idea here in Section V.

We thus learn from Theorems 2.1 and 2.2, under the general framework, that as long as $P_d(\delta)$ and $P_d(\delta_i, \delta_j)$ are defined from the estimation functions satisfying the constraints given in (2), they are probability distributions meeting the three criteria. Consequently, the difficulty becomes:

- to construct $\psi_{\Xi}(t)$ and $\gamma_{\Xi}(t_i, t_j)$ that can capture the occurrence and co-occurrence information of terms practically appropriate and mathematically meaningful in application contexts;
- to verify the constraints given in (2) for each term pair considered in order to ensure that the probability distributions, when defined from $\psi_{\Xi}(t)$ and $\gamma_{\Xi}(t_i, t_j)$, meeting the three criteria.

Thus, the construction of $\psi_{\Xi}(t)$ and $\gamma_{\Xi}(t_i, t_j)$ and verification of the constraints given in (2), which are relatively simple, are the core of obtaining actual probability distributions $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$. Section IV will return to this issue and provide four useful examples, after formalising the MIT measures and discussing their properties and relations in the next section.

III. The MIT Measures

Suppose we are given two arbitrary distinct terms $t_i, t_j \in V_{\Xi}$. In order to measure the mutual information of terms t_i and t_j, we need to consider the mutual information under each state value (δ_i, δ_j), namely, we need to measure the extent of the contribution made by the individual state values to EMIM:

$$\begin{aligned} I_{\Xi}(\delta_i; \delta_j) &= \sum_{\delta_i, \delta_j = 1,0} P_{\Xi}(\delta_i, \delta_j) \log \frac{P_{\Xi}(\delta_i, \delta_j)}{P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)} \\ &= \sum_{\delta_i, \delta_j = 1,0} \mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) \end{aligned} \tag{5}$$

Note that the above expression can be expressed as a sum of four items. Each of four items,

$$\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) = P_{\Xi}(\delta_i, \delta_j) \log \frac{P_{\Xi}(\delta_i, \delta_j)}{P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)} \tag{6}$$

can be regarded as '**m**utual **i**nformation of **t**erms', t_i and t_j, in support of dependence but rejecting independence under state value (δ_i, δ_j), where $\delta_i, \delta_j = 1, 0$. Thus, we can regard each item as a MIT measure, computing the extent of the contributions made by the corresponding state value to $I_{\Xi}(\delta_i; \delta_j)$.

Now, substituting estimates (3) and (4) into (6), corresponding to respective four state value pairs, $(1, 1)$, $(1, 0)$ $(0, 1)$, $(0, 0)$, we can formalise the general forms of the four MIT measures by a definition below:

Definition 3.1 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are the probability distributions given in Definition 2.2. Then the *general forms of four MIT measures* can be defined as follows.

$$\mathbf{mit}_{\Xi}(t_i, t_j) = \gamma_{\Xi}(t_i, t_j) \log \frac{\gamma_{\Xi}(t_i, t_j)}{\psi_{\Xi}(t_i)\psi_{\Xi}(t_j)}$$

which computes the dependence of terms t_i and t_j for their co-occurrence in Ξ;

$$\mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = \big(\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j)\big) \log \frac{\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j)}{\psi_{\Xi}(t_i)\big(1 - \psi_{\Xi}(t_j)\big)}$$

which computes the dependence of term t_i occurring but term t_j not occurring in Ξ;

$$\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = \big(\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j)\big) \log \frac{\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j)}{\big(1 - \psi_{\Xi}(t_i)\big)\psi_{\Xi}(t_j)}$$

which computes the dependence of term t_i not occurring but term t_j occurring in Ξ;

$$\begin{aligned} \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = \big(1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)\big) \times \\ \log \frac{1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)}{\big(1 - \psi_{\Xi}(t_i)\big)\big(1 - \psi_{\Xi}(t_j)\big)} \end{aligned}$$

which computes the dependence of both terms t_i and t_j not occurring in Ξ.

Clearly, each of the four MIT measures is uniquely determined by the estimation functions $\psi_{\Xi}(t)$ and $\gamma_{\Xi}(t_i, t_j)$.

Next, we give some interesting properties of the four MIT measures by Theorem 3.1 below. The properties derive their importance from the fact that they underpin the methods proposed in the current study and are essential for guiding practical applications.

Theorem 3.1 Suppose the four MIT measures are given in Definition 3.1. Then we have the following properties:

(a) if $\gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) > 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \leq 0 \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) > 0 \end{aligned}$$

(b) if $\gamma_{\Xi}(t_i, t_j) = \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = 0 \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = 0 \end{aligned}$$

(c) if $\gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$ then

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) < 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \geq 0 \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \geq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) < 0 \end{aligned}$$

Proof: The proof of (b) is obvious. Here we prove only (a), and a similar proof can be given for (c). Consider the general forms of the four MIT measures. From $\gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$, we have:

$$\begin{aligned}
&\gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_i) - \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \psi_{\Xi}(t_i)\big(1 - \psi_{\Xi}(t_j)\big)\\
&\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_j) - \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \psi_{\Xi}(t_j)\big(1 - \psi_{\Xi}(t_i)\big)\\
&1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)\\
&\qquad\qquad > 1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \big(1 - \psi_{\Xi}(t_i)\big)\big(1 - \psi_{\Xi}(t_j)\big)
\end{aligned}$$

which correspond respectively to

$$\begin{aligned}
&\frac{\gamma_{\Xi}(t_i, t_j)}{\psi_{\Xi}(t_i)\psi_{\Xi}(t_j)} > 1,\\
&\frac{\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j)}{\psi_{\Xi}(t_i)\big(1 - \psi_{\Xi}(t_j)\big)} < 1,\\
&\frac{\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j)}{\psi_{\Xi}(t_j)\big(1 - \psi_{\Xi}(t_i)\big)} < 1,\\
&\frac{1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)}{\big(1 - \psi_{\Xi}(t_i)\big)\big(1 - \psi_{\Xi}(t_j)\big)} > 1
\end{aligned}$$

On the other hand, $0 < \gamma_{\Xi}(t_i, t_j) \le \psi_{\Xi}(t_i) < 1$ and $0 < \gamma_{\Xi}(t_i, t_j) \le \psi_{\Xi}(t_j) < 1$ for $t_i, t_j \in V_{\Xi}$. Thus, we have

$$\begin{aligned}
&\gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) > 0\\
&\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j) \ge 0\\
&\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j) \ge 0\\
&1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)\\
&\qquad = \big(1 - \psi_{\Xi}(t_i)\big)\big(1 - \psi_{\Xi}(t_j)\big) > 0
\end{aligned}$$

Hence, the four inequalities in (a) hold. □

The properties given in Theorem 3.1 enable us to gain an insight into the signs of the four MIT measures. That is, we have

$$\mathbf{mit}_{\Xi}(t_i, t_j), \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) \begin{cases} > 0 & \gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \\ = 0 & \gamma_{\Xi}(t_i, t_j) = \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \\ < 0 & \gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \end{cases}$$

$$\mathbf{mit}_{\Xi}(t_i, \bar{t}_j), \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \begin{cases} \le 0 & \gamma_{\Xi}(t_i, t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \\ = 0 & \gamma_{\Xi}(t_i, t_j) = \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \\ \ge 0 & \gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) \end{cases}$$

Clearly, the relation between $\gamma_{\Xi}(t_i, t_j)$ and $\psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$ can infer all the signs of $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$. Thus, with the properties given in Theorem 3.1, we can further learn the relations of the four MIT measures from the signs:

- The signs of $\mathbf{mit}_{\Xi}(t_i, t_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j)$ are always the same, so are the signs of $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$;
- The signs of $\mathbf{mit}_{\Xi}(t_i, t_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j)$ are always opposite to the signs of $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$.

The relations tells us a key point of applying $I_{\Xi}(\delta_i; \delta_j)$, which we will explain in Section V.

IV. Example Estimations

As mentioned previously, the construction of the estimation functions and verification of the constraints are the core of defining actual probability distributions. This section presents four estimation methods, as examples, to illustrate our ideas described in the previous section. The first three consider the estimates in individual documents (i.e., $|\Xi| = 1$), and the last one considers the estimate in the set of documents (i.e., $|\Xi| > 1$).

In what follows, we always assume that $2 < |V_d| \le n$ (where $n = |V|$), namely, each document $d \in D$ has at least three distinct terms. Also, for an arbitrary term $t \in V$, we denote

$$p_d(t) = p(t|d) = \begin{cases} \frac{f_d(t)}{||d||} & t \in V_d \\ 0 & t \notin V_d \end{cases}$$

where $f_d(t)$ is the occurrence *frequency* of term t in d and $||d|| = \sum_{t \in V_d} f_d(t)$ as the *length* of d.

A. Estimate in a Single Document

Suppose each document d is represented by a $1 \times n$ frequency matrix

$$\mathbf{m}_d = \big[f_d(t_1), f_d(t_2), ..., f_d(t_n)\big] = \big[f_d(t)\big]_{1 \times n}$$

in which, each element in the matrix satisfies $f_d(t) > 0$ when $t \in V_d$ and $f_d(t) = 0$ when $t \in V - V_d$.

Then, for an arbitrary term $t \in V$, introduce an estimation function:

$$\psi_d(t) = \begin{cases} \frac{f_d(t)}{\sum_{t' \in V_d} f_d(t')} & t \in V_d \\ 0 & t \notin V_d \end{cases} \tag{7}$$

Clearly, we have $0 < \psi_d(t) < 1$ for every $t \in V_d \subseteq V$. Next, for an arbitrary given term $t \in V_d$, define a probability distribution by expression (3):

$$\begin{aligned} P_d(\delta = 1) &= \psi_d(t) = p_d(t) \\ P_d(\delta = 0) &= 1 - p_d(t) \end{aligned} \tag{8}$$

The function $\psi_d(t)$ and distribution $P_d(\delta)$ will be used in the three methods below.

A.1 Method One

For two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_d(t_i, t_j) = \begin{cases} \frac{f_d(t_i) f_d(t_j)}{\varpi} & (t_i, t_j) \in V_d \times V_d \\ 0 & (t_i, t_j) \notin V_d \times V_d \end{cases} \tag{9}$$

where the denominator of γ_d is,

$$\varpi = \sum_{i' < j';\ t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'})$$

which is the sum of all the possible products $f_d(t_{i'}) f_d(t_{j'})$ for $i' < j'$; $i', j' \in \{1, 2, ..., n\}$, Clearly, as $|V_d| \ge 3$, we have $0 < \gamma_d(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_d \times V_d \subseteq V \times V$.

Next, for two arbitrary given terms $t_i, t_j \in V_d$ (where $i \neq j$), define a probability distribution by expression (4):

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 1) &= \frac{f_d(t_i) f_d(t_j)}{\varpi} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{f_d(t_i)}{||d||} - \gamma_d(t_i, t_j) \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{f_d(t_j)}{||d||} - \gamma_d(t_i, t_j) \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) \end{aligned} \tag{10}$$

In order to verify the constraints given in (2):

$$\gamma_d(t_i, t_j) \leq \psi_d(t_i), \psi_d(t_j)$$

for an arbitrary $t \in V_d$, let us denote

$$\varpi_t = \sum_{i' < j';\ t_{i'}, t_{j'} \in V_d - \{t\}} f_d(t_{i'}) f_d(t_{j'}) \leq \varpi$$

Study [?] has proven, for the functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ given in (7) and (9), respectively, we have:

- $\varpi_{t_i} \geq f_d^2(t_i)$ if and only if $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$;
- $\varpi_{t_j} \geq f_d^2(t_j)$ if and only if $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$.

Thus we can write immediately the following theorem [?].

Theorem 4.1 The expression, $P_d(\delta_i, \delta_j)$, defined in (10) is a probability distribution if $\varpi_{t_i} \geq f_d^2(t_i)$ and $\varpi_{t_j} \geq f_d^2(t_j)$.

The above theorem tells us, when the estimation functions given in (7) and (9) are used, that $P_d(\delta_i, \delta_j)$ given in (10) is a probability distribution if two conditions $\varpi_{t_j} \geq f_d^2(t_j)$ and $\varpi_{t_i} \geq f_d^2(t_i)$ are satisfied simultaneously. The conditions can also be verified by $p_d(t_i) \geq \gamma_d(t_i, t_j)$ and $p_d(t_j) \geq \gamma_d(t_i, t_j)$, respectively, which may be easier to compute in practical application. Next, we give the property of the MIT measures by the following corollary.

Corollary 4.1 For the four MIT measures derived from expressions (8) and (10), four inequalities,

$$\begin{aligned} \mathbf{mit}_d(t_i, t_j) > 0, \quad \mathbf{mit}_d(t_i, \bar{t}_j) \leq 0 \\ \mathbf{mit}_d(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_d(\bar{t}_i, \bar{t}_j) > 0 \end{aligned}$$

always hold if $\varpi_{t_j} \geq f_d^2(t_j)$ and $\varpi_{t_i} \geq f_d^2(t_i)$.

Proof: By Theorem 4.1, $P_d(\delta_i, \delta_j)$ given in (10) is a probability distribution for terms $t_i, t_j \in V_d$. Also,

$$\begin{aligned} \varpi &= \sum_{i' < j';\ t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'}) \\ &< \sum_{t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'}) = ||d||^2 \end{aligned}$$

from which we have

$$\gamma_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\varpi} > \frac{f_d(t_i)}{||d||} \frac{f_d(t_j)}{||d||} = p_d(t_i) p_d(t_j)$$

Thus, from (a) of Theorem 3.1, four inequalities hold. □

A.2 Method Two

Note that $f_d(t)$ is the number of time(s) that term t occurs in d and that $f_d(t_1) + f_d(t_2) + ... + f_d(t_n) = ||d||$. Thus, the probability that two distinct terms t_i and t_j are simultaneously found in d should be

$$\begin{aligned} & \frac{C^1_{f_d(t_i)} C^1_{f_d(t_j)}}{C^2_{||d||}} \\ &= \frac{[f_d(t_i)]!}{1![f_d(t_i) - 1]!} \frac{[f_d(t_j)]!}{1![f_d(t_j) - 1]!} \Big/ \frac{||d||!}{2!(||d|| - 2)!} \\ &= \frac{2 f_d(t_i) f_d(t_j)}{||d|| \cdot (||d|| - 1)} \end{aligned}$$

Hence, for two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_d(t_i, t_j) = \begin{cases} \frac{2 f_d(t_i) f_d(t_j)}{||d|| \cdot (||d|| - 1)} & (t_i, t_j) \in V_d \times V_d \\ 0 & (t_i, t_j) \notin V_d \times V_d \end{cases} \tag{11}$$

which satisfies $0 < \gamma_d(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_d \times V_d \subseteq V \times V$ as $|V_d| \geq 3$.

Next, for two arbitrary given terms $t_i, t_j \in V_d$ (where $i \neq j$), define a probability distribution by (4):

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 1) &= \frac{2 f_d(t_i) f_d(t_j)}{||d|| \cdot (||d|| - 1)} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{f_d(t_i)}{||d||} \Big(1 - \frac{2 f_d(t_j)}{||d|| - 1} \Big) \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{f_d(t_j)}{||d||} \Big(1 - \frac{2 f_d(t_i)}{||d|| - 1} \Big) \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) \end{aligned} \tag{12}$$

We may give two conditions of $P_d(\delta_i, \delta_j)$, such that it satisfies the constraints given in (2) by the following theorem.

Theorem 4.2 The expression, $P_d(\delta_i, \delta_j)$, defined in (12) is a probability distribution if $f_d(t_i) \leq \frac{||d|| - 1}{2}$ and $f_d(t_j) \leq \frac{||d|| - 1}{2}$.

Proof: From $f_d(t_j) \leq \frac{||d|| - 1}{2}$, we have $1 \geq \frac{2 f_d(t_j)}{||d|| - 1}$, that is,

$$p_d(t_i) = \frac{f_d(t_i)}{||d||} \geq \frac{2 f_d(t_i) f_d(t_j)}{||d|| \cdot (||d|| - 1)} = \gamma_d(t_i, t_j)$$

A similar proof can be applied to $p_d(t_j) \geq \gamma_d(t_i, t_j)$. □

Next, we can give the property of the MIT measures by the following corollary.

Corollary 4.2 For the four MIT measures derived from expressions (8) and (12), four inequalities,

$$\begin{aligned} \mathbf{mit}_d(t_i, t_j) > 0, \quad \mathbf{mit}_d(t_i, \bar{t}_j) \leq 0 \\ \mathbf{mit}_d(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_d(\bar{t}_i, \bar{t}_j) > 0 \end{aligned}$$

always hold if $f_d(t_i) \leq \frac{||d||-1}{2}$ and $f_d(t_j) \leq \frac{||d||-1}{2}$.

Proof: By Theorem 4.2, $P_d(\delta_i, \delta_j)$ given in (12) is a probability distribution for terms $t_i, t_j \in V_d$. Also, we have $||d|| \cdot (||d|| - 1) < ||d|| \cdot ||d||$, thus,

$$\begin{aligned}\gamma_d(t_i, t_j) &= \frac{2f_d(t_i)f_d(t_j)}{||d||(||d||-1)} > \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-1} \\ &> \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||} = p_d(t_i)p_d(t_j)\end{aligned}$$

Hence, from (a) of Theorem 3.1, the four inequalities hold. □

A.3 Method Three

The probability that term t_j is found in d after term t_i has been found in d, where $i \neq j$, should be

$$P_d(\delta_j = 1 | \delta_i = 1) = \frac{f_d(t_j)}{||d|| - f_d(t_i)}$$

Thus, for two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_d(t_i, t_j) = \begin{cases} \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-f_d(t_i)} & (t_i, t_j) \in V_d \times V_d \\ 0 & (t_i, t_j) \notin V_d \times V_d \end{cases} \tag{13}$$

which satisfies $0 < \gamma_d(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_d \times V_d \subseteq V \times V$ as $|V_d| \geq 3$.

Next, for two arbitrary given terms $t_i, t_j \in V_d$ (where $i \neq j$), define a probability distribution by (4):

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 1) &= \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d|| - f_d(t_i)} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{f_d(t_i)}{||d||}\left(1 - \frac{f_d(t_j)}{||d|| - f_d(t_i)}\right) \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{f_d(t_j)}{||d||}\left(1 - \frac{f_d(t_i)}{||d|| - f_d(t_j)}\right) \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) \end{aligned} \tag{14}$$

We need to find out if there exists any verification condition, such that $P_d(\delta_i, \delta_j)$ satisfies the constraints given in (2), by the following theorem.

Theorem 4.3 The expression, $P_d(\delta_i, \delta_j)$, defined in (14) is a probability distribution.

Proof: Notice that $f_d(t_i) + f_d(t_j) \leq ||d||$. Thus,

$$1 \geq \frac{f_d(t_j)}{||d|| - f_d(t_i)}$$

that is,

$$p_d(t_i) = \frac{f_d(t_i)}{||d||} > \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d|| - f_d(t_i)} = \gamma_d(t_i, t_j)$$

A similar proof can be applied to $p_d(t_j) > \gamma_d(t_i, t_j)$. □

It is clear, unlike Methods 1 and 2, that $P_d(\delta_i, \delta_j)$ in (14) is a probability distribution unconditionally. Next, we give the property of the MIT measures by the following corollary.

Corollary 4.3 For the four MIT measures derived from expressions (8) and (14), four inequalities

$$\begin{aligned} \mathbf{mit}_d(t_i, t_j) > 0, &\quad \mathbf{mit}_d(t_i, \bar{t}_j) \leq 0 \\ \mathbf{mit}_d(\bar{t}_i, t_j) \leq 0, &\quad \mathbf{mit}_d(\bar{t}_i, \bar{t}_j) > 0 \end{aligned}$$

always hold for arbitrary terms $t_i, t_j \in V_d$.

Proof: By Theorem 4.3, $P_d(\delta_i, \delta_j)$ given in (14) is a probability distribution for terms $t_i, t_j \in V_d$. Also, from $||d|| - f_d(t_i) < ||d||$, we have,

$$\begin{aligned}\gamma_d(t_i, t_j) &= \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d|| - f_d(t_i)} > \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||} \\ &= p_d(t_i)p_d(t_j)\end{aligned}$$

Hence, from (a) of Theorem 3.1, the four inequalities hold. □

B. Estimate in a Set of Documents

The above three estimation methods consistently use frequency representation for the individual documents. However, in some probabilistic methods, one would state that the *binary assumption* suffices to specify the dependence of terms. The method discussed here is under this assumption.

By 'binary' it is here meant that each document $d \in D$ is represented by a $1 \times n$ matrix:

$$\mathbf{m}_d = \left[t_1^{\delta_1}, t_2^{\delta_2}, ..., t_n^{\delta_n}\right] = \left[t^{\delta}\right]_{1 \times n}$$

in which, each element in the matrix is a binary number satisfying $t^\delta = 1$ when $t \in V_d$ and $t^\delta = 0$ when $t \in V - V_d$.

Consider a sample set Ξ, satisfying $|\Xi| > 1$. Denote $n_\Xi(t)$ as the number of documents in Ξ in which term t occurs, and $n_\Xi(t_i, t_j)$ as the number of documents in Ξ in which terms t_i and t_j co-occur. It can be easily seen $n_\Xi(t_i, t_j) \leq n_\Xi(t_i), n_\Xi(t_j) \leq |\Xi|$

Then, for an arbitrary term $t \in V$, introduce an estimation function:

$$\psi_\Xi(t) = \begin{cases} \frac{n_\Xi(t)}{|\Xi|} & t \in V_\Xi \\ 0 & t \notin V_\Xi \end{cases} \tag{15}$$

Obviously, we have $0 < \psi_\Xi(t) < 1$ for every $t \in V_\Xi \subseteq V$. Next, for an arbitrary given term $t \in V_d$, define a probability distribution by expression (3):

$$\begin{aligned} P_\Xi(\delta = 1) &= \psi_\Xi(t) \\ P_\Xi(\delta = 0) &= 1 - \psi_\Xi(t) \end{aligned} \tag{16}$$

The function $\psi_\Xi(t)$ and distribution $P_\Xi(\delta)$ will be used in the fourth method below.

B.1 Method Four

For two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_\Xi(t_i, t_j) = \begin{cases} \frac{n_\Xi(t_i, t_j)}{|\Xi|} & (t_i, t_j) \in V_\Xi \times V_\Xi \\ 0 & (t_i, t_j) \notin V_\Xi \times V_\Xi \end{cases} \tag{17}$$

which satisfies $0 < \gamma_\Xi(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_\Xi \times V_\Xi \subseteq V \times V$ as $|V_d| \geq 3$.

Next, for two arbitrary given terms $t_i, t_j \in V_{\Xi}$ (where $i \neq j$), define a probability distribution by expression (4):

$$\begin{aligned}
P_{\Xi}(\delta_i = 1, \delta_j = 1) &= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = \gamma_{\Xi}(t_i, t_j) \\
P_{\Xi}(\delta_i = 1, \delta_j = 0) &= \frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|} \\
&= \psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j) \qquad (18) \\
P_{\Xi}(\delta_i = 0, \delta_j = 1) &= \frac{n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)}{|\Xi|} \\
&= \psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j) \\
P_{\Xi}(\delta_i = 0, \delta_j = 0) &= 1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)
\end{aligned}$$

It is interesting to note that $\psi_{\Xi}(t_i), \psi_{\Xi}(t_j) \geq \gamma_{\Xi}(t_i, t_j)$ as

$$\frac{n_{\Xi}(t_i)}{|\Xi|}, \frac{n_{\Xi}(t_j)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|}$$

for arbitrary $t_i, t_j \in V_{\Xi}$. Hence the estimation functions given in (15) and (17) satisfy the constraints given in (2) and, thus we can give the following theorem.

Theorem 4.4 The expression, $P_d(\delta_i, \delta_j)$, defined in (18) is a probability distribution.

Like Method 3, $P_{\Xi}(\delta_i, \delta_j)$ given in (18) is a probability distribution unconditionally. From Theorem 4.4, we may give the properties of the MIT measures by the following corollary.

Corollary 4.4 For the four MIT measures derived from expressions (16) and (18),

(a') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
\mathbf{mit}_{\Xi}(t_i, t_j) > 0, &\quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \leq 0 \\
\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \leq 0, &\quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) > 0
\end{aligned}$$

(b') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
\mathbf{mit}_{\Xi}(t_i, t_j) = 0, &\quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = 0 \\
\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = 0, &\quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = 0
\end{aligned}$$

(c') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
\mathbf{mit}_{\Xi}(t_i, t_j) < 0, &\quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \geq 0 \\
\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \geq 0, &\quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) < 0
\end{aligned}$$

The Method 4 is the most commonly used in many areas, such as, information retrieval, natural language processing, document classification, sentiment analysis, and many related areas. More discussion on this method, including its properties and potential application problems, can also be found in [**?**].

V. Discussion

Some key points, which are helpful to understand the methods proposed under the general framework, are addressed in this section. These key points are also important to guide practical applications.

First, it should be possible, though it may not be easy, to construct a variety of estimation functions and then to define probability distributions and verify the corresponding constraints for formalising the MIT measures. For suitable choices of the estimation functions practically appropriate for and mathematically meaningful to a specific application problem, the term state distributions, when substituted into measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ $(\delta_i, \delta_j = 0, 1)$ and/or $I(\delta_i; \delta_i)$, can be expected to capture the mutual information of terms. The information may be used to develop a variety of techniques in order to assign dependence values to individual term pairs and, then some decision is made on the values. A summary of the four example estimation methods proposed in this study is given in Table I. It is important to understand that the MIT measures formalised by different estimation methods may have entirely different properties. For instance, let us return to the four example estimations discussed in Section IV and consider an inequality,

$$\gamma_d(t_i, t_j) > p_d(t_i)p_d(t_j) = \psi_d(t_i)\psi_d(t_j)$$

Then some key points regarding the properties and relationships of the MIT measures of the four corresponding Methods 1–4 can be made below.

- Theorems/Corollaries 4.1–4.3 in respective Methods 1–3 tell us, when estimation functions (7), (9), (11) and (13) are used, that the above inequality always holds, and that terms co-occurring in document d must be more or less statistically dependent since it is always $\mathbf{mit}_d(t_i, t_j) > 0$ supporting a dependence assertion.
- Theorem/Corollary 4.4 in Method 4 tells us, when estimation functions (15) and (17) are used, that the above inequality does not always hold, and that terms may or may not be statistically dependent for their co-occurrence since the sign of $\mathbf{mit}_{\Xi}(t_i, t_j)$ might be different from term pair to term pair.

Therefore, we can learn from the Theorems/Corollaries: for two terms making the above inequality hold, some estimation functions ensure them to be more or less dependent for their co-occurrence, whereas other estimation functions cannot guarantee them to be dependent for their co-occurrence. This also clearly indicates, for the same term pairs, that different estimation methods may result in entirely different conclusions regarding the statistical dependence for their co-occurrence.

Second, as we all knew, the MIT measures may influence experimental performance significantly. However, as the probability distributions are normally obtained according to practical application, it seems that only the "form" of the mutual information measure has frequently been the main concern of research in literature, whereas the problem of verification of the probability distributions is often ignored as an unimportant matter. This implicitly means that a function with a form

$$i(x_1, x_2) = P(x_1, x_2) \log \frac{P(x_1, x_2)}{P(x_1)P(x_2)}$$

would be a "mutual information measure" of x_1 and x_2 for their co-occurrence, and that the discussion on the three criteria of $P(x)$ and $P(x_1, x_2)$ in the function are trivial. This is indeed not true. It is important to realise that it is not necessarily that the function, $i(x_1, x_2)$, is a mutual information measure. In fact, $i(x_1, x_2)$ is not a mutual information measure in the

TABLE I. A SUMMARY OF THE FOUR EXAMPLE ESTIMATIONS

Method	Function $\psi(t)$	Function $\gamma(t_i, t_j)$	Conditions
1	$\psi_d(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\varpi}$	$f_d^2(t_i) \leq \varpi_{t_i}$, $f_d^2(t_j) \leq \varpi_{t_j}$
2	$\psi_d(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_d(t_i, t_j) = \frac{2 f_d(t_i) f_d(t_j)}{\|\|d\|\| \cdot (\|\|d\|\| - 1)}$	$f_d(t_i) \leq \frac{\|\|d\|\| - 1}{2}$, $f_d(t_j) \leq \frac{\|\|d\|\| - 1}{2}$
3	$\psi_d(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_d(t_i, t_j) = \frac{f_d(t_i)}{\|\|d\|\|} \frac{f_d(t_j)}{\|\|d\|\| - f_d(t_i)}$	none
4	$\psi_\Xi(t) = \frac{n_\Xi(t)}{\|\Xi\|}$	$\gamma_\Xi(t_i, t_j) = \frac{n_\Xi(t_i, t_j)}{\|\Xi\|}$	none

information-theoretic sense, if $P(x)$ and $P(x_1, x_2)$ are not probability distributions and/or, if $P(x_1)$ and $P(x_2)$ are not marginal distributions of the joint distribution $P(x_1, x_2)$ (even though they may be all probability distributions). It may not even converge if $P(x_1, x_2) \ll f_1(x_1)$ and $P(x_1, x_2) \ll P(x_2)$ do not hold. Therefore, in practical applications, it entirely makes no sense to use some function, looking like a mutual information measure, to compute the mutual information of terms when any one of the three criteria is not satisfied. We emphasize that the verification of $P(\delta)$ and $P(\delta_i, \delta_j)$ meeting the three criteria is the major premise when applying $I(\delta_i; \delta_j)$ to effectively capture the mutual information inherent among terms. A simple but interesting example given in our related study [**?**] may clarify our idea. We here give a brief explanation and details of computation can be found in [**?**]. Suppose we are given a document $d = \{t_1, t_2, t_2, t_2, t_3, t_4\} \in D$. This example considers the estimation functions given in Method 1 and illustrates a specific instance of failing to apply them for two terms $t_1, t_2 \in V_d$:

$$\varpi = \sum_{i' < j';\ t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'}) = 12$$

and, with expressions (7), (8), (9) and (10), we have $\gamma_d(t_1, t_2) = \frac{1}{4}$ and

$$P_d(\delta_1 = 1, \delta_2 = 0) = \psi_d(t_1) - \gamma_d(t_1, t_2) = -\frac{1}{12}$$

It can be easily seen, for the term pair (t_1, t_2), that the corresponding $P_d(\delta_1, \delta_2)$ is not a probability distribution since the constraints given in (2) are not satisfied (i.e., $\psi_d(t_1) < \gamma_d(t_1, t_2)$). Consequently, $P_d(\delta_1 = 1, \delta_2 = 1)$ is not reliable for measuring dependence of t_1 and t_2 for their co-occurrence. The key points regarding the probability distributions are:

- There may be many term pairs, of which the corresponding $P_d(\delta_i, \delta_j)$ is indeed a probability distribution. However, it is possible that not all term pairs have the corresponding probability distribution.
- In order to compute MIT of terms, we must verify the constraints given in (2), that is, we have to check both $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$ and $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$ to be satisfied simultaneously, for each of the term pairs considered.

Thus, those term pairs (rather than two individual terms), of which the corresponding $P_d(\delta_i, \delta_j)$ does not satisfy the constraints, should be discarded immediately and omitted from the computation of MIT.

Third, the estimation functions given in Methods 1-3 can be applied to document representations not only for $\mathbf{m}_d = \left[f_d(t)\right]_{1 \times n}$, but also for a more general case, where each document d can be represented by a $1 \times n$ (weight) matrix:

$$\mathbf{m}_d = \left[w_d(t_1), w_d(t_2), ..., w_d(t_n)\right] = \left[w_d(t)\right]_{1 \times n}$$

in which, each element is a real number, satisfying $w_d(t) > 0$ when $t \in V_d$ and $w_d(t) = 0$ when $t \in V - V_d$. The $w_d(t)$ is called a *weighting function*, which indicates the importance of term t in representing document d. For instance, the weighting function in Methods 1-3 is $w_d(t) = f_d(t)$. The key points regarding the estimation functions are below.

- Methods 1-3 should be applicable to any quantitative document representation.
- $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ should be used to capture the information of occurrence and co-occurrence of terms.
- $w_d(t)$ should be the main component of the estimation functions, it is construed by means of occurrence frequencies and co-occurrence frequencies of terms.

The extension of, for instance, Method 1 can be found in another of our studies [**?**]. It is beyond the scope of the current paper to discuss the issue of document representation in greater detail, and some formal discussion and technical treatment can be found in, for instance studies [**?**], [**?**], [**?**].

Fourth, it is certainly true that the MIT measures given in Definition 3.1 can be used to measure the extent of dependence of terms t_i and t_j. Also, it is certainly true that the larger quantities the measures offer, the higher the extent term t_i is statistically dependent on term t_j (and vice versa). However, the *implications* of the dependence obtained from the individual MIT measures are different. Remember that we always emphasize ‘the dependence *under the state value* (δ_i, δ_j)’. This emphasis is necessary because it clearly indicates that it is the state value (δ_i, δ_j) that supports the dependence. For instance, terms t_i and t_j may depend highly on one another, when t_i occurs but t_j does not occur in some document and, in this case, we are talking about the dependence under the state value $(\delta_i, \delta_j) = (1, 0)$. In a practical application, what we generally concentrate on is the statistics of co-occurrence of terms. That is, the dependence with which we are really concerned is state value $(\delta_i, \delta_j) = (1, 1)$ of term pair (t_i, t_j). In this case, what we need is to apply only the first item of $I(\delta_i; \delta_j)$ and to verify the constraints given in (2). For instance, for Method 1, we need only use the measure $\mathbf{mit}_d(t_i, t_j)$ and verify the condition:

$$P_d(\delta_i = 1, \delta_j = 1) = P_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\varpi}$$
$$= \gamma_d(t_i, t_j) > \psi_d(t_i) \psi_d(t_j) = \frac{f_d(t_i)}{||d||} \cdot \frac{f_d(t_j)}{||d||}$$

to ensure that t_i and t_j are highly dependent under their co-occurrence.

Fifth, from a high expected mutual information value, we cannot state immediately that state value $(\delta_i, \delta_j) = (1, 1)$ makes a larger contribution to $I_\Xi(\delta_i; \delta_j)$ and, thus we cannot assert that terms t_i and t_j are highly dependent for their co-occurrence in Ξ. This is because, with the relations of the MIT measures learned from their signs, when $\gamma_\Xi(t_i, t_j) < \psi_\Xi(t_i)\psi_\Xi(t_j)$, the positive value $I_\Xi(\delta_i; \delta_j)$ will be dominated by the positive quantities $\mathbf{mit}_\Xi(t_i, \bar{t}_j)$ and $\mathbf{mit}_\Xi(\bar{t}_i, t_j)$. In this case, the higher value the $I_\Xi(\delta_i; \delta_j)$ has, the larger quantities the $\mathbf{mit}_\Xi(t_i, \bar{t}_j)$ and $\mathbf{mit}_\Xi(\bar{t}_i, t_j)$ provide, the more it is indicated that t_i and t_j are highly dependent under state values $(1, 0)$ and $(0, 1)$ and that they should not co-occur in Ξ. We can clarify our viewpoint by an example given in [**?**]. Let us consider Method 4 and suppose $\Xi = \{d_1, d_2, d_3\}$, $V_{d_1} = \{t_1, t_2, t_3, t_4, t_5\}$, $V_{d_2} = \{t_1, t_4, t_5, t_7\}$ and $V_{d_3} = \{t_4, t_7, t_8\}$. Then, we have: $n_\Xi(t_1) = 2$, $n_\Xi(t_2) = 1$ and $n_\Xi(t_1, t_2) = 1$; $n_\Xi(t_5) = 2$, $n_\Xi(t_7) = 2$ and $n_\Xi(t_5, t_7) = 1$. Thus, we obtain (details of computation can be found in [**?**])

$$I_\Xi(\delta_1; \delta_2) \approx 0.1352 - 0.0959 - 0.0000 + 0.1352 = 0.1745,$$
$$I_\Xi(\delta_5; \delta_7) \approx -0.0959 + 0.1352 + 0.1352 - 0.0000 = 0.1745.$$

Clearly, the positive value of $I_\Xi(\delta_1; \delta_2)$ is dominated by both quantities $\mathbf{mit}_\Xi(t_1, t_2)$ and $\mathbf{mit}_\Xi(\bar{t}_1, \bar{t}_2)$, and t_1 and t_2 are highly dependent for their co-occurrence and co-not-occurrence in set Ξ; the positive value of $I_\Xi(\delta_5; \delta_7)$ is dominated by both $\mathbf{mit}_\Xi(t_5, \bar{t}_7)$ and $\mathbf{mit}_\Xi(\bar{t}_5, t_7)$, and t_5 and t_7 are highly dependent for their not co-occurrence in set Ξ. In addition, from this example, we can see that term pairs (t_1, t_2) and (t_5, t_7) have the same expected mutual information and, however, that the implications of for the individual state values are entirely different: Terms t_1 and t_2 provide the information highly supporting for both their co-occurrence and co-not-occurrence; whereas terms t_5 and t_7 provide the information highly supporting for occurrence of one but not occurrence of the other. It should be repeatedly pointed out that all the five different measures, the four MIT measures and the EMIM measure, may give us useful information, but each tells us different aspects about the dependences of terms and, in particular, it is likely that $I_\Xi(\delta_i; \delta_j)$ tells us nothing about the dependences of terms for their co-occurrence.

Sixth, it is worth mentioning that many studies use the following formula:

$$I(t_i; t_j) = P(t_i, t_j) \log \frac{P(t_i, t_j)}{P(t_i)P(t_j)}$$

to estimate the mutual information of terms t_i and t_j. It is 'equivalent' to the MIT measure for the state value $(\delta_i, \delta_j) = (1, 1)$ given in Definition 3.1,

$$\begin{aligned} \mathbf{mit}(t_i, t_j) &= \mathbf{mit}(t_i^{\delta_i=1}, t_j^{\delta_j=1}) \\ &= P(\delta_i = 1, \delta_j = 1) \log \frac{P(\delta_i = 1, \delta_j = 1)}{P(\delta_i = 1)P(\delta_j = 1)} \end{aligned}$$

as we denote $t^\delta = t, \bar{t}$ when $\delta = 1, 0$, respectively. The expression $I(t_i; t_j)$ seems simpler to that of $\mathbf{mit}(t_i, t_j)$. However, we point out, mathematically, that $\mathbf{mit}(t_i, t_j)$ is more appropriate and clearer than $I(t_i; t_j)$ from, for instance, a viewpoint of the probability space: It is obvious to see that $P(\delta_i, \delta_j)$ is over $\Omega \times \Omega$ as its each argument $\delta \in \Omega = \{0, 1\}$, whereas it is easy to cause confusion that $P(t_i, t_j)$ is over $V \times V$ as each of its arguments has a domain $t \in V = \{t_1, t_2, ..., t_n\}$ (rather than $t \in \{0, 1\}$). Also, $I(\cdot\,;\cdot)$, when used to expressed EMIM, is a traditional mathematical symbol, which is the summation of four items (rather than only one) corresponding to four state value pairs of each term pair.

Seventh, it is worth mentioning that there are five information measures widely used in the literature for computing term dependence (or, relatedness): directed divergence [**?**], divergence [**?**], information radius [**?**], Jensen difference [**?**] and the expected mutual information (i.e., EMIM, which is regarded as a special case of directed divergence) [**?**]. The five measures, which are what are generally called *information gain*, are by now familiar to many researchers. A detailed account of the concept of the measures is given in [**?**], and an axiomatic characterization can be found in [**?**]. The five measures are examined in our series of studies: Study [**?**] develops the measurement of term relatedness using the information radius measure, demonstrates how the relatedness measures may deal with some basic concepts of applications, and summarizes important features of, and differences between, the information radius measure and the first two information measures (directed divergence and divergence), from a practical perspective. Study [**?**] addresses the measurement of term relatedness based on the Jensen difference measure and points out, when Shannon entropy is used, that the Jensen difference measure is in fact the information radius measure, and that some formal methods proposed in many past studies in terms of these two measures are in principle the same matter. Study [**?**] proposes a method for estimating probability distributions required in EMIM, and provides examples to illustrate the possibility of failure of applying this method if the verification conditions are not satisfied. Study [**?**] reconsiders the *emim* measure, which is widely used in applications, derived from simplifying EMIM under a binary assumption, and discusses some potential but important problems of applying the *emim* measure. Study [**?**] attempts to establish a unified theoretical framework for applying several information measures to the measurement of term discrimination information and to define relatedness measures according to the discrimination measures, and then discusses some potential problems arising from using the relatedness measures and suggests solutions.

Finally, we would like to point out that the current study is further work of study [**?**], [**?**]: it focuses on the establishment of a general framework for constructing estimation functions in order to define probability distributions required in EMIM for effectively distinguishing potentially dependent term pairs from many others. As this paper concentrates on a formal analysis and discussion, the reader interested in how the mutual information methods, as well as other information measures' methods, may be supported by empirical evidence drawn from a number of performance experiments is referred to those papers referenced.

VI. Conclusions

This study focused on the establishment of a general framework for defining probability distributions required in EMIM, which is crucial and remains an open issue, for effectively distinguishing potentially dependent term pairs from many others. Under the framework,

- the general forms of estimation functions with a set of constraints were introduced;
- the general forms of probability distributions under term state values were defined;
- the general form of MIT measures for computing the mutual information of terms was formalised;
- the general properties of the MIT measures were studied and the general relations between the MIT measures were revealed.

Four estimation methods were proposed to clarify and illustrate our ideas presented in this study by

- interpreting the mathematical meanings of the estimation functions within practical application contexts;
- discussing verification conditions for satisfying the constraints in order to ensure that probability distributions meet the three criteria;
- presenting the properties and relationships of the MIT measures given in the individual methods.

The key points of this study were pointed out and emphasised, some of them are:

- The different implications of the dependence obtained from the individual MIT measures and the EMIM measure should be carefully distinguished from one another.
- The estimation functions should be constructed using weighting functions capable of capturing the occurrence and co-occurrence information of terms.
- It is possible of failure of using the estimation functions to define probability distributions if the constraints are not satisfied.

Under the general framework, the probability distributions, when defined from the estimation functions satisfying the constraints, will meet the three criteria. Thus, the issue of defining the probability distributions becomes the issue of constructing the estimation functions and verifying the constraints, which is relatively simple for practical applications. Due to its generality, the general framework is applicable to many areas of science, involving statistical semantic analysis of features (concepts, terms, phrases, words, etc.) and quantitative representations of objects (documents, abstracts, sentences, queries, etc.).

References

[1] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell System and Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[3] A. Akadi, A. Abdeljalil El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, no. 4, pp. 116–121, 2008.

[4] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168–1174, 2010.

[5] H.-W. Liu, J.-G. Sun, L. Liu, and H.-J. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, pp. 1330–1339, 2009.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[7] G. Wang, F. Lochovsky, and Q. Yang, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2004, pp. 342–349.

[8] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movies," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo (ICME'06)*, 2006, pp. 1013–1016.

[9] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[10] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Journal of the American Society for Information Science*, vol. 16, no. 1, pp. 22–29, 1990.

[11] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," in *Proceedings of the 29th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 115–122.

[12] S. Gauch, J. Wang, and S. M. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 250–269, 1999.

[13] M. Kim and K. Choi, "A comparison of collocation-based similarity measures in query expansion," *Information Processing & Management*, vol. 35, no. 1, pp. 19–30, 1999.

[14] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, 2000.

[15] D. Cai and T. McCluskey, "A simple method for computing term mutual information," *Journal of Computing*, vol. 4, no. 6, pp. 1–6, 2012.

[16] R. M. Losee, Jr., "Term dependence: A basis for Luhn and Zipf models," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 12, pp. 1019–1025, 2001.

[17] D. Cai, "Reconsideration of potential problems of applying EMIM measure for text analysis," *Journal of Computing (accepted)*, 2014.

[18] ——, "Determining semantic relatedness through the measurement of discrimination information using Jensen difference," *International Journal of Intelligent Systems*, vol. 24, no. 5, pp. 477–503, 2009.

[19] D. Cai and C. J. van Rijsbergen, "Learning semantic relatedness from term discrimination information," *Expert Systems with Applications*, vol. 40, no. 1, 2008.

[20] D. Cai, "An information theoretic foundation for the measurement of discrimination information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1262–1273, 2010.

[21] R. Sibson, "Information radius," *Z. Wahrsch'theorie and verw. Geb*, vol. 14, pp. 149–160, 1969.

[22] C. R. Rao, "Diversity: Its measurement, decomposition, apportionment and analysis," *Sankhya: Indian Journal of Statistics*, vol. 44, pp. 1–22, 1982.

[23] A. Rényi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1961, pp. 547–561.