Lecture Notes in Electrical Engineering 488

Rayner Alfred · Hiroyuki Iida
Ag. Asri Ag. Ibrahim · Yuto Lim
Editors

Computational Science and Technology

4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017

Springer

Lecture Notes in Electrical Engineering

Volume 488

About this Series

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, MetaPress, Springerlink ****

Lecture Notes in Electrical Engineering (LNEE) is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering
- Engineering

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer's other Lecture Notes series, LNEE will be distributed through Springer's print and electronic publishing channels.

For general information about this series, comments or suggestions, please use the contact address under "service for this series".

To submit a proposal or request further information, please contact the appropriate Springer Publishing Editors:

Asia:

China, *Jessie Guo, Assistant Editor* (jessie.guo@springer.com) (Engineering)

India, *Swati Meherishi, Senior Editor* (swati.meherishi@springer.com) (Engineering)

Japan, *Takeyuki Yonezawa, Editorial Director* (takeyuki.yonezawa@springer.com)
(Physical Sciences & Engineering)

South Korea, *Smith (Ahram) Chae, Associate Editor* (smith.chae@springer.com)
(Physical Sciences & Engineering)

Southeast Asia, *Ramesh Premnath, Editor* (ramesh.premnath@springer.com)
(Electrical Engineering)

South Asia, *Aninda Bose, Editor* (aninda.bose@springer.com) (Electrical Engineering)

Europe:

Leontina Di Cecco, Editor (Leontina.dicecco@springer.com)
(Applied Sciences and Engineering; Bio-Inspired Robotics, Medical Robotics, Bioengineering; Computational Methods & Models in Science, Medicine and Technology; Soft Computing; Philosophy of Modern Science and Technologies; Mechanical Engineering; Ocean and Naval Engineering; Water Management & Technology)

(christoph.baumann@springer.com)
(Heat and Mass Transfer, Signal Processing and Telecommunications, and Solid and Fluid Mechanics, and Engineering Materials)

North America:

Michael Luby, Editor (michael.luby@springer.com) (Mechanics; Materials)

More information about this series at http://www.springer.com/series/7818

Rayner Alfred · Hiroyuki Iida
Ag. Asri Ag. Ibrahim · Yuto Lim
Editors

Computational Science and Technology

4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017

Springer

Editors
Rayner Alfred
Knowledge Technology Research Unit,
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu
Malaysia

Hiroyuki Iida
School of Information Science
Japan Advanced Institute of Science
and Technology
Nomi, Ishikawa
Japan

Ag. Asri Ag. Ibrahim
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu
Malaysia

Yuto Lim
School of Information Science, Security
and Networks Area
Japan Advanced Institute of Science
and Technology
Nomi, Ishikawa
Japan

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-10-8275-7 ISBN 978-981-10-8276-4 (eBook)
https://doi.org/10.1007/978-981-10-8276-4

Library of Congress Control Number: 2018933372

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature
Singapore Pte Ltd. part of Springer Nature
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Computational science and technology is a rapidly growing multi- and interdisciplinary field that uses advanced computing and data analysis to understand and solve complex problems. The absolute size of many challenges in computational science and technology demands the use of supercomputing, parallel processing, sophisticated algorithms, and advanced system software and architecture. The ICCST17 conference provides a unique forum to exchange innovative research ideas and recent results and share experiences among researchers and practitioners in the field of advanced computational science and technology.

Building on the previous three conferences that include Regional Conference on Computational Science and Technology (RCSST 2007), the International Conference on Computational Science and Technology (ICCST 2014), and the Third International Conference on Computational Science and Technology 2016 successful meetings, the Fourth International Conference on Computational Science and Technology (ICCST17) program offers practitioners and researchers from academia and industry the possibility to share computational techniques and solutions in this area, to identify new issues, and to shape future directions for research, as well as to enable industrial users to apply leading-edge large-scale high-performance computational methods. This volume presents a theory and practice of ongoing research in computational science and technology. The focuses of this volume is on a broad range of methodological approaches and empirical reference points including artificial intelligence, cloud computing, communication and data networks, computational intelligence, data mining and data warehousing, evolutionary computing, high-performance computing, information retrieval, knowledge discovery, knowledge management, machine learning, modeling and simulations, parallel and distributed computing, problem-solving environments, semantic technology, soft computing, system-on-chip design and engineering, text mining, visualization and Web-based and service computing. The carefully selected contributions to this volume were initially accepted for oral presentation during the Fourth International Conference on Computational Science and Technology (ICCST17) held on November 29–30, 2017, in Kuala Lumpur, Malaysia. The level of contributions corresponds to that of advanced scientific works, although several

of them could be addressed also to non-expert readers. The volume brings together 43 chapters.

In concluding, we would also like to express our deep gratitude and appreciation to all the program committee members, panel reviewers, organizing committees, and volunteers for your efforts to make this conference a successful event. It is worth emphasizing that much theoretical and empirical work remains to be done. It is encouraging to find that more researches on computational science and technology are still required. We sincerely hope the readers will find this book interesting, useful, and informative and it will give then a valuable inspiration for original and innovative research.

Contents

Sequential and Global Learning Styles as Pathways to Improve Learning in Programming

Sin-Ban Ho[1(✉)], Sek-Kit Teh[1], Gaik-Yee Chan[1], Ian Chai[1], and Chuie-Hong Tan[2]

[1] Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
{sbho,gychan,ianchai}@mmu.edu.my, nicholas.teh93@gmail.com
[2] Faculty of Management, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
chtan@mmu.edu.my

Abstract. Programming knowledge is increasingly important to facilitate code reuse. Nevertheless, comprehending another programming language is not simple because of its complexity and clarification needs. Prior work focused on different learning styles to aid programming, but it was important to identify which ones were more effective. This research highlights findings in assessing the different documentation styles, including sequential and global documentation styles. Organizing an observation of 125 intermediate undergraduates participated in cloud hosting computation and file content programming exercises, this empirical investigation revealed that sequential documentation exhibits a positive impact in obtaining programming knowledge, significantly pertaining faster completion time, higher multiple choice comprehension, and fewer difficulties. This concludes that sequential documentation solutions can lead intermediate undergraduates with sequential learning styles to faster growth in gaining programming knowledge.

Keywords: Knowledge management · e-Learning analytics
Modeling and simulations · Learning style · Documentation

1 Introduction

Programming knowledge is important in Computer Science. Since object-orientation was introduced, object-oriented programming languages like Python [1, 2] have emerged as a major way of organizing object-oriented code for reuse. Object-oriented programming (OOP) comprises of a set of classes that work together to solve problems in a domain.

Another way of looking at OOP knowledge is to think of them as prefabricated parts. When one has prefabricated parts, one can put together a new product much faster than if one had to build every piece from scratch. However, this also means that one must know how the parts are intended to be used [3, 4]. The crucial problem is that a developer who is new to a programming language may not be aware of the internal structure of the design. Due to this, acquiring OOP knowledge from documentation often has a steep learning curve [2, 4].

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 1–10, 2018.
https://doi.org/10.1007/978-981-10-8276-4_1

2 Motivation of the Study

Literature on pedagogical documentation has progressed rapidly in recent years. Each philosophy applies different models in mixing examples, texts and diagrams. Applications of cloud hosting and file contents alphabetizing are chosen as the basis for our study using Python [5–8]. Hence, this stage of research work is to find the outcome of novices in pursuing the beginners' level stage of programming in the context of the Python programming language.

Two documentation styles are proposed in this paper, the sequential and global learning styles [9, 10]. These documentation styles are aimed to provide exemplars in learning and instruction that guide developers how to build applications using an object-oriented scripting technology.

The central idea of the first style, the sequential style, is that people learn best when their own needs and interests direct their learning. The concepts, examples, and test are arranged immediately and directly one after another (see Fig. 1). For this reason, sequential documentation presents instructions in small chunks and allows readers to choose the order in which to read them, based on what seems important to them, although they also do often have a particular starting point. Each page in sequential documentation web pages usually links to other pages for related tasks or information that the reader might need in order to complete the task at hand.

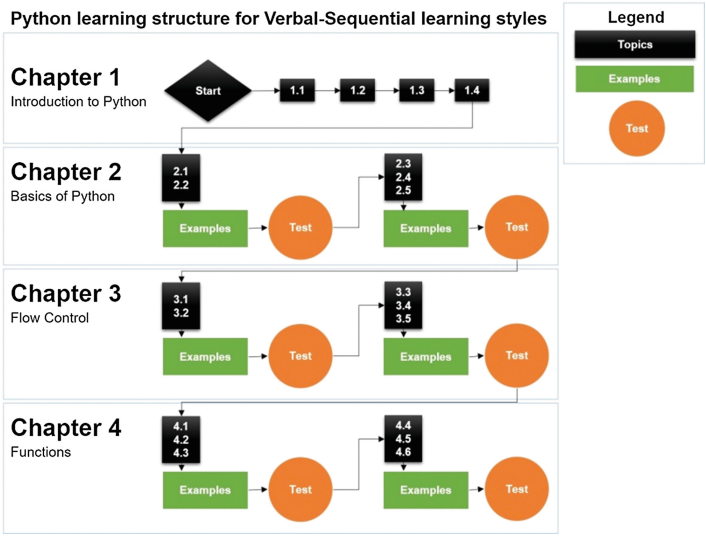


Fig. 1. Python learning structure for sequential documentation style

Sequential documentation generally follows these guidelines [9–12]: Training in linear steps: motivate with following linear stepwise paths to find solutions. Absorbing logically and directly connected pieces: testing them immediately after each concept chunk and example. Reasoning and improvising: instead of the trouble of relating to

numerous different aspects of the same course or to different courses, let people be challenged with something relevant, so that they may know a lot about specific aspects of a course. Coordinating related components: instead of presenting ambiguous and lengthy steps, allow their readers to progress on the task interactively. Supporting error recognition: do not assume people will follow your instructions flawlessly; expect mistakes and give resources to overcome them. Exploiting prior knowledge: use analogies and avoid jargon. This mirrors the findings of [3] regarding patterns documentation, [11] in respect to minimalist documentation, and not to be confused with step-by-step documentation [4].

The second style, global documentation, originates from the concept of learning in large leaps, i.e., a holistic thinking process [9]. A whole picture is gained after absorbing the learning material almost randomly [10]. The rationale is to learn enough material without jumping into too much detail in a particular aspect of a course. The Global style guides learners to find connections among different areas, where the whole picture is presented first, as shown in Fig. 2. We use the situation of giving a big picture to build the learning materials based on the expectations of the audience.

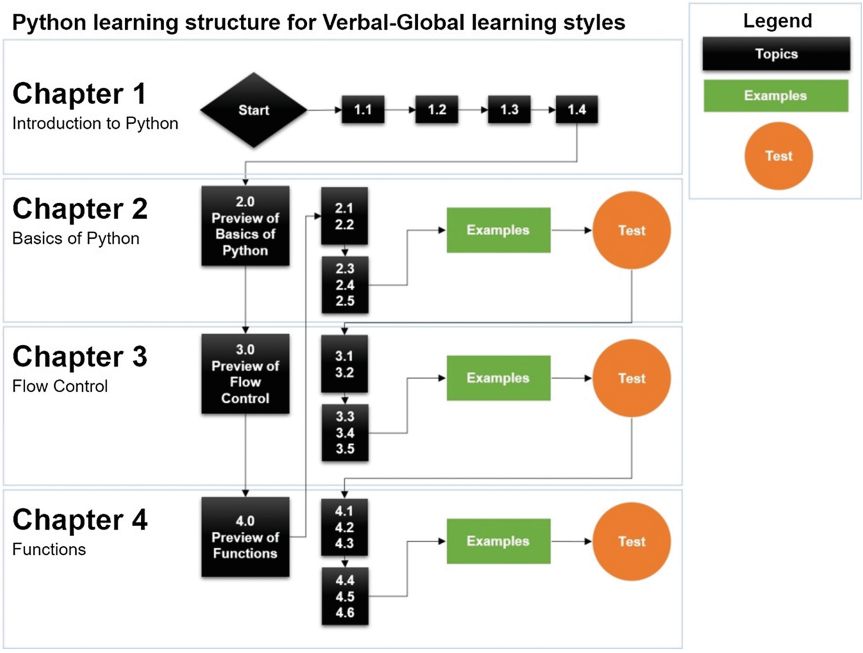


Fig. 2. Python learning structure for global documentation style

In contrast to the sequential style of Fig. 1, the global style in Fig. 2 presents the previews of each chapter, i.e. Chapter 2 preview, Chapter 3 preview, and subsequently, Chapter 4 preview before navigating back to the detailed topics of Chapter 2, i.e. topics 2.1, 2.2, 2.3, and so on. The sequential documentation in Fig. 1, on the other hand, presents each topic linearly, i.e. in the sequence of topics 2.1 and 2.2, followed by

chunked pieces of examples and the test. The small tests are conducted immediately after each aspect is covered. For example, in Fig. 1, there is a small test after topics 2.1 and 2.2, and another small test after topics 2.3, 2.4, and 2.5. In contrast, the global style in Fig. 2 emphasizes a more major test only after all topics of a chapter are covered. For instance, there is a major test of Chapter 2 only after all topics 2.1, 2.2, 2.3, 2.4, and 2.5 are presented in the global style. With this, we observe two main differences between Figs. 1 and 2. Firstly, the preview order and secondly, the grouping of major tests in the global style.

3 Experiments

The sample sizes (number of students) in the two groups are different due to the different class sections arrangement. Furthermore, the exercise-based investigation was conducted in two different semesters. Only one type of the documentation set is uniquely presented to the participants. The formulated hypothesis guided us to test out the documentation sets first for their readability, soundness, and usability. After that, we rolled them out to collect the field data. We analyzed the collected data through suitable statistical analysis techniques.

3.1 Documentation Procedure

The students would use the documentation and write Python source code, which allows faster application development than programming languages such as C++. The tasks outcome would have a running cloud hosting and file content alphabetization applications.

An idea of the sequential documentation is organized in [13]. The whole procedure of the exploratory study proceeded with the additional background information section amended at the beginning of each piece of subtask to formulate the global documentation [14]. Furthermore, the small chunked examples and tests are integrated into a larger example and test towards the end of each chapter in the global style.

Table 1 shows the documentation quantitative characteristics, including their relative length measured in bytes, as supported by Beizer [15]. The documents in the experiment are quantitatively characterized through this methodology.

Table 1. Quantitative characteristics of the documentation

Relative characterization	Sequential	Global
1. Total length (kilobytes, KB)	430 KB	436 KB
2. Information that is relatively available	Chunked pieces of tests after every topic	Background information in overview and collated test
3. Total document files	28 files	29 files
4. Number of sections	10 sections	7 sections
5. Number of paragraphs	48 paragraphs	38 paragraphs

3.2 Hypothesis

We use standard significance testing to explicitly determine the impacts of the documentation styles. We state the hypothesis as follows.

$E1H_0$ - There is no difference between sequential and global documentation for the users in performing the given Python exercise. We derive the interpretations through this hypothesis rejection or non-rejection.

3.3 Experimental Design

The experimental design consists of a factor (independent variable), and seven dependent variables. The factor is the documentation philosophy. Meanwhile, the dependent variables refer to the semi completion time, completion time, comprehension of the exercise, workings, and total difficulties faced.

Factor:

- **Documentation style (*doctype*):** Two documentation groups are used, which were mentioned in Sect. 2, each having a similar aim in completing the designated exercise.

Dependent variables:

- **Semi Completion time (*semiTime*):** Duration for the participants to complete their first cloud hosting programming.
- **Completion time (*complTime*):** Duration to complete the entire work task.
- **Comprehension:** The participants need to determine the coding variables, line of code, and function, which fulfill the given exercise. Several multiple choices questions (***comprMcq***) and structured questions (***comprStruc***) are given to evaluate their code understanding.
- **Workings:** This variable assesses how well the subjects followed the instructions to assign default settings for the cloud hosting computation (***workingHost***), and file content alphabetizing (***workingFile***).
- **Total difficulties faced (*totalDiff*):** Certain documentation parts let the participants discover solutions with the documentation. The participants subsequently record the number of difficulties they encountered.

3.4 Participants

This study involves 125 participants, who have spent between two years to four years undergraduate at the university. They are computer science university students who pursue the software evolution course in the university. The average student age is 22 years old. Two different documentation groups are needed to assess our experiment hypothesis. Since the two different groups are organized according to their laboratory sessions, the number of students in each group is different.

The lectures provide the students basic principles of software evolution and OOP. The practical laboratory sessions supplemented the lectures so that the participants have the chance to practice what they have learnt via the numerous coding tasks through the on-line documentation. The participants represented intermediate undergraduates as

they have mostly undergone the previous courses such as Software Engineering Fundamentals (*SEf*), pre-Python Foundation course (*prePython*), and Data Structures and Algorithms (*DataStruct*). As they are still pursuing the software evolution course, they are not regarded as advanced users.

3.5 Validity

To understand the two groups further, we collect the participants' grades in previous programming courses. We categorized the grade into zero point ('None'), when the participant has not attempted the course, one point ('F') for fail, two points ('C') for poor, three points (grade 'B') for average, and four points (grade 'A') for best score. Furthermore, we also considered the grade intervals, i.e. 2.33 points for 'C+', 2.67 points for 'B−', 3.33 points for 'B+', and 3.67 points for 'A−'.

Table 2 depicts the results of Pearson Chi-square performed, where there is no significant difference in the two documentation groups academically, since all the *p*-values > 0.050. As such, the two groups are balanced pertaining the courses that they had taken in prior semesters, such as *SEf, DataStruct*, *prePython*, and Cumulative Grade Point Average (*CGPA*).

Table 2. Pearson Chi-square tests results of previous achievement of *SEf*, *DataStruct*, *prePython*, and *CGPA*.

Documentation group	Sequential	Global	*p*-value
N (participants)	83	42	not applicable
SEf: Mean (Std. Dev.)	3.01 (0.721)	2.91 (0.543)	0.161
DataStruct: Mean (Std. Dev.)	2.67 (1.213)	2.00 (1.372)	0.115
prePython: Mean (Std. Dev.)	0.08 (0.489)	0.13 (0.862)	0.173
CGPA: Mean (Std. Dev.)	2.97 (0.498)	2.74 (0.450)	0.414

4 Data Analysis and Results

We conduct statistical analyses via the Statistical Package for Social Science (SPSS) on the 125 responses gathered. The data is analyzed in determining which group let the subjects compute cloud hosting (semiTime) and finish the fastest (complTime), as well as comprehend the most in answering multiple choice questions (comprMcq) and structured questions (comprStruc). We accumulate the number of difficulties recorded at intervals (totalDiff). We also collect test scores in their inner workings knowledge of the cloud hosting computation (workingHost), and file content alphabetization (workingFile). We evaluate the dependent variables normality to avoid assuming their distribution is normal. Table 3 shows the normality test result, where two dependent variables, namely semiTime and complTime are normally distributed, with p-values > 0.050. Hence, rather than the means, we use medians as the expected values for the other dependent variables.

Table 3. Results of Kolmogorov-Smirnov normality test

Category	*p*-value	Category	*p*-value
1. *semiTime*	0.061	5. *workingHost*	0.000*
2. *complTime*	0.112	6. *workingFile*	0.000*
3. *comprMcq*	0.001*	7. *totalDiff*	0.000*
4. *comprStruc*	0.012*		

Note: * Statistically significant at 0.050 level (with $p < 0.050$)

In Table 4, some items of the sequential column are bold-faced to show that this group has better performance than global style. In order to discuss more detail in the results of Table 4, for example, in ***complTime***, the sequential group took 21 min 53 s to complete the exercise. Meanwhile, the global group took a longer time of 38 min 19 s to complete the similar exercise. On the other hand, in terms of comprMcq, the sequential group has a better median of 3.00 correct answers (out of 5), compared to the global group, which has only a median of 2.50 correct answers.

Table 4. The categories descriptive statistics

Dependent variable (Category)	Mean		Std. Dev.	
	Seq.	Global	Seq.	Global
1. *semiTime* (hh:mm:ss)	**0:11:30**	0:18:52	0:06:00	0:10:21
2. *complTime* (hh:mm:ss)	**0:21:53**	0:38:19	0:10:30	0:19:11
	Median		Std. Dev.	
	Seq.	Global	Seq.	Global
3. *comprMcq* (scale: 0-6)	**3.00**	2.50	1.695	1.656
4. *comprStruc* (scale: 0-5)	**3.00**	2.50	1.529	1.630
5. *workingHost* (scale: 0-8)	8.00	8.00	0.472	0.821
6. *workingFile* (scale: 0-7)	7.00	7.00	0.469	0.790
7. *totalDiff*	**1.00**	2.00	2.950	2.600

Table 5 indicates the separate multivariate tests results. We conducted these F-tests to see the specific dependent variables that are significant across the categories. We obtained the p-values through between-subjects effects tests via MANOVA (Multivariate Analysis of Variance) [16]. These results, through Wilks' Lambda = 0.689, $F(2,122) = 27.472$ ($p < 0.001$), implied highly significant differences in the mean scores.

Table 5. Multivariate effects of the documentation style on dependent variables.

Category	F	p-value	Category	F	p-value
1. *semiTime*	25.252	0.000**	2. *complTime*	53.487	0.000**

Note: ** Statistically significant at 0.050 level with $p < 0.050$ (2-tailed)

Pertaining semiTime and complTime within Table 4, the sequential participants complete their entire task faster than the other global group. In Table 5, when we look for 0.050 standard significance level (95% probability), the sequential group gives the

evidence of being much faster for semiTime and complTime. Participants using sequential are faster more significantly than the global especially on the Python topic in this exercise, which do not have advanced pointers concept as in C++ programming language. To generalize the results throughout the whole Python programming content chapters, a more technical understanding such as arrays can be considered as our future work.

Since the subsequent five dependent variables are not normally distributed over the comparison of two groups, the non-parametric Mann-Whitney test [17] is used. In Table 6, with the p-values more than 0.05, comprStruc, workingHost, and workingFile have no significant differences between the two groups. The participants worked well in assigning default settings irrespective of which type of documentation was presented to them.

Table 6. Mann-Whitney (MW) test results on the categories.

Categories	Mean rank					
	Sequential	Global	*MW-U*	*Z*	*Wilcoxon W*	*p*-value
1. *comprMcq*	69.39	50.37	1212.500	−2.817	2115.500	0.005**
2. *comprStruc*	62.54	63.92	1704.500	−0.205	5190.500	0.838
3. *workingHost*	64.22	60.60	1642.000	−0.828	2545.000	0.408
4. *workingFile*	64.23	60.56	1640.500	−1.139	2543.500	0.255
5. *totalDiff*	55.00	78.81	1079.000	−3.574	4565.000	0.000**

Note: ** Statistically significant at 0.050 level with $p < 0.050$ (2-tailed)

Regarding comprMcq and totalDiff, the participants in the sequential group indicate significantly better outcome than the global group at the 5 per cent level. Therefore, this supports that the $E1H_0$ hypothesis for these variables in Sect. 3 is rejected. This rejection means that the sequential and global groups are different in promoting learning to the participants. Most of the undergraduates are often from the sequential learning style, as found by [18]. As such, the sequential documentation style suits the majority intermediate undergraduates, who typically have a sequential learning style.

5 Conclusion

In summary, the strong measurements showing the shorter duration of semi completion time, the faster completion time, with the higher multiple choice questions (MCQ) comprehension, and fewer total difficulties faced confirm the advantage of the sequential documentation. Based on our experiment, we discover that Python programming learners perform better in terms of knowledge acquisition using sequential documentation compared to global learners. As majority of the participants exhibit sequential learning characteristics, we can conclude that students' learning performance is related to their learning styles. Most of the undergraduates are often from the sequential learning style, as found by [18].

Therefore, knowing the learning styles of each student can help identify their learning preference, which can eventually be utilized in instructional documentation to improve

students' learning performance. The aim of suiting documentation according to the students' learning style is to harness their learning performance. This will have a positive impact on the result of less text context preferable with sequential documentation, which is shown effective at least in the context of basic Python programming tasks.

For future work, we can consider advanced topics of Python as a whole, and use Structural Equation Modelling (SEM) for data analysis. SEM refers to a multivariate statistical technique, which aims to explain further the relationship of multiple variables. The main benefit of SEM over other multivariable techniques is its ability to examine a series of dependence relationships simultaneously [19–21]. With this, SEM allows all hypothesized relationships to fit together into a single model so that they can be simultaneously evaluated. This can give higher accuracy than individual causal path testing of the proposed pathways to acquire effective programming knowledge.

Acknowledgments. This research work was financially supported by the Fundamental Research Grant Scheme, FRGS/1/2015/SS06/MMU/02/1.

References

1. Schneider, D.I.: An Introduction to Programming Using Python, pp. 208–237, 299–315. Pearson Education Limited, Harlow (2016)
2. Lutz, M.: Learning Python, 5th edn, pp. 862–868. O'Reilly Media Inc., Sebastopol (2013)
3. Chai, I.: Pedagogical framework documentation: how to document object-oriented frameworks: an empirical study. PhD dissertation, University of Illinois at Urbana-Champaign, IL (2000). http://www.cs.uiuc.edu/research/techreports.php?report=UIUCDCS-R-99-2077. Accessed 25 Sep 2017
4. Ho, S.B.: Framework documentation with patterns: an empirical study. PhD thesis, Multimedia University, Cyberjaya, Selangor, Malaysia (2008)
5. Gaddis, T.: Starting Out with Python, 3rd edn. Pearson Education Limited, Upper Saddle River (2015)
6. Unpingco, J.: Python for Probability, Statistics, and Machine Learning. Springer, Switzerland (2016)
7. Nelli, F.: Python Data Analytics, pp. 13–34. Springer, New York (2015)
8. Briggs, J.R.: Python for Kids: a Playful Introduction to Programming, pp. 193–217. No Starch Press Inc., San Francisco (2013)
9. Felder, R.M., Spurlin, J.: Applications, reliability and validity of the index of learning styles. Int. J. Eng. Educ. **21**(1), 103–112 (2005)
10. Graf, S., Viola, S.T., Kinshuk: In-depth analysis of Felder-Silverman learning style dimensions. J. Res. Technol. Educ. 40(1), 79–93 (2007)
11. Carroll, J.M.: Minimalism Beyond the Nurnberg Funnel. MIT Press, Cambridge (1998)
12. Dollmat, K.S., Ho, S.B., Chai, I.: A minimalist approach in creating a guide for Visual Basic 2010. In: Proc. IEEE Student Conference on Research and Development (SCOReD 2010), Kajang, Selangor, Malaysia, pp. 154–158 (2010). https://doi.org/10.1109/SCORED.2010.5703992
13. Example of the documentation fragment which was presented in the sequential documentation group. http://pesona.mmu.edu.my/~sbho/Pythoncmd/Pt1.1.3.html. Assessed 25 Sep 2017

14. Example of the documentation fragment that is available in the global style, but not available in the sequential documentation. http://pesona.mmu.edu.my/~sbho/Pyglobal. Assessed 25 Sep 2017
15. Beizer, B.: Software is different. In: Patel, D., Wang, Y. (eds.) Comparative Studies of Engineering Approaches for Software Engineering, vol. 10, pp. 293–310. Baltzer Science Publishers, Norwell (2000)
16. Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W.: Applied Linear Statistical Models. McGraw Hill, Boston (1996)
17. Leech, N.L., Barrett, K.C., Morgan, G.A.: IBM SPSS for Intermediate Statistics: Use and Interpretation, 5th edn. Routledge, Taylor & Francis, New York (2015)
18. Ho, S.B., Tan, C.H.: Local population: a study in the influence of learning styles in computing field. Aust. J. Basic Appl. Sci. **9**(22), 1–7 (2015)
19. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis, 7th edn. Pearson Prentice Hall, Upper Saddle River (2010)
20. Finch, W.H., Immekus, J.C., French, B.F.: Applied Psychometrics Using SPSS and AMOS. Information Age Publishing Inc., Charlotte (2016)
21. Ho, R.: Handbook of Univariate and Multivariate Data Analysis with IBM SPSS, pp. 421–505. CRC Press, Taylor & Francis Group, Boca Raton (2014)

Vulnerability Reports Consolidation for Network Scanners

Nicholas Ming Ze Lee, Shih Yin Ooi(✉), and Ying Han Pang

Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia
lmz.nicholas@gmail.com, {syooi,yhpang}@mmu.edu.my

Abstract. Vulnerability scanning is one of the vital process conducted by many penetration testers and security consultants as to assess the security of an organizational network. However, when multiple vulnerability scanners are used, reports of varied sources have to be compiled via manual means. It is an uncomplicated but lengthy process, where vulnerabilities of different reports have to be examined thoroughly in order to assess them. Thus, this paper describes an approach of creating a report consolidation tool in order to merge similar vulnerabilities and to unify results of differing scanner.

Keywords: Vulnerability management · Vulnerability merging · Scanner

1 Introduction

The ever-increasing popularity of Internet has not only made it a cornerstone for information sharing, but also paves way to numerous new opportunities. It has now become imperative, especially for organization to secure their network perimeters and any access points which could turn out to be the root of security breach. However, the amount of tests required to analyze and identify every misconfigurations and weaknesses in a system are too much of a work. Coupled with both internal and Internet-facing devices in the network, the amount increases exponentially. Thus, a software which is known as vulnerability scanner is used to automate the process of identifying potential security holes.

Vulnerability scanning is often conducted in a penetration test in order to quickly identify and quantify the exposure to weaknesses [1]. Thereafter, an exploitation will be attempted in order to access the risks associated with the vulnerabilities found. This will also enable the organization to pinpoint the more critical weaknesses and to provide suitable mitigations for them.

Due to the nature of the vulnerability, each scanner of differing vendor employs a different set of algorithms and are specialized in different types of use case. Hence, by using several vulnerability scanners altogether, one scanner could tackle the limitations of the other scanners. It is especially beneficial to smaller-scale companies, where the budget of leasing and purchasing a commercial scanner can be cut off greatly by utilizing a variety of open source and free vulnerability scanners. However, at the cost of using multiple scanners, a problem arises at the end of the scanning phase. Be it security

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 11–20, 2018.
https://doi.org/10.1007/978-981-10-8276-4_2

consultants, penetration testers, or a private individual, a report documenting the list of findings and methods is expected to be delivered. A tedious job would then be required to manually scrutinize dozens of reports.

Thus, this paper explores the work done on developing the proposed framework with the objective of parsing and merging vulnerabilities addressed from different network vulnerability scanners.

An overview and literature review on related topics are featured in Sect. 2. Section 3 contains the design of the proposed system. Section 4 presents the implementation result of the proposed system, followed by the conclusion of this paper in Sect. 5.

2 Background

2.1 Types of Vulnerability Scanner

Vulnerability scanner can be categorized into two groups – network and host based vulnerability scanner [2].

A host-based scanner is installed and runs from within the target host itself. This enables the scanner to access to low level data and is able to provide a greater insight on a vulnerable system. Some common risks and vulnerabilities that can be detected by host-based scanner include backdoors, non-compliant policies, weak passwords and inadequate file access permissions.

A network-based scanner scans and examines live systems over the network. Typically, an unabridged network vulnerability scanner is able to perform banner grabbing and scans for related vulnerabilities and misconfigurations. There are some scanners, however, specialized in handling specific task of a complete network scanning suite, such as the port scanner and application scanner [3].

The main functionality of a port scanner is to identify open and closed ports on the target. Some notable features included OS fingerprinting, services and applications identification, and version scanning. While port scanner is able to gather information of the target host, it does not detect nor identify any vulnerabilities. On the other hand, an application scanner is used to assess configurations and security features of specific application. Database servers and web applications are the more common applications being assessed.

Network-based scanning can be done easily as compared to a host-based scanning. A host-based scanning requires the installation of the scanner in every system to be assessed, while a network scanner can be launched remotely and assess multiple systems on the network. In certain cases, the merit of using host-based scanner is that, a greater level of security check can be performed, since a network scanner does not have a direct access to the file system of the target.

2.2 Architecture of a Vulnerability Scanner

Typically, a vulnerability scanner consists of four main components: user interface, scan engine, scan database and report module [4].

- User interface. It allows users to operate and configure the scanner, two main types of user interface are used by vulnerability scanners – graphical user interface (GUI) and command-line interface (CLI). Alternately, some scanners allow the use of its API (Application Programming Interface). By sending specially crafted messages, user is able to launch an automated scan easily without the need of user interface.
- Scan engine. The core module which, based on plugins installed, carries out specified scanning tasks. Some scanners identify vulnerabilities by detecting running services and its version, while others attempt exploitation on the scanned target.
- Scan database. Data stored in the database are used by scanner to aid in scanning and reporting. While contents of the database varied for each vendor, some of the more common ones are vulnerability information, configuration data, scanning results and Common Vulnerabilities and Exposures identifier (CVE-ID).
- Report module. Most Scanners now allow users to customize and sort the contents of scan result. From a high-level summary report to a detailed technical report, different levels of view and formats can be provided for employees of a different echelon.

2.3 Vulnerability Source and Database

Vulnerability feed and database plays an important role in unifying vulnerabilities from various technologies and keeping us informed firsthand of new security flaws. Security professionals and developers rely on these feeds to keep an organization or software application secure.

Maintained by MITRE Corporation, the Common Vulnerabilities and Exposures or CVE in short, is a dictionary of publicly known security vulnerabilities. It has become the de facto standard many organizations have referred to as a source for vulnerability intelligence [5]. Rather than being a vulnerability database itself, CVE uses a standardized identifier to facilitate data sharing and association across separate security tools and databases. As such, it does not include an in-depth technical details and solutions for each vulnerability [6].

One such vulnerability database which addresses the lack of analysis on the CVE entries is the NIST National Vulnerability Database (NVD). NVD data is freely available, it is built on top of CVE and synchronizes with every new vulnerability added to the CVE dictionary. Additional capabilities include a fine-grained search engine and enhanced information [7] for each CVE entry. The additional information include but are not limited to technical details, vulnerability fixes and CVSS severity.

Common Vulnerability Scoring System (CVSS) is a prevailing open industry standard used to calculate and produce a numerical score reflecting the severity of a vulnerability. The scores can then be used as a guidance for an organization to plan and prioritize on the vulnerabilities which poses a substantial amount of risk [8]. CVSS is made up of 3 metric groups [9], with each representing a different area of concern. The first group is the base metric, it reflects the intrinsic qualities of the vulnerability which remain unchanged with time and user environments. The temporal metric group otherwise, expresses the qualities that change with time such as the availability of exploit

techniques at present time. The environmental metric group then defines the qualities of the vulnerability based on the user's environment.

Recently in April 2016, the popular and comprehensive Open Sourced Vulnerability Database (OSVDB) announced the shutting down of its services [10]. OSVDB was around for 14 years, dedicated on cataloguing vulnerabilities of various types for non-commercial uses. The shutdown has causes some of the vulnerabilities to lose their only identifier, as well as references used in many of the security products such as Metasploit to point to a non-existing resource. Although a modest attempt on the replacement of the identifier has already begun, a coordinated effort would be required in order to build a complete historical vulnerability database [11].

2.4 Existing Free Vulnerability Management and Reporting Tools

Dradis Community Edition [12] is an open source vulnerability management and reporting tool. It is one of the more active and frequently updated project. The Community Edition is maintained by a small dedicated team and is sponsored by Security Roots. It uses plugin architecture to integrate more scanner tools, output file formats and customized functionalities. Managing of imported or custom data is also possible through the web interface of Dradis. Some templates for importing plugins are bundled with Dradis for easier creation and modification. The commercial version of Dradis is another alternative for a more consistent support and quality reporting.

Another free vulnerability reporting tool is the MagicTree by Gremwell [13]. Although the last update for the tool is year 2013 and has remained in the same version ever since. Most of its operations remain intact. MagicTree has main functionalities similar to that of Dradis', which includes aggregating vulnerabilities data, querying imported data, and generating report based on supplied templates. It also supports the execution of shell commands, and allows XML results to be combined together with the imported data. For example, Nmap port scanner can be executed and its result obtained directly from MagicTree. Besides, the notion that imported data are stored and organized in a tree structure is the reason why it is named MagicTree.

3 Solution Design

3.1 Process Flow

The program accepts vulnerability reports in XML format. Before they are parsed, plugins and templates which reside in respective repository are scanned and imported into the main program. Each plugin is created to associate to a specific nature of a report or a vulnerability scanner. In it contains detection signatures, parsing instructions and hexadecimal color code.

- Detection signatures comprise of a root tag and a list of XPath expressions. They are used as a green light to assign a particular plugin in handling specific vulnerability report.

- Parsing instructions are a set of Python codes responsible in obtaining relevant data from the assigned report.
- Hexadecimal color code is used by templates for output styling. Each color serves as an identity to distinguish which vulnerability scanner the data originated from.

Root tags are obtained from both supplied vulnerability reports and detection signature in plugins to perform comparison. If there's a match, further tests using XPath queries are executed to verify the report. A report will only be assigned to a plugin once all conditions are met.

The parsing instructions are then called from associated plugin to perform data extraction. Each report is parsed block by block in an iterative manner using approach published by IBM [14]. Subsequently, all previous nodes which are processed will then be freed. This is to ensure memory consumption is kept low instead of loading the entire report into memory. A simple pseudocode on parsing Nessus report is shown as below, which adheres to the structure [15].

```
iterate ReportHost:
  host_info.append(host_name)
  iterate HostProperties:
    host_info.append(host_ip, host_os)
    new dataframe(host_info)
  iterate ReportItem:
    temp_dict = {attributes}
    iterate ReportItem.child:
      if match element:
        temp_dict[element] = element.data
  dataframe.add(temp_dict)
  while ReportHost.getprevious():
    delReportHost.getparent()[0]
```

Extracted data are stored in a two-dimensional data frame structure with attributes as shown in Table 1. When all data are extracted from supplied reports, a cleanup process on the extracted data is initiated. Any redundant data found are removed and an attempt to merge vulnerabilities of different scanners is made.

Thereafter, user-selected template is executed to build the consolidated report. Several common templates have been created for this project at the time of writing, such as complete database dump in CSV format, and sort based on vulnerabilities and scanners in Microsoft Word Document format.

Table 1. General attributes for each instance of data.

Attribute	Description
Name	Vulnerability name
Port	Port and protocol used
Reference	Reference identifier related to found vulnerability
Reference URL	URL for the reference
Severity	Severity score
CVSS score	CVSS base or temporal score
Risk	Risk factor ranging from low to critical
Summary	Short summary on the vulnerability
Description	Full description of the vulnerability
Solution	Provided remedies for the vulnerability
Impact	Consequences of the vulnerability on the system
Extra	Extra technical information such as footprint and scanner output
PCI	PCI audit scan result

3.2 Programming Language and Libraries Used

Since performance is not the main focus in this paper, Python scripting language, or Python 3 specifically, is used primarily in building the framework. There are two versions of Python available – Python 2 and Python 3. Python 2 is the legacy version and will reach the end-of-life support in year 2020, while Python 3 is under active development and will be "the future of the language" [16].

There are a few advantages of choosing Python as the development language, the first one being the portability it offers. Since Python is a scripting language, it can be run on any machine that has a Python interpreter installed. This allows the Python code to be executed independently on different platforms without the need of modification. Besides, Python comes with a large number of standard libraries which can be imported as needed instead of requiring all functionalities directly. When performance is a concern, modules written in compiled language such as C can also be used to circumvent the complications of interpreted language. In addition to portability and extensibility, Python is one of the more popular language, it is favored for its decent development speed and easily understood high-level language without compromising any efficacies. Due to its popularity, there are a generous amount of third party modules being actively developed and are available to be used.

In this project, several third-party libraries are used in order to utilize the wide availability of the libraries the community has offered, and also to ease the development of the project without the need to re-implement the functionalities. Some notable libraries are listed as below.

- lxml – lxml is a popular library which is used to parse and process vulnerability reports in XML format. Few other libraries which offer the same functionalities such as cElementTree and the built-in ElementTree are taken into consideration as well. However, after running several tests using parsing techniques posted by IBM [14],

both cElementTree and ElementTree are not chosen due to their inferior performance as compared to lxml's.

- pandas – Built on top of NumPy library, pandas offers fast and flexible data structure which is designed to work with many kinds of data. It is used to access and store the extracted data easily, as well as a built-in function to enable the exporting of the stored data into CSV format.
- Python-docx – The library allows a DOCX format report to be built and generated easily. It is primarily used by template in the final stage of the program.

3.3 Data Processing and Consolidation

Unless otherwise specified, the execution of the program will filter out redundant information generated by some of the vulnerability scanners, namely Nessus and OpenVas. The excessive information mostly originated from the same vulnerability due to different ports such as the port 80 and port 443 which are reserved for HTTP and HTTPS respectively. The process is done by querying duplicate vulnerability name entries stored in the data frame. If attributes of the duplicate pair have unalike content, both contents will be obtained and appended to one another in a newly created record.

Next, consolidation of vulnerabilities from different scanners will be executed. Since scanners of multiple vendors are involved, even though when a system has only one weakness, several results addressing the same vulnerability might be obtained. Moreover, they would have a dissimilar name from one another even though their contexts are similar. The method used for clustering is through linking references associated to the vulnerability. CVE reference is used in this program as it is publicly available and amongst the prevailing standard. Besides, it also highlights the specific instance within a product or a system, thus reducing possibility of incorrectly clustering unrelated results. While the consolidation approach relying on Common Weakness Enumeration (CWE) reference might proves useful, the result will not be as precise as compared to CVE. The reason is that CWE reference highlights the class of a vulnerability instead of the specific instance a vulnerability in. While CWE can be used, more efforts/techniques will be required to refine the broad clustering. Thus, CVE method is used as a proof of concept for this paper.

The program will first create two temporary dictionaries. One contains all unique CVEs key with their associated vulnerabilities of same host, and the other containing the contrariwise. Next, each unindexed key in CVE dictionary is then obtained and processed. Any unprocessed vulnerabilities associated to the CVE are stored and their unindexed CVEs obtained for further operation. The recursion is stopped when all CVEs keys are indexed. All vulnerabilities which are returned from the corresponded recursion will be assigned the same cluster. A sample process is illustrated in Fig. 1.

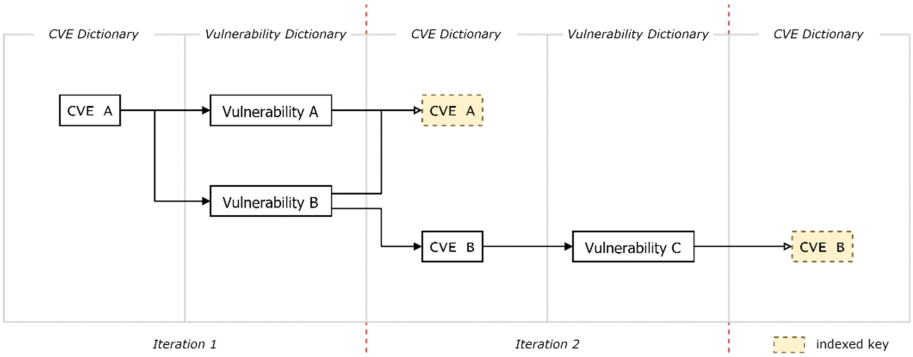


Fig. 1. An example showing the clustering of three similar vulnerabilities.

4 Implementation Result

A simple command-line interface is created in the implementation phase. Input files and reporting template are the mandatory parameters required by the program. A sample demonstration of the command-line usage is shown in Fig. 2 below.

```
>>> exec('python vulnMin.py -e example\ -t SortbyScanner -o example.docx -m')
>> Template SortbyScanner is selected
>> Output file will be saved as example.docx
>> example_01.xml is assigned to.. Nexpose plugin (default)
       > Extracting..                                                    SUCCEED
>> example_02.nessus is assigned to.. Nessus plugin (default)
       > Extracting..                                                    SUCCEED
>> example_03.xml is assigned to.. OpenVAS plugin (default)
       > Extracting..                                                    SUCCEED
>> example_04.xml is assigned to.. Nexpose plugin (default)
       > Extracting..                                                    SUCCEED
>> Consolidating data..                                                 SUCCEED
>> Generating report..                                                  SUCCEED
       > Report example.docx is created
>>>
```

Fig. 2. Demonstration on command-line execution of the program

A database snapshot of the consolidated vulnerabilities is shown in Fig. 3. The generated result is obtained by consolidating reports of three vulnerability scanners which are deployed against an Apache server host.

Scanner	IP Address	Operating System	Host Name	name	port	ref	refURL	severity	cvss(b/v)
Nessus	192.168.31.	Microsoft Windows	LAST-PC	PHP 7.0.x < 7.0.9 Multiple Vulnerabilities (httpoxy)	8080/tcp 443/tcp	BID:91821 BID:92051	http://php.net /ChangeLog-	4	10.0 CVSS2#AV:N/AC:
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-5385	80/tcp 443/tcp	BID:91821 CERT:79789		5	5.1 AV:N/AC:H/Au:N
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6289	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-	https://bugs.ph	7	6.8 AV:N/AC:M/Au:
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6290	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-		8	7.5 AV:N/AC:L/Au:N/
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6291	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-	https://bugs.ph	8	7.5 AV:N/AC:L/Au:N/
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6292	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-		4	4.3 AV:N/AC:M/Au:
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6294	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-	https://bugs.ph	8	7.5 AV:N/AC:L/Au:N/
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6295	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-	https://bugs.ph	8	7.5 AV:N/AC:L/Au:N/
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6296	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-		8	7.5 AV:N/AC:L/Au:N/
Nexpose	192.168.31.	Windows FreeBSD 6.2-	Last-Pc	PHP Vulnerability: CVE-2016-6297	80/tcp 443/tcp	APPLE:APPLE- SA-2016-09-		7	6.8 AV:N/AC:M/Au:
OpenVAS	192.168.31.	cpe:/o:microsoft:wi ndows		PHP Multiple Vulnerabilities-05 July16 (Windows)	80/tcp 443/tcp	CVE:CVE- 2016-6288	http://php.net /ChangeLog-	7.5	7.5 AV:N/AC:L/Au:N/
OpenVAS	192.168.31.	cpe:/o:microsoft:wi ndows		PHP Man-in-the-Middle Attack Vulnerability July16 (	80/tcp 443/tcp	CVE:CVE- 2016-5385	http://www.kb .cert.org/vuls/i	5.1	5.1 AV:N/AC:H/Au:N
Nessus	192.168.31.	Microsoft Windows	LAST-PC	SSL/TLS Diffie-Hellman Modulus <= 1024 Bits (Logjar	443/tcp	BID:74733 CPE:cpe:/a:o	http://weakdh.c	1	2.6 CVSS2#AV:N/AC:

Fig. 3. Snapshot of CSV data storing results of parsed vulnerabilities reports.

While in Fig. 4, an example on one of the clustered vulnerabilities using CVE reference is shown.

Cluster #4	**Vulnerability**	**Nessus**	**Nexpose**	**OpenVAS**
	HTTP TRACE / TRACK Methods Allowed	✓		
	HTTP TRACE Method Enabled		✓	
	http TRACE XSS attack			✓
Reference Type	**Reference ID**	**Nessus**	**Nexpose**	**OpenVAS**
CVE	CVE-2003-1567	✓		✓
CVE	CVE-2004-2320	✓		✓
CVE	CVE-2004-2763	✓	✓	
CVE	CVE-2005-3398	✓	✓	
CVE	CVE-2006-4683	✓	✓	
CVE	CVE-2007-3008	✓	✓	
CVE	CVE-2008-7253	✓	✓	
CVE	CVE-2009-2823	✓	✓	
CVE	CVE-2010-0386	✓	✓	
CERT-Bund	CB-K14/0981			✓
CWE	16	✓		
OVAL	OVAL1445		✓	
DISA_SEVERITY	Category II		✓	
OSVDB	877	✓		
DISA_VMSKEY	V0011706		✓	
BID	37995	✓		
APPLE	APPLE-SA-2009-11-09-1		✓	
XF	14959		✓	
IAVM	2005-T-0043		✓	
CERT	867593	✓	✓	
DFN-CERT	DFN-CERT-2014-1018			✓

Fig. 4. Example of clustered vulnerabilities and their associated references.

5 Conclusion

Research on various topics has been conducted as to discern the components and methods needed to carry out the project. The overall design and structure are laid out using Python programming language to aid in the merging of vulnerabilities on differing scanners. As of writing, plugins for 3 popular network vulnerability scanners are

developed and they are Nessus, Nexpose and OpenVAS. Future work on consolidating vulnerabilities which lack specific reference is to be explored. A further study is also required to integrate the support of scanners targeted at different applications, such as in the case of web application vulnerability scanner.

Acknowledgement. This research work was supported by a Fundamental Research Grant Schemes (FRGS) under the Ministry of Education and Multimedia University, Malaysia (Project ID: MMUE/160029).

References

1. Scarfone, K., Orebaugh, A.: Technical Guide to Information Security Testing and Assessment Recommendations of the National Institute of Standards and Technology. NIST Special Publication 800, pp. 1–80 (2008)
2. Guirguis, R.: Network- and Host-Based Vulnerability Assessments: An Introduction to a Cost Effective and Easy to Use Strategy (2003)
3. Nilsson, J.: Vulnerability scanners (2006)
4. An Overview Of Vulnerability Scanners. The Government of the Hong Kong Special Administrative Region (2008)
5. Risk Based Security: CVE/NVD: The High Price of "Free" (2015)
6. The MITRE Corporation: Common vulnerabilities and exposures. https://cve.mitre.org/
7. National Institute of Standards and Technology: National vulnerability database. https://nvd.nist.gov/
8. Kim, G., Oh, J., Seo, D., Kim, J.: The design of vulnerability management system. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **13**, 19 (2013)
9. Czagan, D.: Common vulnerability scoring system. IEEE Secur. Priv. **2015**, 1–2 (2013)
10. OSVDB Shut Down Permanently (2016). http://www.securityweek.com/osvdb-shut-down-permanently
11. Gold, J.: Open-source vulnerabilities database shuts down (2016). http://www.networkworld.com/article/3053613/open-source-tools/open-source-vulnerabilities-database-shuts-down.html
12. Security Roots: Dradis community edition. https://dradisframework.com/ce/
13. Gremwell: MagicTree. http://www.gremwell.com/what_is_magictree
14. Daly, L.: High performance XML parsing in Python with lxml, pp. 1–10 (2010)
15. Nessus v2 File Format. https://static.tenable.com/documentation/nessus_v2_file_format.pdf
16. Python Software Foundation: Python. https://www.python.org/

A Performance Comparison of Feature Selection Methods for Sentiment Classification

Lai Po Hung(✉), Rayner Alfred, and Mohd Hanafi Ahmad Hijazi

Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
janelai627@gmail.com, {ralfred,hanafi}@ums.edu.my

Abstract. Document sentiment analysis is the task of determining whether a document has a positive, negative or neutral sentiment. It is made up of subtasks including feature extraction, feature selection and sentiment classification. Feature selection is the task of selecting relevant features that can aid the classifier to produce better results. This paper focuses on comparing the classification performances based on several feature selection methods used to select relevant features and also to minimize the document-term matrix representation of the documents. The purpose of applying feature selection besides selecting relevant features is also to reduce the number of features to preserve the efficiency of the whole system. In this work, the experiment setup is designed in order to investigate the effectiveness of several selected feature selection methods in improving the sentiment analysis results. Based on the findings from the experiment, although common feature selection methods such as Document Frequency (DF), Information Gain (IG) and Chi-Squared Statistics (CSS) are found to be able to produce high sentiment analysis accuracy, the Categorical Probability Proportional Difference (CPPD) method is found to be more effective as it produces higher performance accuracy in classifying the documents based on the sentiments. Although, the Categorical Proportional Difference (CPD) method produces acceptable classification results, it is weak in reducing the number of features. In short, the CPPD method enables the sentiment analysis task to be conducted with higher accuracy rate couples with high feature reduction rate too.

Keywords: Sentiment analysis · Feature selection · Feature extraction
Information Gain · Chi-Squared Statistics · Categorical Proportional Difference

1 Introduction

Public opinion contributes a significant impact on the operations of business and product perception. Sentiment analysis is the task of classifying a document into positive, negative or neutral sentiment groups. This task is important in opinion mining. Opinions are useful resources as highlighted above, and the internet has enabled an enormous amount of opinion to be circulated. Harvesting and summarizing large amount of data is better done with technology than manual labor. Sentiment analysis plays an important role in which it analyses the opinion so that a summary of general opinion polarity can be produced. Sentiment analysis is normally done by applying a rule based system or a

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 21–30, 2018.
https://doi.org/10.1007/978-981-10-8276-4_3

machine learning based system. Machine learning based systems are more popular because they are easier to apply and more versatile. A machine learning algorithm is trained to recognize the underlying patterns in documents in order to classify documents. There are three essential steps involved in machine learning based sentiment analysis and they are feature extraction, feature selection and the machine learning classifier [1]. Features represent a single unit used in the analysis process when classifying documents into the corresponding polarities. Huge amount of features will burden the overall system with heavy processing load. Irrelevant features will cause over fitting or under fitting models of classifiers to be trained and produced. It is ideal for the system if the feature set is considerably small but informative and accurate. Feature extraction and selection is responsible for producing the desired feature set. Feature extraction involves extracting features from the text that can be recognized and analyzed by the classifier. These features include N-Grams, Part of Speech, Named-Entity Recognition, SentiWordNet [1, 2, 25].

Feature selection performs the tasks of filtering the extracted features for useful features and removing irrelevant features. Feature selection also helps reduce the feature set size to preserve the efficiency of the system. Feature Selection brings two benefits to the system. Firstly, using smaller feature set allows the system to perform efficiently as it takes less time to finish its run but maintains its accuracy. Complex classifying algorithms that involve many processes with large amounts of features, will require feature selection to ensure that the algorithm does not become impractical due to the large amount of processing load and time required [3]. Secondly, irrelevant features can negatively impact the performance of the system as they affect the quality of model produced during the modelling process [4]. This paper focuses on comparing the classification performances based on several feature selection methods used to select relevant features and also to minimize the document-term matrix representation of the documents. These feature selection methods include some of the common methods used such as Document Frequency (DF), Information Gain (IG) and Chi-Squared Statistics (CSS), and also less common methods such as Total Term-Frequency Inverse Document Frequency (TF-IDF), Categorical Proportional Difference (CPD), Probability Proportional Difference (PPD) and Categorical Probability Proportional Difference (CPPD). There are minimal to none performance comparison performed on the less common methods as they are relatively new, and TF-IDF is normally used for feature weighting, however we chose to test the total TF-IDF as a feature selection method instead.

This paper contains several parts. Section 2 discusses previous work in feature selection for sentiment analysis. Section 3 explains the experimental setup and Sect. 4 provides the results obtained and discusses the findings based on the experimental results. Section 5 concludes this paper by discussing future works to be explored which can bring improvement the process of sentiment analysis.

2 Related Works

There are several feature selection methods that can be used to perform the feature scoring [5]. These feature selection methods are described next.

Document Frequency (DF) is a simple method that is scalable to the size of the dataset. This method selects based on the counts of documents in a training set where the feature term occurs. It removes feature terms with document frequencies which are under or over given thresholds. The assumption applied here considers feature terms that are too rare and too common occurring in training documents to be less helpful in improving accuracy of classification in sentiment analysis [6].

Information Gain (IG) is a method that values the usefulness of a term for classification based on the information it can provide to help the classifier discern between classes. It measures the amount of information a term carries by comparing the entropy of the term absence and term presence in a document [5]. The formula is shown below,

$$IG(f) = [-\sum_{i=1}^{|c|} P(c_i) \log P(c_i)] - [-P(f) \sum_{i=1}^{|c|} P(c_i|f) \log P(c_i|f)] - [-P(f') \sum_{i=1}^{|c|} P(c_i|f') \log P(c_i|f')] \quad (1)$$

where $P(c_i)$ is the probability of the class occurring, $P(f)$ is the probability of the feature term presence and $P(f')$ is the probability of the feature term being absent [7].

Mutual Information (MI) is differs from IG as it does not consider the absence of term [8]. MI measures the dependencies between two variables. When the MI score is higher, it means that the variables are more dependent on each other as shown below, where $P(f,c)$ is the probability of the feature term and class occurring together, $P(f)$ is the probability of feature term occurrence, $P(c)$ is the probability of class occurrence.

$$MI(f, c) = \log \frac{P(f, c)}{P(f) \times P(c)} \quad (2)$$

The drawback of MI is that rare terms with equal conditional probabilities will have higher scores than those of common terms. Considering the drawback, MI scores of features that have great difference in frequencies cannot be compared [7].

Chi-Squared Statistics (CSS) can measure the association strength between a feature term and a class by calculating how dependent the term and class is on each other. If a term and class is dependent on each other, then the term is useful in identifying the class it belongs to. The formula and explanation [1] is as below,

$$\chi^2(f, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3)$$

where A is the frequency that f and c occurs together, B is the frequency that f occurs without c, C is the frequency that c occurs without f, D is the frequency where neither c nor f occurs and lastly N is the total number of documents.

Simeon and Hilderman **proposed the Categorical Proportional Differences (CPD)** method to measure the usefulness of a term to differentiate between different categories [9]. The measurement is based on the ratio of the frequencies of a word occurring across different categories of documents. The calculation is as shown,

$$CPD(f, c) = \frac{A - B}{A + B} \tag{4}$$

where A is the frequency of the word and category appearing together and B is the frequency of the word occuring without the category. When CPD score approaches maximum score of 1, this shows that the feature occurs more frequently in documents of a particular category only and is helpful to distinguish between categories. CPD is quite a recent method and has been used in a few experiments only [9, 10]. The benefit of CPD is that it can eliminate common terms with high document frequency but are not important such as stop-words, based on their equal occurrence in all classes of documents [11]. On the other hand, the weakness of CPD is that it might retain irrelevant rare terms due to their low document frequencies.

A variation of the CPD method that combines CPD and another method called the Probability Proportional Difference (PPD) method has also been proposed, in which the authors proposed as a measure of the feature terms' relation with a particular category [11]. A high PPD value shows that the term belongs to a certain class, whereas, a low PPD value shows that it appears in multiple classes and is not very helpful for classification. The PPD calculation for sentiment analysis is computed as follows,

$$PPD(f) = \frac{nPos(f)}{sPos + S} - \frac{nNeg(f)}{sNeg + S} \tag{5}$$

where *nPos(f)* represents the number of positive documents where the feature is present, *nNeg(f)* represents the number of negative documents where the feature is present, *sPos* represents the number of features in the positive class, *sNeg* represents the number of features in the negative class and *S* represents the total number of features. The combined method proposed is the Categorical Probability Proportional Difference (CPPD) method which selects feature terms that achieves both thresholds set for CPD and PPD. The advantage of this method is that CPD and PPD compensate for the weaknesses of each other. When CPD selects irrelevant terms with low frequency, PPD will eliminate the term, whereas when PPD selects irrelevant terms with high frequency, CPD will eliminate the term.

Table 1 summarizes several works performed which focus on performing a comparison of the performance of various feature selection methods. The popular methods that have always been used are DF, IG and CSS, these methods are methods that are matured and frequently used by researchers for text classification related works.

DF method is the simplest method of selection but it is weaker in terms of performance. IG and CSS are more commonly used among classification tasks because they have better performances. Three of the mentioned methods measure the usefulness of a feature through consideration of the presence and absence of the feature among documents in a feature set. The IG and CSS methods are also aggressive selectors adding to their favor, the two of them are commonly applied in works of text classification [12–16].

Table 1. Previous works that performed comparison on feature selection methods

Authors	Feature selection methods	Classifiers	Dataset	Best performing method
Yang and Pedersen [7]	DF, IG, MI, CSS, Term Strength	*k*-Nearest Neighbor(*k*-NN), Linear Least Squares Fit	Reuters-22173 collection, OHSUMED collection of medical journals	IG, CSS
Tan and Zhang [17]	DF, MI, CSS, IG	Centroid classifier, *k*-NN, Naïve-Bayes, Winnow classifier, Support-Vector-Machine(SVM)	ChnSentiCorp-Chinese sentiment documents	IG
Sharma and Dey [6]	DF, IG, Gain Ratio, CSS, Relief-F algorithm	Naïve Bayes, Max Entropy, SVM, Decision Tree, *k*-NN, Winnow classifier, AdaBoost classifier	Cornell Movie Review Dataset	Gain Ratio
O'Keefe and Koprinska 10]	CPD, SentiWordNet Subjectivity Scores, SentiWordNet Proportional Difference	Naïve Bayes, SVM	Cornell Movie Review Dataset	CPD
Sarkar and Goswami [8]	IG, MI, CSS, Symmetrical Uncertainty	Naïve Bayes, SVM,Decision Tree, *k*-NN	CNAE-9, SMS Spam Collection, Preview of hotel data	IG
Zhao et al. [18]	DF,IG, CSS, MI	Naïve Bayes, Max Entropy, SVM, Winnow classifier, C4.5 Decision Tree Classifier	Chinese news comments	CSS

3 Experimental Setup

In this paper, the experimental setup is designed to investigate the effectiveness of feature selection methods in influencing the classification accuracy of the sentiment analysis and also the ability to reduce the feature count. Based on the best performing methods listed in Table 1, seven feature selection methods are implemented and evaluated in this work, and these methods include the DF, IG, CSS, Total TF-IDF, CPD and its variation, PPD and CPPD.

The dataset used here is obtained from the Multi-Domain Sentiment Dataset [19]. The dataset consist of English language product reviews from different domains. The reviews chosen in this work are obtained from the domains of Books, DVD and Electronics. For each domain, 1000 reviews are obtained and the reviews are preprocessed and represented as features consisting of a combination of unigrams and bigrams.

In this work, the TF-IDF method is used as a feature weight in the term document matrix for classification and the Total TF-IDF value for a feature in all reviews is used as a feature selection metric. TF-IDF value rare terms higher based on the observation that rare terms can be more selective. The TF-IDF is defined as follows,

$$TFIDF = TF \times IDF = (1 + \log_2 tf) \times (\log_2 \frac{N}{n_i}) \tag{6}$$

where, *tf* is term frequency in a document, *N* is the total number of documents in the dataset and n_i is the number of documents where the term occurs.

Naïve Bayes classifier is used in this experiment because its application is simple and fast and its performance in text classification is quite highly regarded by many researchers [6, 20–23]. Naïve Bayes classifiers assume that variables have conditional independence and calculate the probability of a document or feature being assigned to a particular class (positive or negative). This classifier is built based on the Bayes theorem [1]. In this experiment, all classification tasks will be conducted by using the classifiers obtained from the WEKA (Waikato Environment for Knowledge Analysis) system developed by the University of Waikato. Weka is a data mining tool that contains a group of machine learners and other processing applications. The features can be run in WEKA after converting the file format to ARFF format. The training and testing of classifiers based on cross fold validation is carried out within WEKA [24] (Fig . 1).

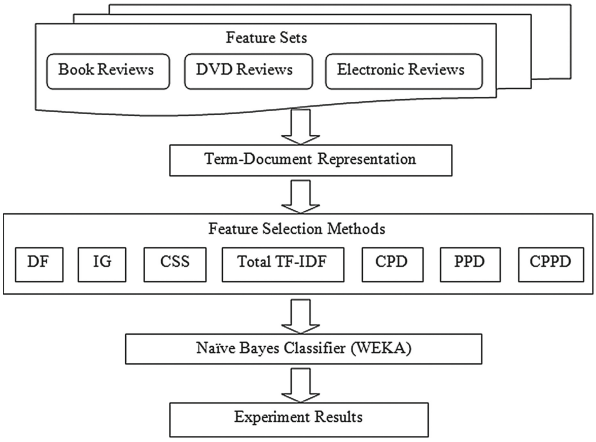


Fig. 1. The experimental setup designed to investigate the effectiveness of feature selection methods on the classification accuracy of sentiment analysis

The feature selection methods outlined previously assigns each feature a value that reflects their usefulness for classification. Prior to classification, only features with values that exceed certain thresholds are chosen and used. In the experiments conducted, 10 threshold values are generated for each feature selection method in each dataset experiment where the values are in increasing order. The 10 thresholds are generated by arranging all unique values calculated by a feature selection method in increasing order, and selecting the threshold value found within 10 equal intervals in position. Results for the threshold that has the best sentiment analysis accuracy of each method will be selected and shown as the results in Sect. 4 in order to compare the performances of the feature selection methods.

4 Results and Discussion

Table 2 outlines the results obtained from running sentiment analysis on three different dataset domains with different feature selection methods. The best threshold is obtained by taking the threshold scores of the feature selection methods that produce the best classification accuracy. The feature count represents the number of features selected after applying the threshold and the reduction rate provides the percentage of features reduced by the threshold and finally the accuracy rate presents the classification accuracy of sentiment analysis performed with the features selected.

Table 2. A performance comparison of feature selection methods results on document sentiment analysis for three topics: Book, DVD and Electronic reviews datasets

Feature selection method	Best threshold	Feature count	Reduction rate (%)	Accuracy rate (%)
Book				
No feature selection	–	113273	0.0	62.1
Document frequency	35	1308	98.8	63.1
Information gain	0.00331	1116	99.0	76.2
Chi-Square	7.23037	199	99.8	75.7
Total TFIDF	186.12464	570	99.5	63.4
CPD	1.0	98262	13.3	**98.6**
PPD	0.00021	27	**99.9**	70.7
CPPD	0.00002	2002	98.2	97
DVD				
No feature selection	–	110715	0.0	72.7
Document frequency	18	1313	98.8	72.6
Information gain	0.00331	1150	99.0	84.9
Chi-Square	4.96159	620	99.4	84.5
Total TFIDF	167.88627	678	99.4	72.9
CPD	1.0	96042	13.3	**99.1**
PPD	0.00006	477	**99.6**	79.9
CPPD	0.00002	2032	98.2	96.3
Electronics				
No feature selection	–	66524	0.0	74.1
Document frequency	15	1247	98.1	75
Information gain	0.00293	1929	97.1	82.6
Chi-Square	6.49176	300	99.5	82.5
Total TFIDF	119.00907	950	98.6	74.7
CPD	1.0	56816	14.6	**90.8**
PPD	0.00012	285	**99.6**	81.8
CPPD	0.00002	5237	92.1	**90.8**

Based on the comparison results shown in Table 2 above, it can be concluded for this experiment that CPD is the best feature selection method out of the others since it produces the highest accuracy for the sentiment analysis classification results. However CPD's reduction rate is the lowest of all. While others achieve reduction rate of 90% and above, CPD is the only one that achieved reduction rate of less than 20%. The PPD managed to reduce the most number of features from 100% to only less than 1% of the

features compared to the other methods. The weaker methods that produced lower accuracies are DF and Total TF-IDF. Based on the observations of the results, it can be summarized that the pair of methods (DF and Total TF-IDF), (IG and CSS) and (CPD and CPPD) methods produce similar results to each other. This observation means that these pairs of methods have similar behaviors as shown in their formulas.

The DF and total TF-IDF methods rely on the presence of features as the basis to their measurement and their results are lower compared to the others. DF counts how many document does the feature appear in, but it does not discern whether the appearance of the feature is equally distributed among documents of different classes or more skewed towards document of a certain class. As a result, it cannot find features that are discriminative. Since the TF-IDF method weights features more by their value in a document, the poor result obtained suggests that the importance of a feature in a single document does not emulate the importance of the feature in the whole dataset. The two methods can be used as good selectors if the focus of the task is on the reduction of features. This is because they are quite aggressive in feature reduction, in which more than 90% of the features in datasets are reduced and the classification accuracies are acceptable although they are weaker when compared to the other feature selection methods.

The IG and CSS methods also produce similar results and calculation parameters. In order to compute the score of feature by using the IG and CSS methods, the frequency of feature appearance and absence in a certain class need to be taken into consideration. In other words, when measuring the importance of a feature in the dataset, the two methods compare whether the feature is dispersed equally among classes or not. This is probably why they performed better than the two previous methods (e.g., DF and TF-IDF). IG and CSS methods are considered as very good feature selection techniques because they are aggressive in feature reduction and at the same time with the reduced features they can produce acceptable classification accuracy that are quite high. For these reasons they are very popular methods used in classification tasks.

The CPD and CPPD methods yield the best results in the experiments. CPPD is basically a method that is designed based on two previously defined methods, CPD and PPD. Both of these methods compare the presence and absence of particular features. However, the CPD method tends to eliminate common terms, whereas PPD method tends to remove rare terms. Based on the results obtained, it is found that keeping rare terms yield better classification accuracies compared to keeping common terms as the CPD method outperforms the PPD method. The CPD method is able to produce high accuracies because it can select terms that are more exclusive to a certain class only, however the downfall of it is that it is very poor in feature reduction since its reduction rate is less than 20% in all the three datasets. This is probably due to fact that CPD retains the extremely rare features that appear in only one document out of the whole datasets. These features are either extremely informative or useless. The PPD method on the other hand, produces lower accuracies compared to the CPD method but it is very aggressive in feature reduction. The PPD method was created to behave the opposite of CPD as it eliminates those rare features that occur in only one document. Thus, the CPPD method, which is the result of combining both the CPD and PPD methods, enjoys the best of both worlds as it manages to reduce the features by more than 90%, and at the same time

retains quite high performance too. So, when considering the traits of a good feature selector which is increased performance and high feature reduction, the CPPD method is considered to be a better feature selection method for classifying document into classes of sentiment polarities.

5 Conclusion

This paper presents an assessment of several feature selection methods that can be applied to select relevant features for the purpose of classifying documents based on sentiment analysis. An experiment has been conducted to compare seven feature selection methods that include DF, IG, CSS, Total TF-IDF, CPD and its variation, PPD and CPPD. Three datasets have been used to perform the assessment which are the reviews taken from three main topics (e.g., Book, DVD and Electronics). The reviews are represented as combinations of unigram and bigram and a Naïve Bayes classifier is used to perform the sentiment classification. Based on the results obtained, the best sentiment analysis result can be obtained when the CPD method is applied to reduce the number of features presented to the Naïve Bayes classifier in classifying documents based on the sentiment polarities (e.g., positive, negative and neutral polarities). However, the number of features reduced is minimal and the large set of features selected is not efficient to be used in the classification task. On the other hand, the PPD method yields lower classification accuracies with higher percentage of features reduced. The CPDD method is found to be very effective in balancing the classification accuracy and the percentage of features reduced. CPPD combines CPD and PPD elements and this method is able to create a balanced performance between CPD and PPD methods since it can achieve high sentiment classification accuracy and also high percentage of feature reduction rate in achieving that accuracy. Since, feature extraction process may also influence the classification accuracy of classifying document into sentiment polarities, future works may include the task of classifying documents into positive, negative and neutral sentiments based on types of methods used for feature extraction. These feature extraction methods include N-gram, stemmed, Part of Speech, and Named-Entity Recognition based methods.

References

1. Hung, L.P., Alfred, R., Hijazi, M.H.A., Ibrahim, A., Asri, A.: A review on the ensemble framework for sentiment analysis. Adv. Sci. Lett. **21**(10), 2957–2962 (2015). http://dx.doi.org/10.1166/asl.2015.6494
2. Alfred, R., Leow, C.L., Chin, K.O., Anthony, P.: Malay named entity recognition based on rule-based approach. IJMLC **3**(4), 300–306 (2014)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval the Concepts and Technology Behind Search, 2nd edn. Pearson Education Limited, England (2011)
4. Vanaja, S., Ramesh, K.K.: Analysis of feature selection algorithms on classification: a survey. Int. J. Comput. Appl. **96**(17) (2014). 0975–8887
5. Hung, L.P., Alfred, R., Hijazi, M.H.A.: A review on feature selection methods for sentiment analysis. Adv. Sci. Lett. **21**(10), 2952–2956 (2015). http://dx.doi.org/10.1166/asl.2015.6475

6. Sharma, A., Dey, S.: A performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. IJCA Special Issue Adv. Comput. Commun. Technol. HPC Appl. **3**, 15–20 (2012)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997, pp. 412–420 (1997)
8. Sarkar, S.D., Goswami, S.: Empirical study on filter based feature selection methods for text classification. Int. J. Comput. Appl. **81**(6), 38–43 (2013)
9. Simeon, M., Hilderman, R.: Categorical proportional difference: a feature selection method for text categorization. In: Proceedings of the 7th Australasian Data Mining Conference, vol. 87, pp. 201–208 (2008)
10. O'Keefe, T., Koprinska, I.: Feature selection and weighting methods in sentiment analysis. In: Proceedings of the 14th Australasian Document Computing Symposium, Australia (2009)
11. Agarwal, B., Mittal, N.: Categorical probability proportion difference (CPPD): a feature selection method for sentiment classification. In: Proceedings of the 2nd Workshop on Sentiment Analysis where AI Meets Psychology (SAAIP 2012), Mumbai, pp. 17–26 (2012)
12. Azhagusundari, B., Thanamani, A.S.: Feature selection based on information gain. Int. J. Innovative Technol. Explor. Eng. (IJITEE) **2**(2), 18–21 (2013). ISSN 2278–3075
13. Hassan, A., Abbasi, A., Zeng, D.: Twitter sentiment analysis: a bootstrap ensemble framework. In: International Conference of Social Computing, pp. 357–364 (2013)
14. Koncz, P., Paralic,J.: An approach to feature selection for sentiment analysis. In: 15th International Conference on Intelligent Engineering Systems, Poprad, Slovakia (2011)
15. Shang, L.: A feature selection method based on information gain and genetic algorithm. In: International Conference on Computer Science and Electronics Engineering (2012)
16. Uchyigit, G.: Experimental evaluation of feature selection methods for text classification. In: 9th International Conference on Fuzzy Systems and Knowledge Discovery (2012)
17. Tan, S., Zhang, J.: An empirical study of sentiment analysis for Chinese documents. Expert Syst. Appl. **34**, 2622–2629 (2008)
18. Zhao, Y., Dong, S., Lee, L.: Sentiment analysis on news comments based on supervised learning method. Int. J. Multimedia Ubiquit. Eng. **9**(7), 333–346 (2014)
19. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. Association of Computational Linguistics (ACL) (2007)
20. Wang, G., Sun, J., Ma, J., Xu, K., Gu, J.: Sentiment classification: the contribution of ensemble learning. Decis. Support Syst. **57**, 77–93 (2014)
21. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, Association of Computational Linguistics, pp. 79–86 (2002)
22. Su, Y., Zhang, Y., Ji, D., Wang, Y., Wu, H.: Ensemble learning for sentiment classification. In: Ji, D., Xiao, G. (eds.) Proceedings of 13th Chinese Conference on Chinese Lexical Semantics (CLSW 2012), pp. 84–93. Springer, Heidelberg (2012)
23. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. **181**, 1138–1152 (2011)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
25. Lai, P.H., Alfred, R.: A performance comparison of feature extraction methods for sentiment analysis. In: Król, D., Nguyen, N., Shirai, K. (eds.) Advanced Topics in Intelligent Information and Database Systems. ACIIDS 2017. SCI Book Series vol. 710 (2017)

A Real Time Road Marking Detection System on Large Variability Road Images Database

B. S. Khan(✉), M. Hanafi(✉), and S. Mashohor(✉)

Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, Jalan UPM, 43400 Serdang, Selangor, Malaysia
bahadur_00_shah@hotmail.com, {marsyita,syamsiah}@upm.edu.my

Abstract. For no less than two decades, the development of autonomous systems has led to the development of embedded applications permitting to enhance the driving comfort and limit the hazard level of dangerous zones. One of the first embedded system is a lane detection system, which was implemented using road marking detection algorithms with the aim to produce a system that is able to detect various shapes of road markings on the images that are captured under various imaging conditions. Generally, the road images were captured using a camera, which has been placed inside a vehicle at a fixed position. In this paper, a road markings detection system that tackles the problems of detecting road markings on the images captured under various weather and illumination conditions is proposed. The proposed system consists of inverse perspective transform method, which is used to convert an image into a bird's-eye view image, an image normalization method, namely CLAHE that tackle various illumination conditions and Sobel edge detection method for identifying the road marker. We demonstrate the usefulness of the constructed algorithm by performing experiments on our Large Variability Road Images database (LVRI) that consists of 22,500 road images with the accuracy of 96.53%.

Keywords: Road marks detection · Bird's eye view image · CLAHE · Sobel edge detection · LVRI database

1 Introduction

Road markings are defined as the lane borders markings and painted arrows on the road surface. In the area of autonomous vehicles and driver support technologies, a large number technique has been introduced to detect road markings in order to improve driving safety and reduce car accidents. Till now a large number of road marking detection systems have been developed using various types of sensors, such as radar and infrared sensors, inductive loop, and microwave detectors. However, the main issues with these sensors are high installation and maintenance expense. The problems are tackled by introducing, video sensors which are inexpensive and as well as slight traffic disruption. Hence, due to the advantages, the proposed algorithm is developed via video camera sensors as in [1–3]. In recent years, the study as detecting road marking has received great attention from the scholars. In [4–8] scholars had applied threshold

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 31–41, 2018.
https://doi.org/10.1007/978-981-10-8276-4_4

segmentation method for road marking detection. Usually a successful road marking segmentation is highly dependent on the choice of thresholds. It can be expected to be successful in high contrast environments. Its advantages are simple and the applicability is strong. However, under uneven illumination conditions, threshold selection would be a major challenge. Its disadvantages are low capacity of resistance, noise and less sensitive to the intensity variation of the gray image. In [9], the authors developed ROI (region of interest) based lane detection method where, the lanes were detected using the Canny edge detector. However, the Canny edge detector raises the noise on the road screen when shadow is present. Furthermore, a robust noise reduction algorithm is required to deal with those problems or farther detection algorithm required. Lane-marking detection based on HSI color is presented in [10]. However, the HSI model works well in simple free roads with clear images, but under the complex road image with heavy traffic scenario conditions such kind of systems might get an inconsistent detection result. In [11], a real-time road marker detection by utilizing improved Hough transform and RANSAC line fitting is developed. However, under complex conditions, such as heavy traffic scenario or poor weather condition produced inconsistent detection result and the method only able to detect lanes and not arrow. In [12–15], the authors detected straight lanes using Hough transform and admitted had difficulty to detect continuous (non-straight) curves and claimed that Hough space alone is not efficient enough to detect arrow. In [16–18] the authors developed a robust algorithm to work with extreme shadow conditions based on an illumination invariant combination with a region growing algorithm for road segmentation. For road markings detection purpose, the method is not explicit, it takes a devious way to acquire road markings by seeking road region first and extra computation time would be required.

2 Methodology

The system consists of several stages, as shown in Fig. 1.

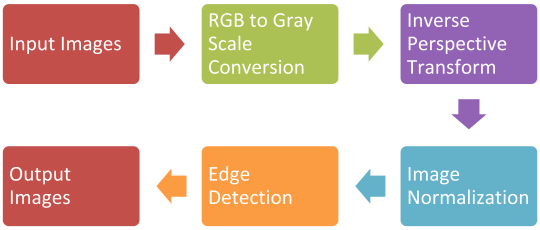


Fig. 1. The flowchart of the proposed method

The first step in our approach is to convert RGB images into grayscale images in order to reduce processing time. Then, the grayscale images are transformed into the bird's-eye view images using an Inverse Perspective Transform method in order to reduce distortion caused by Perspective Transform while capturing images and then

normalize the images to tackle various illumination. The last stage is to detect the road marking utilizing Edge detection method which is robust in occlusions.

2.1 Inverse Perspective Transform

Inverse perspective transformation (IPT) is a geometrical transformation technique that projects a 3-dimensional object from a 2-dimensional perspective view and constructs a new image on an inverse 2-dimensional planar by re-maps each pixel to a new position. This will result in the bird's eye view of the image and thus, removes the perspective effect [19–21]. To proceed with inverse perspective transformation, RGB images converted to grayscale images. Basically, the method is divided into three stages [22]. The first stage is to represent the image in a shifted coordinate system in 2-dimensional plane where the x and y coordinates of an image are shifted using Eqs. (1) and (2).

$$X = x - (image\ size/2) \tag{1}$$

$$Y = y - (image\ size/2) \tag{2}$$

where (x, y) represent the coordinates of the segmented image. Then, the image is rotated using a simple matrix multiplication operation, as shown in Eq. (3). The location of the rotated image is represented as (p, q, r), which is the positional coordinates in a three-dimensional plane, and they are calculated from X, Y and Z values which are the positional coordinates of two-dimensional planes. However, the initial value for Z is declared as zero due to conversion from two-dimensional plane to three-dimensional plane [22, 23].

$$\begin{bmatrix} p \\ q \\ r \\ 1 \end{bmatrix} = R \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3}$$

where R is the rotation matrix in a three-dimensional plane, as given by Eq. (4). The rotation can be performed around X, Y and Z axis by using the different rotation matrix as given by (4)–(6). ω, φ, and κ represent the rotation angle in degree of X-axis, Y-axis and Z-axis, respectively. The rotation matrices, for X-axis, Y-axis and Z-axis can be represented as R_x, R_y, and R_z, respectively. The rotated angle can be found by knowing the angle at which camera is mounted. In our work, the top view images were produced by mounting the camera at the angle of 35° from the road and the images can be rotated only in Z-axis, at $\varphi = 35°$.

$$R_x = \begin{bmatrix} \cos(\kappa) & \sin(\kappa) & 0 & 0 \\ -\sin(\kappa) & \cos(\kappa) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

$$R_y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\omega) & \sin(\omega) & 0 \\ 0 & -\sin(\omega) & \cos(\omega) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

$$R_z = \begin{bmatrix} \cos(\varphi) & 0 & -\sin(\varphi) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\varphi) & 0 & \cos(\varphi) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

Equations (7) and (8) were used to project the image on a two-dimensional plane.

$$u = \frac{f * p}{f - r} + (image\ size/2) \tag{7}$$

$$v = \frac{f * q}{f - r} + (image\ size/2) \tag{8}$$

where (u, v) is the projected coordinate and f represents the focal length [22]. The example of transformed image produced by the inverse perspective transform method is shown in Figs. 3(b), 4(b), and 5(b).

2.2 Image Normalization

Under different illumination conditions, the problem of detecting the road images is tackled by normalizing the road using Contrast Limited Adaptive Histogram Equalization (CLAHE). First, the method computes several histograms, each relating to a particular section of the image and uses them to redistribute the lightness values of the image. It is subsequently reasonable for enhancing the local contrast and improving the definitions of edges in every area of an image. An improved algorithm of Adaptive Histogram Equalization (AHE) is known as CLAHE [24]. CLAHE is developed to prevent over amplification of noise raised by AHE. In this method, an image is divided into several small regions called tiles. The division of the tiles is based on the user-defined tile size. In our work, the number of tiles are 8 rows × 8 columns, total is equal to 64. In each contextual, the numbers of pixels are equally divided at each gray level and contrast limiting is applied to the histogram. The histogram bin is clipped at the predefined value known as clip limit and in this work (Clip Limit) value of 0.01 is used. In the event that any histogram bin is above the clip limit, those pixels are clipped and distributed equally among other histogram bins and number of histogram bin value is 256 used in this work. Figures 3(c), 4(c) and 5(c) shows examples of a road image that has been normalized using CLAHE.

2.3 Edge Detection

An image gradient is referred as a directional change in the color or intensity in an image [25]. An edge in an image appear when the gradient is most prominent and the Sobel operator makes utilization of this fact to discover the edges in an image. Sobel operator discover the approximate image gradient of each pixel by convolving the image with a pair of 3×3 kernels [25]. The kernels as shown in the Eq. (9) estimate the gradients in the horizontal (G_x) and (10) vertical (G_y) directions and the magnitude of the gradient is simply the sum of the two gradients, as shown in Eqs. (11) and (12).

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{9}$$

$$G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{10}$$

$$\text{Magnitude} \quad |G| = \sqrt{G_x^2 + G_y^2} \tag{11}$$

$$\text{Direction} = atan(y/x) \tag{12}$$

The Sobel operator has a bigger convolution mask that smoothers the input image to a more prominent degree and makes the operator less sensitive to noise. The operator predominantly delivers extensively higher output values for comparative edges with the Roberts Cross [26]. Figures 3(d), 4(d) and 5(d) shows examples of a road image edge detection using Sobel edge detection.

2.4 Data Acquisition Method

The road images were captured using a single camera sensor as in [2, 3] and objective is to detect road and road marking that are situated in front side of the vehicle. Therefore, the installation of a smartphone camera was on the front windshield of the experimental vehicle as shown Fig. 2.

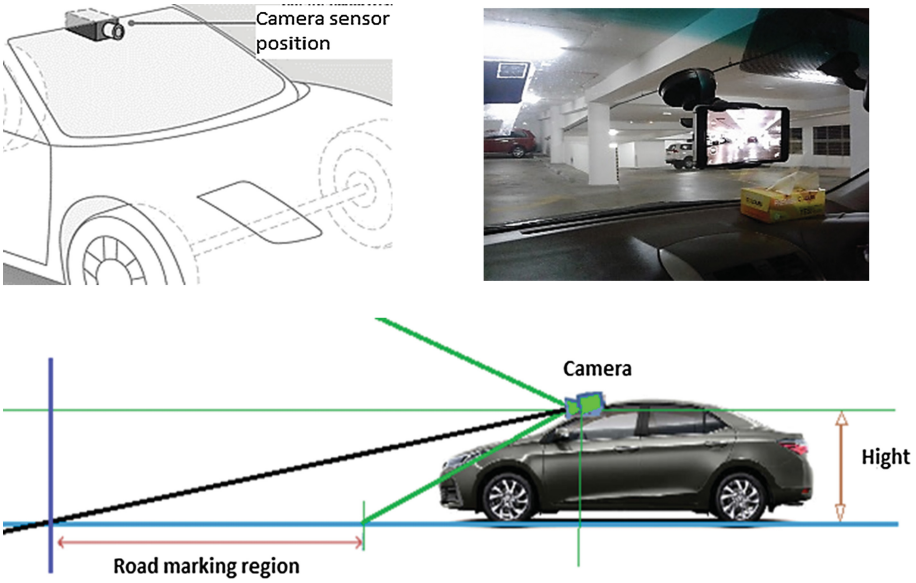


Fig. 2. Camera setup position

The images used in the experiment were extracted from videos, which's were recorded using Samsung Galaxy Alpha (SM-G850F). The recorded video is a full HD Video with a resolution of 1920 × 1080 (.mp4) and was captured at 30 fps in auto video capture mode without video stabilization. The total number of videos used in the experiment is 15 and the total number of images extracted from the videos is 22,500. The images are under various imaging conditions such as variation in illuminations, traffic and weather conditions. During the data collections, the host car was driven by following the two second safety rule. Safe following distance is very important while driving a vehicle and safe following distance also required to be established for autonomous driving. The safe and correct following distances are hard to establish for car [27]. Consequently, the two second safety standard is used to gain a safe following distance at any speed [28, 29]. The standard is that a driver should remain no less than two seconds behind any vehicle that is straightforwardly in front of the driver's vehicle.

2.5 Results and Discussions

The algorithm was tested on our LVRI datasets that consists of 22500 road images that consist of captured images in day and night with different illumination and occlusions such as shadow, complex background, traffic, slight rain, rains and snow. Images with the after-rain effect are also collected. The total number of images for each situation is shown in Table 3. The images that were captured at day and night by a camera, which was mounted on the dashboard, the location for the data collection is around Selangor and Kuala Lumpur. The rest of the images (Slight rain, Rain, After rain, Snow) in the dataset were downloaded from the internet and the images consist of reflection effect,

complex background and were captured during the day and night under various illuminations condition and occlusions. In order to detect road marking, input RGB images converted to grayscale images which were transform in to bird's eye view using inverse perspective transform and normalized using CLAHE and then road markings detected using Sobel edge detection. The results of detected road markings in percentages are shown in Table 1, where road marking detection result of daylight with various illumination condition contains shadow and complex background is 97% and for night images is 94%. The images during the day and night with occlusions such as slight rain, rain, after rain and snow, detection result is respectively 99%, 91.7%, 99% and 98.5%. In the datasets, comparatively higher rate of shadow presence in 'Day' images rather than 'Slight rain' and 'After rain' images, that is why 'Slight rain' and 'After rain' have higher detection than 'Day'. Overall road marking detection result is 96.53%. However, the proposed algorithm has difficulty in detecting lane markings painted in soft yellow color, hence road markings detection rate reduced to 96.53%. Figures 3, 4 and 5 shows some of the examples of the detected road markings during the day and night with different illumination and occlusions. It is shown that for the images in the LVRI datasets, the results are promising. Computation time is also calculated for read input images from datasets, RGB to gray scale conversion, applying inverse perspective transform (IPT), normalization using CLAHE, Sobel edge detection and then detected result for single image in milliseconds (ms), which is generated by MATLAB code profiler. Self-time is the total time spent in a function, not including any spent time in any child functions called and total time is the total time spent on a function includes the timing in all of the child functions called are shown in Table 2. Intel core i5 4th (2013) generation processor was used to compute the result. The proposed algorithm also had applied in several video files in with low 640 × 480 resolution and its work fine in real-time simulation.

Table 1. The detection of road marking in the LVRI datasets

Set	No of image	Detection	Percentage
Day	1500	1467	97%
Night	12000	11803	94%
Slight rain	1500	1495	99%
Rain	4500	4127	91.7%
After rain	1500	1490	99%
Snow	1500	1477	98.5%
Total	22500	21859	96.53%

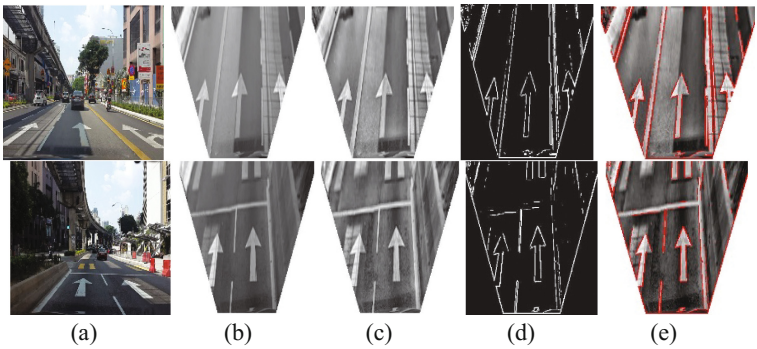


Fig. 3. Road images in daylight with different illumination contain shadow, where (a) Input images, (b) Inverse prospective transform images in gray scale, (c) CLAHE normalization, (d) Sobel edge detection and (e) Output images.

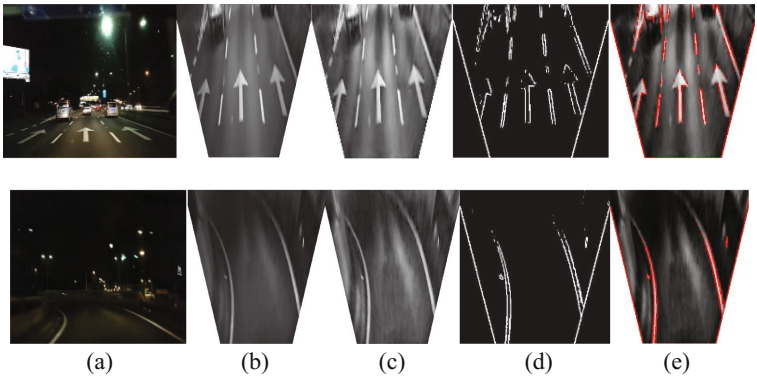


Fig. 4. Road images during night with different illumination contain very low light images, where (a) Input images, (b) Inverse prospective transform images in gray scale, (c) CLAHE normalization, (d) Sobel edge detection and (e) Output images

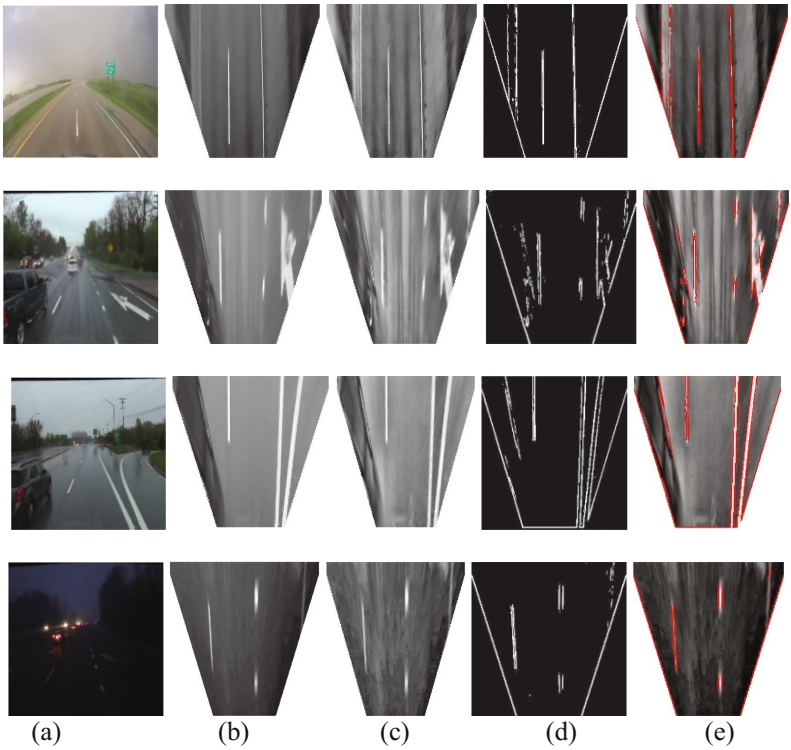


Fig. 5. Road images in day and night with occultations contain slight rain, rainy, after rain, and snow falling images, where (a) Input images, (b) Inverse prospective transform images in gray scale, (c) CLAHE normalization, (d) Sobel edge detection and (e) Output images

Table 2. Computation time per image (MATLAB)

Computation	Total time in ms	Self-time in ms
Read image	87 ms	11 ms
Gray scale conversion	54 ms	7 ms
IPT	680 ms	10 ms
Applying CLAHE	58 ms	2 ms
Edge detection	11 ms	2 ms
Others	290 ms	131 ms
Detected result	15 ms	12 ms

2.6 Conclusion

The proposed framework of a road marking detection approach for autonomous vehicle application, which consists of inverse perspective transform method, image normalization and edge detection method. In the inverse perspective transform method, the two-dimensional plane images transformed into three-dimensional plane images and using the rotation matrix, images were rotated 35° and then images were projected into two-dimensional plane, which was an image with a top view of road images. Various illumination conditions were tackled by normalizing the road images using CLAHE. The last stage in the algorithm is the edge detection method, which was used to detect road marking, such as forward arrow left-side arrow, right-side arrow, lanes and signs printed on road. It is shown that the techniques proposed in this paper are promising as the percentage of the true detection for the road images that are captured under various imaging conditions and with a complex background is 95.5%, and the percentage of the true detection for the images without the complex background is up to 99%. The limitation of the method is to detect lane mark with having soft yellow color. The proposed areas for future improvement are incorporate with HSI color analysis into the algorithm, in order to tackle the problem of detecting lane mark in soft yellow color and implementation of fuzzy logic algorithm to recognize road mark, road boundary and front vehicles.

References

1. Song, W., Fu, M., Yang, Y., Wang, M., Wang, X., Kornhauser, A.: Real-time lane detection and forward collision warning system based on stereo vision. In: Intelligent Vehicles Symposium (IV), pp. 493–498. IEEE (2017). ISBN 978-1-5090-4804-5
2. Jong, P., Hyun, C., Se-Young, O.: Real-time vehicle detection in urban traffic using AdaBoost. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, 18–22 October 2010
3. Kortli, Y., Marzougui, M., Atri, M.: Efficient implementation of a real-time lane departure warning system. In: International Image Processing, Applications and Systems (IPAS), pp. 1–6 (2016). ISBN 978-1-5090-1645-7
4. Wang, J., Mei, T., Kong, B.., Wei, H.: An approach of lane detection based on inverse perspective mapping. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), Qingdao, 8–11 October 2014
5. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**, 285–296 (1975)
6. Li, Z., Cai, Z.-X., Xie, J., Ren, X.-P.: Road markings extraction based on threshold segmentation. In: Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1924–1928, 29–31 May 2012
7. Huang, J., Liang, H., Wang, Z., Mei, T., Song, Y.: Robust lane marking detection under different road conditions. In: International Conference on Robotics and Biomimetics (ROBIO) (2013)
8. Wang, H., Shao, S.L.: Lane markers detection based on consecutive threshold segmentation. Adv. Mater. Res. **317–319**, 881–885 (2011)

9. Chang, C.-Y., Lin, C.-H.: An efficient method for lane-mark extraction in complex conditions. In: International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing (2012). ISBN 978-1-4673-3084-8
10. Sun, T.-Y., Tsai, S.-J., Chan, V.: HSI color model based lane-marking detection. In: 2006 IEEE Intelligent Transportation Systems Conference, Toronto, 17–20 September 2006
11. Aly, M.: Real time detection of lane markers in urban streets. In: IEEE Intelligent Vehicles Symposium (2008)
12. Moon, H.C., Min, K.M., Kim, J.H.: Vision system of unmanned ground vehicle. In: International Conference on Control, Automation and Systems (2008). In COEX
13. Leng, Y.-C., Chen, C.-L.: Vision-based lane departure detection system in urban traffic scenes, in Control. In: 11th International Conference on Automation Robotics Vision (ICARCV), pp. 1875–1880 (2010)
14. Liu, G., Wo, F., Markelic, I.: Combining statistical Hough transform and particle filter for robust lane detection and tracking. In: IEEE Intelligent Vehicles Symposium, University of California (2010). ISBN 978-1-4244-7868-2
15. Suchitra, S., Satzoda, R.K., Srikanthan, T.: Detection & classification of arrow markings on roads using signed edge signatures. In: Intelligent Vehicles Symposium, Alcalá de Henares, pp. 796–801 (2012)
16. Dornaika, F., Álvarez, J., Sappa, A.D., López, M.: A new framework for stereo sensor pose through road segmentation and registration. IEEE Trans. Intell. Transp. Syst. **12**(4), 954–966 (2011)
17. Álvarez, J., López, M., Baldrich, R.: Shadow resistant road segmentation from a mobile monocular system. In: Martí, J., et al. (eds.) IbPRIA 2007, Part II, Part of the Lecture Notes in Computer Science book series (LNCS), vol. 4478, pp. 9–16 (2007)
18. Alvarez, J.M., López, A., Baldrich, R.: Illuminant-invariant model-based road segmentation. In: IEEE Intelligent Vehicles Symposium, Eindhoven University of Technology, Eindhoven, pp. 1175–1180, 4–6 June 2008
19. Broggi, A.: An image reorganization procedure for automotive road following systems. In: Proceedings of the International Conference on Image Processing, vol. 3, pp. 532–535 (1995)
20. Hill Jr., F.S.: Computer Graphics: Using OpenGL. Prentice-Hall, New Jersey (2001)
21. Ballard, D.H., Brown, C.M.: Computer Vision. Prentice-Hall, New Jersey (1982)
22. Venkatesh, M., Vijayakumar, P.: Transformation Technique, 5 May2012. ISSN 2229-5518
23. Li, H., Feng, M., Wang, X.: Inverse perspective mapping based urban road markings detection. IEEE (2012). ISBN 978-1-4673-1857-0/12
24. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. **39**(3), 355–368 (1987)
25. Vincent, O.R., Folorunso, O.: A descriptive algorithm for Sobel image edge detection. In: Proceedings of Informing Science & IT Education Conference (InSITE) (2009)
26. Bawa, S., Singh, K.: Edge based region growing, Thapar Institute of Engineering & Technology thesis report (2006)
27. Izzati, N., Pandak, J.: Safety distance awareness system for Malaysian driver. Report submitted (2013)
28. The two-second rule: Road safety authority (Government of Ireland). Accessed 13 Dec 2011
29. The 2 second rule. http://www.drivingtesttips.biz/2-second-rule.html. Accessed 25 Sept 2012

Time Delay Modeling for Energy Efficient Thermal Comfort Control System in Smart Home Environment

Yuto Lim(✉) and Yasuo Tan

Japan Advanced Institute of Science and Technology (JAIST),
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{ylim,ytan}@jaist.ac.jp
http://www.jaist.ac.jp/

Abstract. The design of the control system is a crucial point for improving the thermal comfort level in the smart home environment. Compared to the conventional temperature control system, a thermal comfort control (TCC) system can provide a better human comfort. Due to system complexity, the TCC system is usually designed as a hybrid system. To ensure the design of a highly energy efficient thermal comfort control (EETCC) system, Cyber-Physical Systems (CPS) can offer numerous opportunities. This paper addresses the time delay modeling issues of the EETCC system. Following the execution model of the programming temporally integrated distributed embedded systems (PTIDES), the real-time requirements of execution tasks can be guaranteed. In addition, as the task dependency relations widely exist in the practical applications, by using a directed acyclic graph and the proposed schemes, these task dependency relations can also be dealt properly.

Keywords: Smart homes · Cyber-Physical Systems
Thermal comfort control · Time delay model
Task dependency relation

1 Introduction

Nowadays, the advanced sense-control-actuate technologies allow us to live more comfortable, cost-effective, more convenient, more secure and safe, and more assisted caring and watching in the smart home environment. In the meantime, the demands such as control loads, anticipating home appliance and equipment failures, efficient energy consumption, and so on increase dramatically. In particular, to control the home thermal comfort at the desired level while optimizing the energy consumption at the same time is a challenging issue. This is because (i) the dynamic change of outside environment that can be used as one of the thermal resources for controlling the home thermal comfort level; and (ii) the preference of thermal comfort level for each occupant is different.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 42–52, 2018.
https://doi.org/10.1007/978-981-10-8276-4_5

Some related works of thermal comfort control system in smart home environment can be found in [1–3]. Oliveria et al. [1] present a load management mechanism that allows inhabitants to adjust power consumption according to expected comfort and energy price variations. Molina et al. [2] introduce an optimization and control algorithm for residential temperature regulation. Concepts from system identification, model predictive control (MPC) and genetic algorithms are incorporated with to compromise between comfort and cost in the presence of time-varying electricity prices. Vazquez and Kastner [3] propose a smart home application based on habit profiles for the thermal comfort support. It pursues to keep the environment as close as possible to the desired air quality and thermal comfort conditions by means of low energy cost strategies taking into account the management of shading devices, automated windows and dampers.

Unlike these works, to develop the practical application for thermal comfort control system in smart home environment, in our previous work [4], based on the concept of Cyber-Physical Systems (CPS), an energy efficient thermal comfort control (EETCC) system was proposed. The EETCC system is defined as a hybrid system, which combines a supervisory controller and proportional-integral (PI) controllers. Predictive Mean Vote (PMV) index is adopted to evaluate home thermal comfort level. We correlate the room thermal environment with multiple actuators (objects), which can potentially change the home thermal environment. Our proposed EETCC system enables to monitor and maintain the desired PMV value dynamically with three actuators: air-conditioner, window and curtain. In this paper, the objective is to address the time delay modeling problem of the EETCC system using the programming temporally integrated distributed embedded systems (PTIDES) model. The real-time requirement of tasks can be guaranteed by using a directed acyclic graph and the proposed schemes to determine the holding time for the right order task must be executed first.

The rest of this paper is structured as follows. Smart homes, thermal comfort, and CPS are briefly described in Sect. 2. In Sect. 3, the EETCC system is explained. Section 4 describes how to use the PTIDES model for the EETCC system. Section 5 explains the simulation setup, result and discussion. Concluding remarks are given in Sect. 6.

2 Research Background

2.1 Smart Homes

In the past decades, the smart home concept evolved rapidly with the Internet growth in which the idea of a house is full of interconnected devices. The term of smart homes was first officially used by the American Association of House Builders in 1984 with the aim of pushing forward the inclusion of the necessary technologies into the design of new homes by a special interest group called *Smart House* [5]. Smart house is instrumented with various sensors and networking infrastructure and the major problem in this area is the integration

and interaction among heterogeneous applications provided from different vendors. Smart homes is a home-like environment that possesses ambient intelligence and automatic control, which allow it to respond to the behavior of residents and provide them with various facilities [6]. The major targets of smart homes are improving thermal comfort, dealing with medical rehabilitation, monitoring mobility and physiological parameters, and delivering therapy. Thus, smart homes require a strong integration of technology and services through home networking for a better quality of living (QoL).

The current implementation of smart home environment, i.e., *iHouse*, which is shown in Fig. 1. *iHouse* that stands for Ishikawa, internetted, inspiring, and intelligent House is an advanced experimental environment for future smart homes in Japan and it has been implemented according to Standard House Design by Architectural Institute of Japan. The location of *iHouse* is built at Nomi city, Ishikawa prefecture. *iHouse* consists of sensors, electronic devices and home appliances that are connecting to each others by utilizing ECHONET Lite version 1.1 and ECHONET version 3.6. This configuration network emanates more than 300 sensors and actuators.

Fig. 1. Experimental smart home environment – *iHouse*

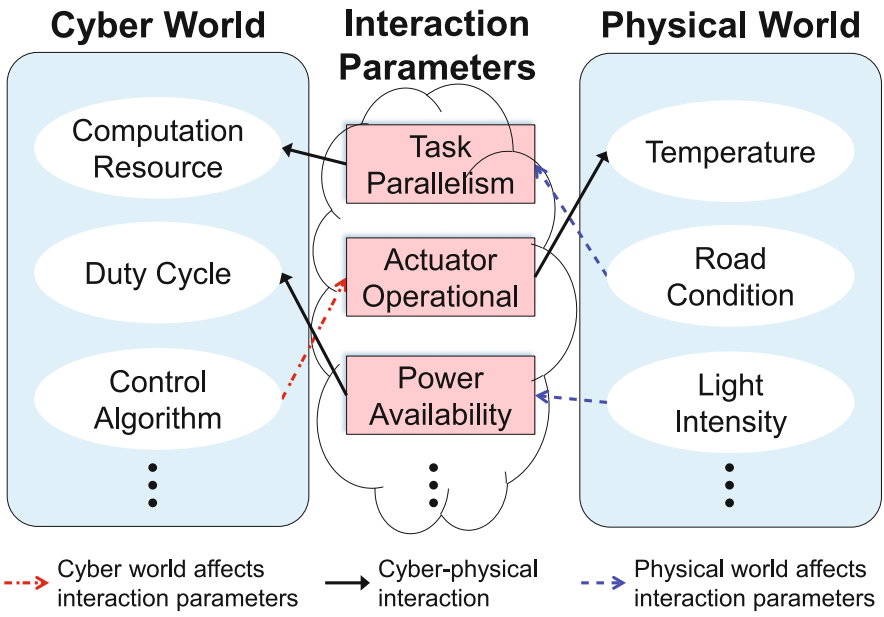


Fig. 2. Concept model of CPS

2.2 Thermal Comfort

Thermal comfort is a condition of mind that expresses satisfaction with the thermal environment [7]. A thermally comfortable home environment makes for occupant health and comfort. A control system that is based on the thermal comfort can much significantly improve building energy efficiency than ones maintaining one or more of thermal factors like air temperature, humidity, and air velocity at the desired level. Thus, improving thermal comfort and designing its control system have becoming an important concern.

2.3 Cyber-Physical Systems

Cyber-Physical Systems (CPS) in [8,9] is defined as a tight integration of computation, communication, and control for active interaction between physical and cyber elements in which embedded devices, such as sensors and actuators, are wireless or wired networked to sense, monitor and control the physical world. With CPS becoming more popular, many researches have devoted to their development. Researches in [10,11] have contributed to the modeling of CPS. By referring to [10], a concept model of CPS is depicted in Fig. 2. CPS consists of two worlds: cyber world and physical world. Elements in cyber world are properties of computing units which are embedded in the system. Such elements include computation resource, duty cycle of a sensor, control algorithm, etc. Elements in physical world are physical properties, such as temperature, road condition,

and intensity of sunlight. The two worlds are connected through interaction parameters that are associated with both the computing and physical properties. Examples of interaction parameters include task parallelism, operation of actuators, and available power energy for computing units.

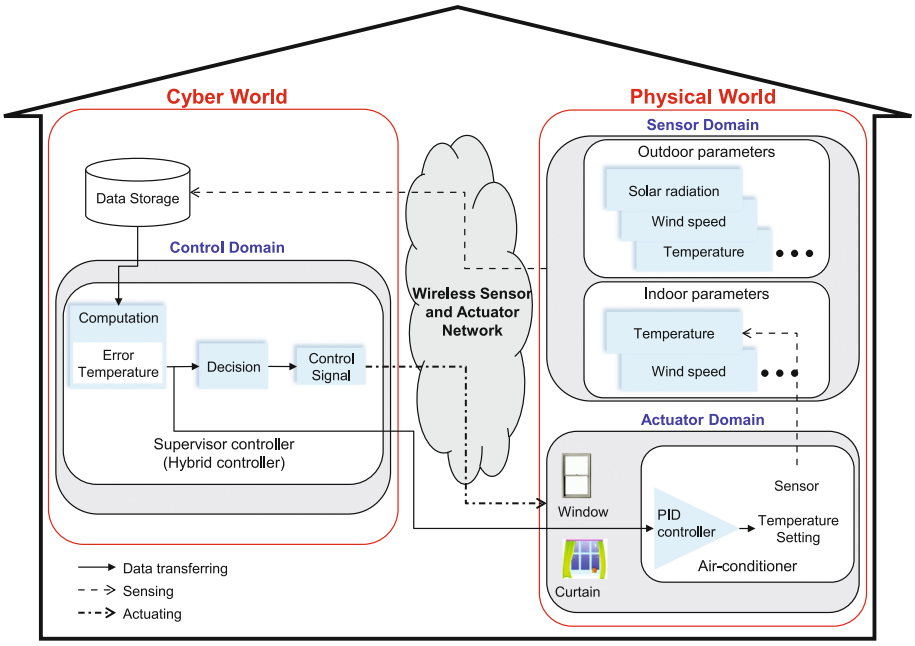


Fig. 3. Architecture of EETCC system

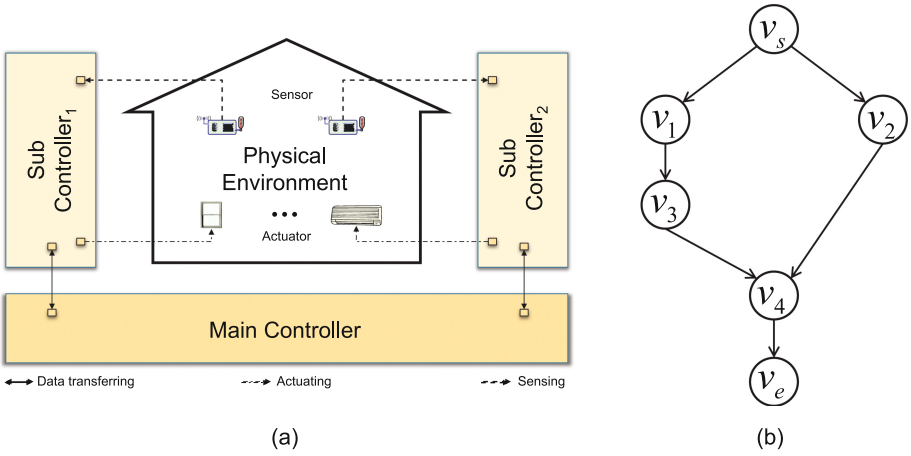


Fig. 4. (a) Schematic control diagram of EETCC system and (b) Task dependency graph

3 Energy Efficient Thermal Comfort Control System

A basic architecture of the energy efficient thermal comfort control (EETCC) system is shown in Fig. 3. In this architecture, both cyber world and physical world are defined. To connect these two worlds, a communication network of wireless sensor and actuator network (WSAN) is used. In particular, the WSAN comprises two components: sensors and actuators. The sensors in the physical side send the environment factor (temperature, air speed, etc.) both for room and outside periodically to a supervisory controller, which is using EETCC algorithm and a proportional-integral (PI) controller on the cyber side. Subsequently, the supervisory controller computes an error PMV value, which is a difference value between the current PMV value and the desired PMV value. Based on the error PMV value and outside environment factors, the supervisory controller computes a control signal to activate the appropriate states of those actuators (i.e., air-conditioner, window and curtain) in order to influence the level of thermal comfort. The operations of window and curtain are open or closed when air-conditioner is triggered. Two steps are needed for the cascade PI controller to decide a control input. First, the supervisory controller computes the error PMV, which is fed to the primary PI controller. Then, the primary PI controller calculates the error temperature, which is used as the desired temperature value of the secondary PI controller. Subsequently, based on this value, the secondary PI controller generates an air-conditioner setting temperature as a control input to operate air-conditioner.

3.1 State Description

The EETCC system is organized in a set of six states, which are represented as numerical number 0–5, respectively, as shown in Table 1. Considering the energy efficiency as a first priority, turning on the air-conditioner while opening the window at the same time is not allowed in the EETCC system.

4 PTIDES Model for EETCC System

A programming temporally integrated distributed embedded systems (PTIDES) [12,13] is a model for a deterministic modeling paradigm suitable for CPS applications at any scale. It follows discrete-event (DE) semantics. For a PTIDES model, the physical realization must satisfy the following assumptions: *(i)* Clocks are synchronized with a known bound on the synchronization error; *(ii)* Every communication channel has a known bound on its latency; and *(iii)* The time taken by any computation that may affect the physical world has a known bound. More details can be referred to [12,13].

The PTIDES model can be used to address the time delay modeling problem of EETCC system. Without loss of generality, Fig. 4(a) that illustrates a general control schematic diagram of EETCC system consists of sensors, actuators, sub controllers, and a main controller are connected through network channels (wired

Table 1. State of EETCC system

Status	Air-conditioner	Window	Curtain
State 0	×	×	×
State 1	×	×	◯
State 2	×	◯	×
State 3	×	◯	◯
State 4	◯	×	×
State 5	◯	×	◯

× means OFF/closed; ◯ means ON/open

or wireless). The state of EETCC system is controlled by the main controller (a.k.a. supervisory controller), as MC. The main controller decides the state of EETCC system, which is based on the physical environment. Actuators (e.g., air-conditioner, window and curtain) are controlled by two sub controllers, represented by a primary sub controller, as SC_1 and a secondary sub controller, as SC_2, respectively. Following some control algorithms (e.g., PI controller), both sub controllers can properly operate the actuators. For example, turning on or off the air-conditioner. When a signal or a data frame transfers from one element to another element (e.g., from a sub-controller to the main controller), there exists a data transfer delay. The symbols δ_1 and δ_2 are used to represent the data transfer delay from SC_1 to MC and from SC_2 to MC, respectively. This data frame triggers a task that needs to be processed at the receiving side. As a result, a task dependency relation widely exists in practical applications. A task dependency graph is used to illustrate the relationship of the task dependency. Figure 4(b) shows a task dependency graph with six tasks and six direct dependency relations.

4.1 Task Dependency

Definition (task dependency graph). *A* task dependency graph is a directed acyclic graph. $G = (V, E, v_s, v_e)$, *where* V is a task (node) set, $E \subseteq V \times V$ *is* a dependency relation (edge) set, with $(v_i, v_j) \in E$, $v_i \neq v_j$, *where* $v_i, v_j \in V$. $v_s \in V$ *is* the start task, and $v_e \in V$ *is* the end task.

An edge (v_i, v_j) in the task dependency graph means task v_j can start to execute only after task v_i has been completed. $v_i \prec v_j$ is used to illustrate this dependency relation. The dependency relation is transitive. That is, $v_i \prec v_j, v_j \prec v_k \implies v_i \prec v_k$.

4.2 Time Label Assignment

In the PTIDES model, task execution order follows the time label assignment to tasks. Task τ_i will be executed before the execution of task τ_j when the time label assignment to task τ_i, represented by Γ_i, is less than the time label assignment

to task τ_j, represented by Γ_j. In this paper, only one processor is considered in the EETCC system, in which it can be used by both sub controllers to execute their tasks. Since the tasks may have dependency relations with each other, if $\tau_i \prec \tau_j$, it should be guaranteed that task τ_j can start to be executed only after task τ_i has been completed. Note that, as elements in the system are connected through network channels, the data transfer delay should be considered.

Considering a scenario that the current state of EETCC system is at the State 2. Now, the main controller MC decides that the EETCC system should change to State 4. This means that the secondary sub controller SC_2 should turn on the air-conditioner, and the primary sub controller SC_1 should close the window. In the status of the air-conditioner is turned on, opening window is not allowed. Thus, the EETCC system needs to ensure that only after the task (represented by τ_1) closing the window has been completed, the task turning on the air-conditioner (represented by τ_2) can start, that is $\tau_1 \prec \tau_2$. Assuming the data transfer delays from MC to SC_1 and to SC_2 are δ_1 and δ_2, respectively, while the data transfer delays from SC_1 to window and from SC_2 to air-conditioner are δ_1' and δ_2', respectively. The time label assignment to task τ_1, represented by Γ_1, and the time label assignment to task τ_2, represented by Γ_2, should satisfy:

$$\Gamma_1 < \Gamma_2 \tag{1}$$

This equation makes sure that when both tasks τ_1 and τ_2 are arriving in the main controller, task τ_1 will be executed first.

To illustrate the time label assignment to tasks, τ_1 and τ_2 that are issued by the main controller at system time t_0 are assumed. Then, the time instant of task τ_1 arriving at the primary sub controller SC_1 is $t_0 + \delta_1$. Let use symbol s_1 and s_2 represent the start execution time of task τ_1 and τ_2, respectively. The EETCC system should satisfy $s_1 < s_2$, which results

$$t_0 + \delta_1 < s_2 \tag{2}$$

This equation guarantees that when task τ_2 starts to be executed, the task τ_1 has arrived in the primary sub controller SC_1.

4.3 Safe-to-Process Analysis

The environment sensors keep on sensing physical environmental factors. A task may be triggered by the sensing operation. As different sensors are connected with different sub controllers, when the sensed data are transferred to the main controller, the transfer delay can be different as well. For example, when a sensor is connected with a sub controller SC_i, the sensing data at system time ts_i, the time instant when the data is transferred to the main controller, represented by t_i, equals to $ts_i + \delta_i + \delta_i'$, where δ_i represents the transfer delay when the data frame is transferred from the sub controller to the main controller SC_i and δ_i' represents the transfer delay when the sensed data is transferred from the sensor to the sub controller SC_i.

When two data frames are transferred to the main controller MC from two sub controllers, SC_1 and SC_2, the EETCC system needs to guarantee the execution of the tasks triggered by the sensed data, represented by τ_1 and τ_2, respectively, can be performed following the sensed time instant. This is called safe-to-process analysis [13]. According to this time requirement, if task τ_1 executes before task τ_2, it means the sensed time instant ts_1 with respect to τ_1 is earlier than the sensed time instant ts_2 with respect to task τ_2. Based on the analysis above, when task τ_1 executes before task τ_2, the following equality is as below:

$$ts_1 < ts_2 + \delta_2 + \delta_2' - \delta_1 - \delta_1' \tag{3}$$

Now considering the case of $t_2 < t_1$ when two tasks τ_1 and τ_2 are arriving at the main controller. This means that task τ_2 will be executed before task τ_1. As a result, it is not safe to process the task of the smallest timestamp because of the problem of task dependency. Therefore, a scheduler of the main controller should wait for a specific time upon the arrival of task τ_1, so that task τ_1 must be executed first. The specific time o is defined as an offset time for first arrival task to be waited after the second arrival task has been received. The scheduler can use two different schemes to determine the length of the specific time. First scheme is based on the root mean square (RMS) of a set of n time different values $\{x_1, x_2, \ldots, x_n\}$ of both task arrivals when the case of $t_2 < t_1$. The first scheme is called as a RMS-based scheme. The specific time length of the RMS-based scheme is given by

$$o_{rms} = t_2 + \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} \tag{4}$$

Meanwhile, second scheme is based on double average (DA) of a set of n inter-arrival time values $\{i_1, i_2, \ldots, i_n\}$ of task that must be completed first. The second scheme is named as a DA-based scheme. The specific time length of the DA-based scheme is given by

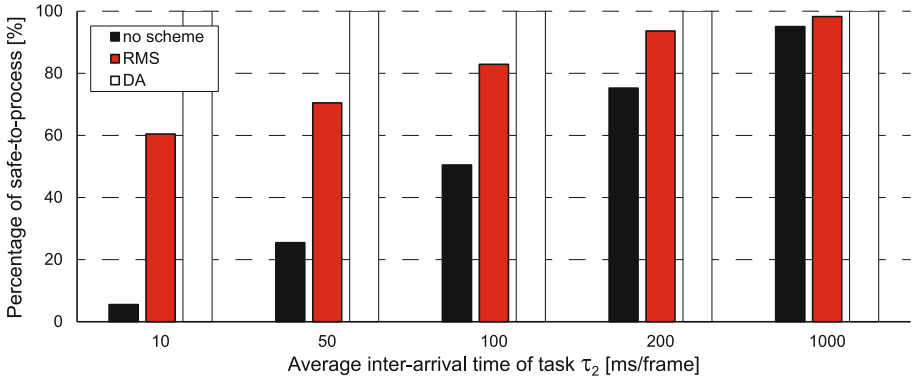


Fig. 5. Safe-to-process performances of RMS-based and DA-based schemes

$$o_{da} = t_2 + \frac{2(i_1 + i_2 + \cdots + i_n)}{n} \tag{5}$$

5 Numerical Evaluation

This section examines how RMS-based and DA-based schemes are implemented for comparison of safe-to-process performances under random arrival processes of average inter-arrival time of two tasks. The simulation scenario is based on Fig. 4(a), in which two sub controllers are assumed to send their data frames as tasks, respectively to the main controller. A data frame that has a fixed length will be received by the main controller according to a memoryless Poisson process with an average inter-arrival time of Δt ms per frame. In our simulation, the average inter-arrival time of task τ_1 is fixed at 100 ms per frame, whereas the average inter-arrival time of task τ_2 varies from 10 to 1000 ms per frame. Each sub controller sends one million data frames and the percentage of safe-to-process is computed. All the simulation results is collected using a 64bit Intel Core i7-5600U vPro CPU 2.60 GHz with 16 GB of memory.

In Fig. 5, the percentage of safe-to-process is very low when the average inter-arrival time of task τ_2 is at 10 ms per frame with no scheme. When the RMS-based scheme is applied, the percentage is increased to 60.43%. On the other hand, the DA-based scheme can guarantee 100% of safe-to-process regardless of any average inter-arrival time of task τ_2. The disadvantage of the DA-based scheme is that it requires more time to ensure the tasks to be completed.

6 Concluding Remarks

In this paper, we have addressed the time delay modeling issues of the EETCC system by using the time label assignment of the PTIDES model. Following the task execution of the PTIDES model, the real-time requirements of tasks can be guaranteed. This is because the task dependency relations can be dealt by using the directed acyclic graph and the DA-based scheme. An important future work is to evaluate the PTIDES model in our experimental smart home environment, i.e., *iHouse* to further investigate the actual factors that can cause the issues of the time delay in the EETCC system.

Acknowledgment. This work was supported in part by a Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS), Japan. Grant number is 15K00120.

References

1. Oliveria, G.D., Jacomino, M., Ha, D.L., Ploix, S.: Optimal power control for smart homes. In: International Federation of Automatic Control (IFAC) World Congress, vol. 44, no. 1, pp. 9579–9586 (2011)
2. Molina, D., Lu, C., Sherman, V., Harley, R.G.: Model predictive and genetic algorithm based optimization of residential temperature control in the presence of time-varying electricity prices. IEEE Trans. Ind. Appl. **49**(3), 1137–1145 (2013)
3. Vazquez, F.I., Kastner, W.: Thermal comfort support application for smart home control. In: Novais, P., Hallenborg, K., Tapia, D., Rodrguez, J. (eds.) Ambient Intelligence - Software and Applications. Advances in Intelligent and Soft Computing, vol. 153, pp. 109–118. Springer, Heidelberg (2012)
4. Cheng, Z., Shein, W.W., Tan, Y., Lim, A.O.: Energy efficient thermal comfort control for cyber-physical home system. In: IEEE International Conference on Smart Grid Communications, pp. 797–802 (2013)
5. Harper, R.: Inside the Smart Home. Springer, London (2003)
6. De Silva, L.C., Morikawa, C., Petra, I.M.: State of the art of smart homes. Elsevier J. Eng. Appl. of Articial Intell. **25**(7), 1313–1321 (2012)
7. ANSI/ASHRAE Standard 55-2010: Thermal Environmental Conditions for Human Occupancy (2010)
8. Lee, E.A.: CPS foundations. In: ACM/IEEE Design Automation Conference (DAC), pp. 737–742 (2010)
9. Wan, K., Man, K.L., Hughes, D.: Specification, analyzing challenges and approaches for cyber-physical systems (CPS). Eng. Lett. **18**(3), 305–308 (2010)
10. Banerjee, A., Venkatasubramanian, K.K., Mukherjee, T., Gupta, S.K.S.: Ensuring safety, security, and sustainability of mission-critical cyber-physical systems. In: Proceedings of the IEEE, vol. 100, no. 1, pp. 283–299 (2012)
11. Derler, P., Lee, E.A., Vincentelli, A.S.: Modeling cyber-physical systems. In: Proceedings of the IEEE, vol. 100, no. 1, pp. 13–28 (2012)
12. Zou, J., Matic, S., Lee, E.A., Feng, T.H., Derler, P.: Execution strategies for PTIDES, a programming model for distributed embedded systems. In: IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), pp. 77–86 (2009)
13. Zhao, Y., Liu, J., Lee, E.A.: A programming model for time-synchronized distributed real-time systems. In: IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), pp. 259–268 (2007)

Energy Management Techniques for RF-Enabled Sensor Networks Based on Internet of Things

Shaik Shabana Anjum[1(✉)], Rafidah Md Noor[1], Ismail Ahmedy[1], Mohammad Hossein Anisi[2], and Norazlina Khamis[3]

[1] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
anjumjavid@siswa.um.edu.my, {fidah,ismailahmedy}@um.edu.my
[2] School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
m.anisi@essex.ac.uk
[3] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
norazlina@ums.edu.my

Abstract. The vision of ubiquitous computing is based on the fact that future computers will merge up with the surrounding environment in IoT domain. Radio Frequency Identification (RFID) and Wireless Sensor Networks (WSN) are two important cornerstones for pervasive computing as they combine the physical and virtual world, thereby bridging the gap between cyber space and physical world of real things. RFID enables the identification and detection of entities while WSN are used to sense the condition of the environment or object. The integration of RFID and WSN have paved way for the existence of RSN (RFID Sensor Networks), thereby providing extended capabilities, scalability, portability, lower cost and novel perspective towards a broad range of applications. This paper presents a brief introduction about the evolution of RSN, major issues in RSN and energy management with regards to Energy Harvesting (EH), energy request and transfer. It also investigates into the problems encountered for efficient energy transmission. Differently from the classic schemes in the literature that deals with scalability, security and communication protocol aspects, the proposed methodology focuses on energy management issue which is of utmost importance for wide area RSN. Furthermore, the paper provides insights into the preliminary experimental evaluation and its comparative analysis with existing schemes pertaining to performance metrics.

Keywords: IoT · RFID · WSN · Energy management · Energy harvesting Energy transfer

1 Introduction

Internet of Things (IoT) can be defined as a paradigm where every day physical objects are connected through internet. The successful evolution of IoT vision, has paved way for computing portables and smart-phones as an extension of past traditional scenarios. The evidences of such evolution can be witnessed in the growing presence of 4G-LTE

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 53–63, 2018.
https://doi.org/10.1007/978-981-10-8276-4_6

(4th Generation Long Term Evolution) and Wi-Fi. Such a seamless integration of physical objects into information network provides intelligent and ubiquitous services providing a promising future in fields of surveillance, health care, security, transport, food safety, object monitoring and control. In IoT environment, the connected web is a highly distributed network consisting of dynamic services, information providers and data consumers who share the information. Compared to the traditional desktop realm, IoT based services system faces major risks of handling the dynamic services and heterogeneity of devices. To meet this challenge, it is envisioned that sensor networks and RFID technology will become increasingly integral to the human environment, where communication and information systems will be invisibly integrated. This need makes WSN and RFID as two major aspects of IoT.

Radio Frequency Identification (RFID) and Wireless Sensor Networks (WSNs) represent two most prominent technologies that have a wide range of applications. These two ubiquitous computing based technologies have gained considerable attention in potential research and development fields. RFID applications include supply chain management, manufacturing, search and rescue and on the other hand, WSN technology is used for sensor mote deployment to monitor air pollution and for battlefield surveillance. The integration of RFID and WSNs has paved way for the evolution of RFID sensors Networks (RSNs). RFID is used to track or to locate an object identity without providing any traces about the physical environment of the object. WSNs on the other hand are networks of small interconnected devices that are incorporated to collect information by sensing the environmental conditions of the surroundings like temperature, light, humidity, pressure, vibration and sound. These two technologies provide extended capabilities, enhanced efficiency, cost effectiveness and eventually bridge the gap between real and virtual world, on an integrated perspective. The requirements for the development of RSN include accurate communication, reliability, energy efficiency, network maintenance survivability, tolerable latency and criticality of the application. Energy efficiency is one of the most attention seeking limitations because both sensor nodes and RFID tags comprises of scarce resources. The practical issue lies in the fact of periodical recharging of these nodes, which puts forth the challenge of effective deployment of large scale RSNs. An auxiliary solution for efficient integrated system is energy harvesting and recharging. Different sources of energy exist in different forms (e.g. light, vibration, air, and electromagnetic waves) [2] as shown in Fig. 1. These sources can be harvested and used either to extend battery life of a sensor node or power a sensor node directly without any storing techniques. Reducing or eliminating the problem of limited lifetime will enable node designers to enhance the functionality of a node by adding extra features and components. A sensor node comprises of four basic components with additional units being added depending on application requirements [1]. The basic components of sensor node consist of a sensing unit used for acquiring data from the environment and converting it to digital data, a processing unit for processing raw data to store the results, and a transceiver unit for sharing data with other nodes or the end-user, a power unit that consists of an energy sink (battery, capacitor or both) and power management that monitors and routes power to the entire node. The lifetime of a sensor node depends on the capacity of the power resources it is equipped

with. One way of prolonging lifetime of sensor network would be to periodically replace the batteries of all or some of the deployed sensor nodes.

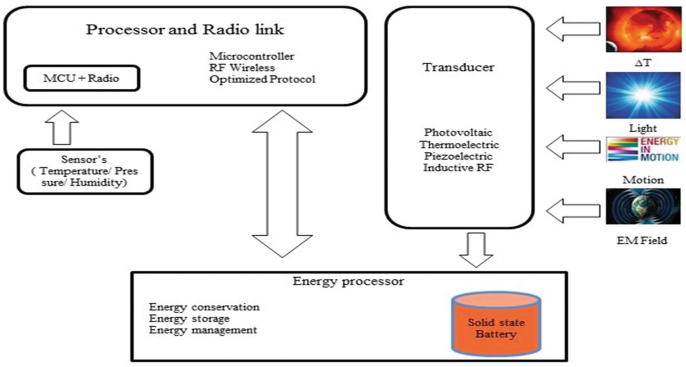


Fig. 1. Components of energy harvesting for sensor node

2 Background Study

2.1 RFID Sensor Networks

The integration of two most prominent wireless technologies RFID and WSN will enhance their effectiveness and bridge the gap between real and augmented world. This eventually results in an integrated technology with lower cost, portability, extended capabilities and scalability [7]. The specific requirements that are to be met to achieve an effective integrated RSN are accurate and reliable communication, energy efficiency, network maintenance survivability and cost considerations. The reliability of RSN is dependent upon the criticality of a specific application. Furthermore, the adoption of intrusion tolerant mechanisms will enable the integrated network to recover from Denial of Service (DoS) attacks. Several research contributions have been reported in the literature for the different types of WSN-RFID integration architectures [7, 9].

2.2 Applications and Challenges of RFID Sensor Networks

RFID and WSN have been used in many applications separately ranging from monitoring, health care to disaster management. The potential of these two promising technologies increases to greater heights when integrated and operated on a shared platform. The possible applications of RSN in healthcare can be body temperature sensing, measuring blood pressure, heart rate, pH values or any other medical ailments. The location of the patient can be traced with the help of RFID and the condition can be monitored by sensors. In the field of supply chain management, RSN have been used extensively for product tracking, asset monitoring and inventory control. These integrated networks are also employed in the field of fire detection, monitoring shipping containers, managing cattle and checking the condition of weapons in battlefield without personal

intervention. They are also used to sense the vibration levels of large cooling fans used in oil-refining processes [5].

2.3 Energy Management in RFID Sensor Networks

The emergence of wireless charging techniques provides a more flexible and promising way to solve the energy constrain problem in Wireless Rechargeable Sensor Networks (WRSNs) [3, 4]. Although many researches have been conducted on wireless charging algorithms, most of them only focus on passively replenishing energy for nodes with insufficient energy [6]. When the energy of a sensor node is depleted, it will no longer fulfil its role in the network unless either the source of energy is replaced or some harvesting mechanism is induced to bridge the energy gap. The existing solutions utilize energy source powered by batteries in sensor nodes of RSN's, but are associated with many drawbacks, like chemical leakages, extreme weather conditions and limited energy density [6]. The problem of finite node lifetime is addressed using energy harvesting [11]. The improvement of energy efficiency of discrete-manufacturing facilities through state of art, and internet of things solution has been addressed by Shrouf and Miragliotta [10]. This enables a high level of awareness, large data collection and flexible installation of energy-related data in real time. In presence of energy harvesting, the preamble length or wait duration can be increased or decreased based on the effective energy at a node, to allow energy-scare nodes to sleep for longer durations. The drawbacks of relating expected energy levels of all nodes in a network, their routing paths and traffic pattern, with duty-cycle and MAC parameters is of potential interest for research direction [11]. The problem of energy sharing in sensors networks has been studied [8] where the authors have proposed a new technique to manage energy available through harvesting, but still needs improvisations by tuning the partition thresholds for clustering the state space.

3 Proposed System Model

The proposed system model comprises of randomly distributed static RSN nodes (W1 to W9) and mobile readers (R1, R2 and R3) grouped together in clusters (cluster 1, 2 and 3) and a master reader M along with sink as depicted in Fig. 2. The RSN nodes perform the sensing task and communicate it to the mobile reader. The reader that is mobile in nature collects the sensed data from these nodes and recharges it through the concept of ambient RF energy backscattering. The RF energy that is harvested every time when the reader comes into the proximity range of the nodes is stored in the capacitor.

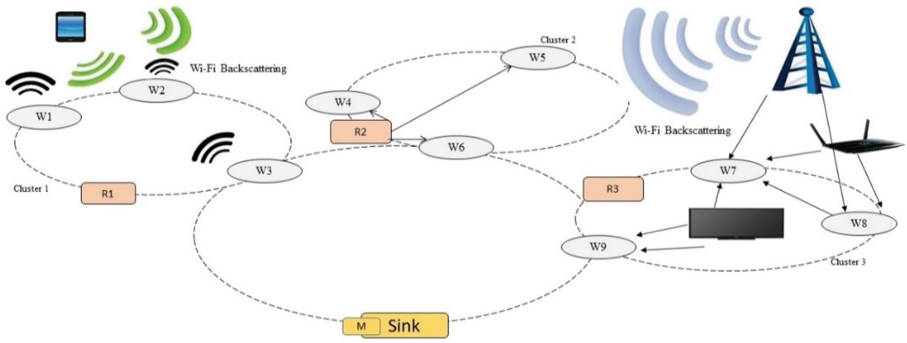


Fig. 2. Sample scenario of the proposed solution

The data collected from each cluster is delivered to the sink through the master reader (M). The sink can be considered as a RFID tag, equipped with larger memory space. As mentioned earlier, the total time required by R, to read sensed data from W is directly proportional to the memory size of W. Hence, the size being in 'B' bits, the time required to recharge the sensor node is also comparatively smaller. Therefore, for larger complex operations in wide area RSN, the nodes would need more energy requiring the mobile readers to make frequent tours for recharging cycles. This can consume more energy and can also cause node failures or packet loss in the network, if the mobile reader does not reach the recharging distance (R_d) from the energy depleted RSN node. To overcome this kind of energy constraint, multiple mobile readers can be deployed in large and wide area RSNs, so that all the 'W' nodes are visited and recharged without time delays. In such a scenario, the multiple mobile readers, will move along the unique tour, but with different timings. Hence, the proposed methodology is designed for the proper time synchronisation of RSN nodes during each path movement and recharging cycles of energy harvesting.

The main novelty of this research is the integration of RFID and WSN using network simulations for harvesting the energy of sensor nodes using RF energy. The proposed algorithm focuses on managing the energy levels of the nodes in RSN. The existing solutions of EH networks do not integrate RFID technology along with sensing capabilities. This research focuses on integrating both identification and sensing technologies at the data link layer. The network density of proposed algorithm is tested by increasing the number of readers and sensor tag nodes. The complexity lies in the fact of preventing node failure and avoiding inadequate distribution of harvested energy because RFID used high frequency whereas radio model of other networks like Adhoc On-Demand Distance Vector (AODV) routing use medium frequency for communication. The proposed scenario is generated and experimented for different network traffic (Poisson and FTP). The further work focuses on energy transfer through Wi-Fi Backscattering. Therefore, this research contributes towards maximum energy optimization by preventing energy losses, node failure and packet loss thereby improving the network lifetime and performance.

The residual energy (R_E) in RSN node is computed as,

$$R_E = 1/2\,C\left(V_x^2 - V_d^2\right) = P_x T_x \tag{1}$$

where 'C' is the capacitance of the RSN capacitor, V_x is the maximum voltage of the embedded capacitor, V_d is the minimum operational voltage of RSN node, T_x is the time for completion of fully charged capacitor, P_x is the received power of RF signal from reader. The value of P_x is calculated from Friis equation as,

$$P_x = P_T G_T G_R \lambda^2 / (4\pi d)^2\, L \tag{2}$$

where, d is the distance between the transmitting antenna and receiving antenna, P_x is the power received, P_T is the power that is transmitted, λ is the wavelength of the RF signal, L is the path loss factor, G_T is the antenna gain of the transmitter and G_R is the antenna gain of the receiver. The mobile reader should ensure the continuous operation of the RSN node and provide guarantee about the fact that the energy stored in the capacitor does not run out. The amount of time spent by the mobile reader between two consecutive tours to recharge the RSN is referred to as period of visit (T_{PV}). This is computed by summing up the total recharging time (T_x) and lifetime of RSN (T_L). T_x can be calculated using Eq. (1), whereas T_L is dependent upon the operational cycles (D) of the RSN node. These operational cycles can be calculated as,

$$D = A_T/(A_T + S_T) \tag{3}$$

where A_T is active mode time interval and S_T is sleep mode time. The total lifetime (T_L) of RSN can be calculated as,

$$T_L = R_E/A_P(D) + S_P(1 - D) \tag{4}$$

where A_P & S_p are power consumed by RSN node in active and sleep modes respectively. According to EPC C1G2 standardisation, the total time spent for read/write operations of RSN node is directly proportional to the memory size of the RSN node (B bits). The authors in this paper aim to provide an event triggered scheduling mechanism to prevent node failure and packet loss in wide area RSNs. In this way, the proposed solution is compared with existing mechanisms such as non-cooperative solution and co-operative tag-based solutions.

Algorithm 1. Event-triggered scheduling algorithm.

```
Step 1: Check status of the node

Step 2: Calculate the average residual energy of all
nodes in time specific intervals.

Step 3: Check the condition if average residual energy is
lesser than the threshold   value.

Step 4: If yes, then energize the RSN node and activate
the energy harvesting module

Step 5: Once energy level exceeds the threshold value
then send beaconing message.

Step 6: Assign duty cycling mechanisms depending upon the
queue size.

Step 7: Check for similar and multiple candidate nodes in
the same cluster for recharging deadlines.

Step 8: Determine the movement of master reader depending
upon the current energy level statistics.

Step 9: Assign weights to the nodes according to the
event triggered or application priority for communication
between cluster mobile reader and master bus reader

Step 10: Send the energy related statistics to the sink
through the master reader.
```

Differently from the existing solutions, the proposed mechanism focuses on energy harvesting aspect based upon Wi-Fi backscattering when the reader is away from the energy depleted RSN node. This ensures that the information is not lost and hence the RSN node remains active during the operational cycle, with minimal power consumption. The algorithm is applied for all the nodes. The node with highest energy and mobile in nature is considered as the cluster mobile reader. The other nodes are tag nodes with energy level more than threshold value. The master reader is the RFID reader which is an agent with functions and variables (data members and member functions). All the three RSN nodes, both cluster and master readers belong to a single network and are synchronized based on time intervals and occurrence of events. Step 1 to 6 and 11 is performed by mobile cluster reader and steps 7 to 10 are carried out by the master reader. The energy harvesting task is done by the RSN nodes. The depiction in Fig. 3 shows the proposed network model. The aspects such as network clustering, tour definition cum

path movement of mobile readers and communication between mobile readers define the recharging and data collection process between RSN nodes and readers. The proposed event-triggered scheduling algorithm focuses on scheduling the movement of mobile readers and the dynamic demand for recharging by the active nodes. This synchronisation is designed in such a manner that there is no energy overflows or node failure in the network.

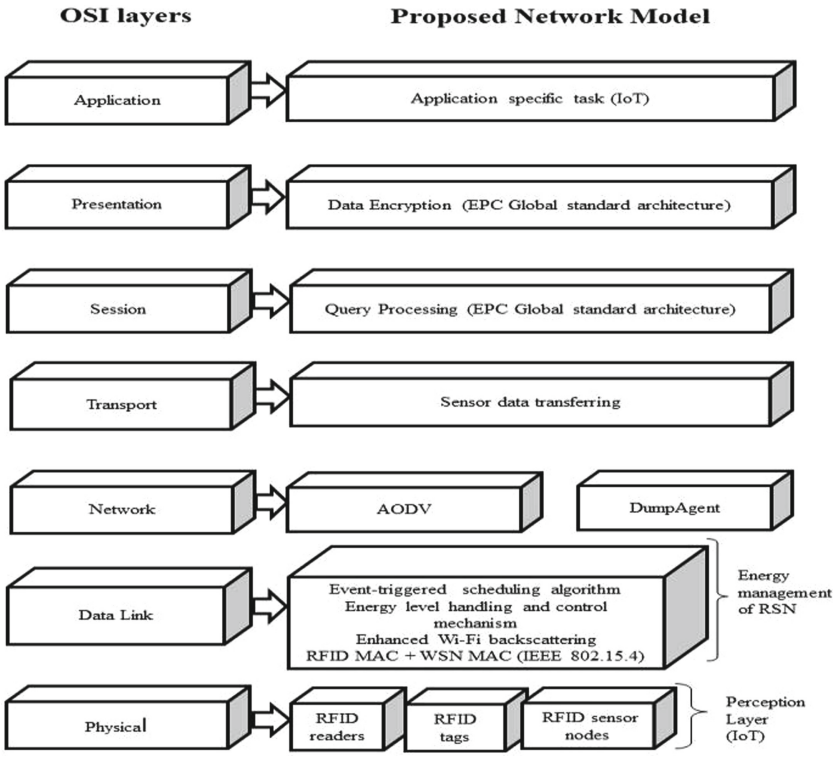


Fig. 3. The proposed network model

4 Performance Evaluation

A simulative analysis has been experimented by using simulation code developed in NS-2, to test the performance of the proposed solutions in wide RSNs as depicted in Fig. 4. The parameters of the network simulation have been tabulated in Table 1. The mechanisms are designed and proposed with the aim that all RSN nodes stay in operational state and are recharged before they fall out below their minimum threshold levels to ensure that no sensed information is lost. Thus, the evaluation of the proposed solution is done in terms of data delivery rate to the sink, the number of required mobile readers, average throughput and energy harvesting rate. It is clear from the plotted graphs as shown in Fig. 5 that the proposed mechanism outperforms the existing solutions in terms of average data delivery

delay and number of readers required for efficient data transmission to the sink. We can observe that the proposed solution guarantees lowest average data delivery delay and requires fewer readers when compared to existing techniques till date for RSNs.

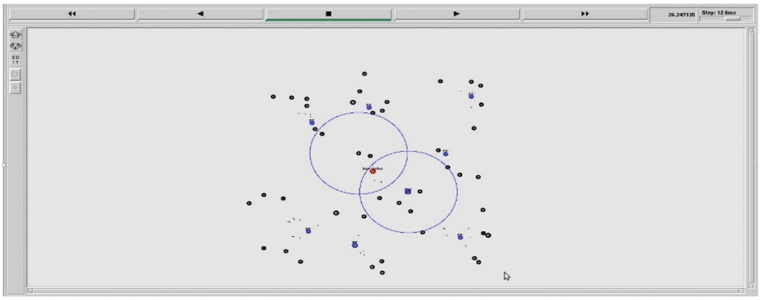
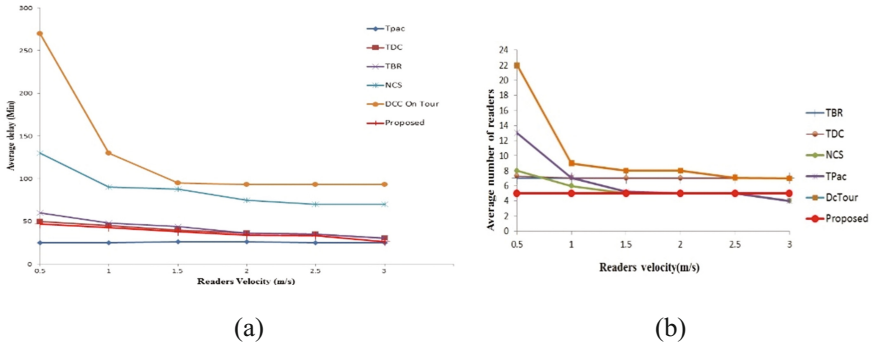


Fig. 4. Screenshot of NS-2 simulation for the proposed scenario

Table 1. Network parameters and values of simulation

Parameter	Value
Networks used	WPAN and RFID (RSN)
Topology used	Random and cluster
Number of nodes	10, 25 and 50
Traffic	Poisson/FTP
Energy level	0.5 for Transmission, 0.3 W for receiving
Harvesting energy	0.05 W to 0.3 W
Radio propagation model	Two-ray ground
Channel type and antenna model	Wireless/Omni antenna
Routing protocol	AODV and Dump agent
Period of simulation	50 s



(a) (b)

Fig. 5. (a) Performance comparison of proposed solution with existing techniques in terms of data delivery delay. (b) Performance comparison of proposed solution with existing techniques in terms of number of readers required

5 Conclusion and Future Work

This paper had provided insights and discussions about the evolution of RSN, its relative applications and challenges followed by energy management techniques in wide area RSNs. The existing solutions such as tag based relay scheme and tag based data channel scheme suffer the drawback of no communication between the RSN nodes, whereas the proposed solution on the contrary utilizes the concept of Wi-Fi backscattering for conditional beaconing by the RSN node and hence ensures nil data loss with minimal energy consumption. The first algorithm of the proposed methodology has been tested through extensive simulation platform and compared with existing techniques, to ensure improved performance and throughput of the network.

The authors are currently working on the energy transfer mechanism based on proposed enhanced Wi-Fi backscattering technique, to prevent data packet loss caused by interferences from multiple devices which emit RF signals. This request/response method ensures minimum energy consumption during waiting delay of recharging cycles by RSN nodes. Further work will evaluate and test scenarios with heterogeneous capabilities of sensing and data transmission with multiple readers and designing of individual path movement of the master bus reader towards the sink based upon aspects such as tightest time constraint and residual energy levels for real time applications in internet of things scenario.

Acknowledgement. This research is supported by Grand Challenge Grant UM.0000007/HRU.GC.SS GC002B-15SUS from Sustainable Science Cluster, University of Malaya.

References

1. Akhtar, F., Rehmani, M.H.: Energy replenishment using renewable and traditional energy resources for sustainable Wireless Sensor Networks: a review. Renew. Sustain. Ener. Rev. **45**, 769–784 (2015)
2. Ferdous, R., Reza, A.W., Siddiqui, M.F.: Renewable energy harvesting for wireless sensors using passive RFID tag technology: a review. Renew. Sustain. Ener. Rev. **58**, 1114–1128 (2016)
3. Han, G., et al.: A grid-based joint routing and charging algorithm for industrial wireless rechargeable sensor networks. Comput. Netw. **4**, 1–10 (2016)
4. He, L., Kong, L., Gu, Y.: Evaluating the on-demand mobile charging in Wireless Sensor Networks. IEEE Trans. Mob. Comput. **14**(9), 1861–1875 (2015)
5. Hussain, S., Schaffner, S., Moseychuck, D.: Applications of Wireless Sensor Networks and RFID in a smart home environment, pp. 153–157 (2009)
6. Karim, F., Zeadally, S.: Energy harvesting in Wireless Sensor Networks: a comprehensive review. Renew. Sustain. Ener. Rev. **55**, 1041–1054 (2016)
7. Liu, H., Bolic, M.: Integration of RFID and Wireless Sensor Networks integration of RFID and wireless (2015)
8. Padakandla, S., Bhatnagar, S.: Energy sharing for multiple sensor nodes with finite buffers. IEEE Trans. Commun. **63**(5), 1811–1823 (2015)
9. Pereira, D.P., et al.: Model to integration of RFID into wireless sensor network for tracking and monitoring animals, pp. 125–131 (2008)

10. Shrouf, F., Miragliotta, G.: Energy management based on internet of things: practices and framework for adoption in production management. J. Clean. Prod. **100**, 235–246 (2015)
11. Sudevalayam, S., Kulkarni, P.: Energy harvesting sensor nodes: survey and implications. IEEE Commun. Surv. Tutor. **13**(3), 443–461 (2011)

Keypoint Descriptors in SIFT and SURF for Face Feature Extractions

SukTing Pui and Jacey-Lynn Minoi(✉)

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
polly_sukting@hotmail.com, jacey@fit.unimas.my

Abstract. The last decade, numerous researches are still working on developing a robust and faster keypoints image descriptors algorithm. In this paper, we will review a few complex keypoint descriptor approaches that are well-known and commonly used in vision applications, and they are Scale Invariant Feature Transform (SIFT) and Speed-up Robust Features (SURF). These methods aim to make the descriptors faster to compute and robust to scale, rotation and noise. We will the results of the experiments on face image data. The extracted keypoints and the regions of interest are analysed and compared against the corresponding facial features. The results have shown SIFT outperformed SURF in terms of speed while the extracted keypoints using SURF descriptors are mainly located on the corners and distinct facial features.

Keywords: Keypoint descriptors · SIFT · SURF · Feature extraction

1 Introduction

Over the past decade, applications in computer vision have dramatically increased and feature extraction is still at the base of many of these applications' problem. Earlier works of research have been focusing on extracting keypoint descriptors for better accuracy and improved speed.

Detecting facial features is a vital step in face recognition for image registration purposes. The accuracy of the feature extraction process is often influenced by face variants due to the changes in orientation, head poses, illumination, facial expressions, occlusion, cluttered background and so on. A number of methods have been studied and improved feature extraction. The commonly used feature extraction approaches are the Scale Invariant Feature Transform (SIFT) [1] and the Speeded Up Robust Features (SURF) [2]. These are known to be the most promising feature extraction methods for its high performance in various applications. Therefore, we will discuss further on the performance of these methods on face feature extraction purposes.

We have conducted a number of experiments on the existing 2D face images. The keypoint descriptors and the computational speed of both methods are compared. Follows on is the analysis of the results based on the region of interests and the extracted keypoint descriptors.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 64–73, 2018.
https://doi.org/10.1007/978-981-10-8276-4_7

The outline of this paper is as follows: In Sect. 2, we present the technicalities of the SIFT and SURF methods. Then, it is followed by a discussion of the conducted experiments and the results of the extracted keypoint descriptors. Finally, the paper concludes with the analysis and future directions of the work.

2 SIFT for Face Feature Points Extraction

SIFT stands for Scale Invariant Feature Transform, which was introduced by David Lowe in 1999 [1]. This approach is aimed to transform image data into scale-invariant coordinates relative to local features. It solved the problems of image rotation, scaling, viewpoint change and affine deformation, change in 3D viewpoint and addition of noise. The distinctiveness of individual features can be matched to a large database of objects. Many features can be generated for even small objects as it can generate a large number of features. And it is close to real time performance efficiency. It was left unchallenged for almost a decade and widely used in panorama stitching, object detection, tracking and so on.

The four major stages of computation in SIFT are scale space extrema detection, keypoint localisation, orientation assignment and keypoint descriptor.

The first stage of SIFT approach is to identify the candidate keypoints. Scale space extrema detection will search the entire scale and image location. The cascade filtering approach is used to identify candidate locations that are then examined in further details. In order to detect the locations that are invariant to scale changes of an image, a continuous function of scale known as scale space, is applied. The Difference of Gaussian (DoG) functions is proposed and used to improve the computational speed. Given the scale-space of an image is defined as a function, $L(x, y, \sigma)$ that is produced from the convolution of Gaussian kernel, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{1}$$

where * is the convolution operation in x and y; σ is defined as the width of the Gaussian filter. This process down-sampled the Gaussian image by a factor of 2 in each octave, and the process is repeated. The DoG images are generated from two nearby scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{2}$$

Each of the pixels in the DoG images is compared to its 26 neighbors 3×3 regions in the scale above and below (see Fig. 1). The keypoints with low contrast, are removed and responses along edges are eliminated. A candidate keypoint (marked as 'X' in Fig. 1) is selected if the pixel is a local maximum or local minimum. The properties of the keypoint are measured to the keypoint orientation, which provides rotation invariance.

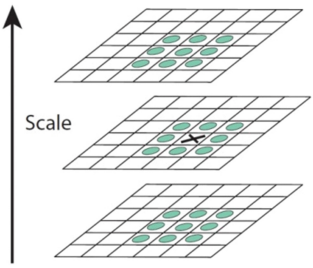


Fig. 1. Keypoint detection in scale-space (image taken from [1])

In the previous stage, a candidate keypoint has been found by comparing a pixel with its neighbors. In keypoint localisation process, the keypoints are refined and localised by rejecting those with low contrast or unused data which fall along edges. The purpose is to search for location, scale and ration of principal curvatures. Brown and Lowe [9] approached a 3D quadratic function to the local sample points. It is to determine the interpolated location of the maximum. As suggested by the author, Taylor expansion is used to accurately locate for the location of x, y and σ of the keypoint. Hessian matrix [9] is to compute the principal curvatures with a ratio of threshold at the location and scale of the keypoints. The quadratic coefficients are computed by approximating the derivatives by using differences of neighboring sample points. The candidate keypoints are chosen from the location of the extremum by rejecting unstable extrema with low contrast.

Following, the orientation assignment stage, the orientation of the keypoint is obtained based on local image gradient. An orientation histogram is formed with 36 bins to cover a 360° range of orientations. Each sample is weighted by gradient magnitude and Gaussian weighted circular window with a $\sigma = 1.5$ times of the scales of keypoint. The peaks in the orientation histogram determine the direction of local gradients. And any peaks within 80% of the highest peak are also used to create a keypoint with that orientation. Thus, there are multiple peaks created at the same location and scale, but in a different direction. This increases the stability during the matching of the keypoint descriptors. As a result, the gradients of each pixel needed to be computed and these computations are time consuming.

Once the orientation is assigned, the feature descriptor is computed as a set of orientation histograms in the pixel neighborhood. In the paper, the best results were achieved by computing the 4×4 array of histograms with 8 orientation bins in each. The descriptors of SIFT that was used is $4 \times 4 \times 8 = 128$ element feature vector. Each of the descriptors is computed by a 16×16 neighborhoods with 16 sub-blocks. The 128 dimensions form the feature vector and the keypoint which is uniquely identified by this feature vector. Then, the feature vector is normalized so that the gradient magnitude changes have no effects on illumination change. Therefore, the keypoint descriptor is invariant to affine changes in illumination.

3 SURF

Speeded Up Robust Features (SURF) is designed by Bay et al. [2] in 2008 based on the ideas of SIFT, but it is employed slightly in a different way in detecting features. It is a fast and robust algorithm for the local similarity invariant representation and feature extraction. This algorithm is invariant to a scale and in-plane rotation features. It is found to be faster and with a lesser computational speed as compared with SIFT without sacrificing the performance. The SURF algorithm consists of two consecutive stages, which are keypoint detection and keypoint description.

The originality of SURF algorithm is to achieve fast and robust descriptors. On keypoint detection stage, it is to locate the keypoint in the image. The Bay et al. [2] detected the keypoints using Hessian matrix approximation instead of DoG as in SIFT. Hessian matrix approximation based detectors are more stable and repeatable [3, 4]. It allows a fast computation of box type convolution filters. For an image I, given a point of $P = (x, y)$, the Hessian matrix $H(P, \sigma)$ in P at scale σ is defined as follows:

$$H(P,\sigma) = \begin{bmatrix} L_{xx}(P,\sigma)\, L_{xy}(P,\sigma) \\ L_{xy}(P,\sigma)\, L_{yy}(P,\sigma) \end{bmatrix} \tag{3}$$

where $L_{xx}(P, \sigma)$ in the equation is the convolution of the image with the second order derivative of the Gaussian. The location, where the Hessian determinant is the maxima form the keypoint. The approximate second order Gaussian derivatives can be evaluated by the integral images at a low computational cost (Fig. 2).

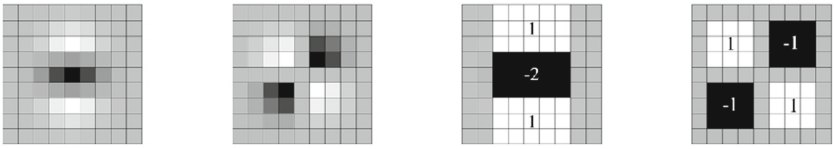


Fig. 2. Gaussian second order derivatives in *y* and *xy* directions (image taken from [2])

With the use of box filters and integral images, the same filter from the output of a previously filtered layer can be applied to the original image at any size with the same speed directly. Therefore, the scale-space is analysed by up-sampling the filter size rather than reducing the image size. In the paper [2], the authors computed 3 octaves with 3 levels. In the first octave, the construction of the box filter starts at 9×9. To change the filter size between two successive scales, an increase of 2 pixels is necessary, which yields an increased filter size by 6 pixels. In the consequent octaves, the filters with size 15×15, 21×21 and 27×27 are applied. For each new octave, the filter size is doubled for the sampling without image pyramiding. This reduces the computation time to compare to the image sub-sampling of the traditional approaches [10].

Once the keypoint localisation is completed, the keypoint descriptors must be uniquely described by a descriptor such that the correspondences between two images can be evaluated. A descriptor can be generated based on the area surrounding a

keypoint. The SURF descriptor is based on Haar wavelet responses and can be calculated efficiently with integral images with 64 dimensions. It consists of constructing a square region centred around the keypoint. The square region is divided into equally with 4×4 sub-regions. As a first step, it finds the orientation using circular window around the keypoint. For such, Haar wavelet responses are calculated in x and y direction in a circular neighborhood with radius 6 s, with s the scale at which the interest point was detected. After finding out the orientation of all keypoints, a square window is aligned to the selected orientation and extracts the descriptor from it. The descriptor vector is normalized to achieve invariance to contrast. All descriptors are then used in local feature matching. In keypoint matching step, the nearest neighbor is defined as the keypoint with a calculation with minimum Euclidean distance for the invariant descriptor vector.

4 Experiments and Results

In the experiments, we examined SIFT and SURF methods on the 2D FERET (color) face dataset [5, 6]. FERET (color) face database consists of 1199 subjects with a total of 14126 images in different appearance through time, different in lightings, clothes and hair, controlled pose variation and facial expression. We randomly selected a set of 200 images consisting of a pair of frontal-views (100 images in FA and 100 images in FB). The results are compared based on the number of extracted keypoints and total computational time. The methods are programmed in C++ on OpenCV.

4.1 Evaluating the Parameters of SIFT

The parameters of SIFT were evaluated based on those used by Lowe [1]. First, a series of Gaussian blurred images from the first image using σ, $k\sigma$, $k^2\sigma \ldots k^{n+1}\sigma$ (as in Eq. 2) are generated. A DoG image is obtained from the difference of neighboring blurred images at scales $k^n\sigma$ and $k^{n+1}\sigma$. In order to test the DoG image scales in a pyramid level, we set $k = \sqrt{2}$. As for sigma, σ, we performed different values by selecting from the range of 1.0 to 2.0, respectively. The number of octave layers was set to 3, which is suggested by Lowe [1], and computed automatically from the image resolution. On average, these results confirm that the σ values with 1.6 could correctly locate keypoints on distinct face regions. The higher value of σ is blurring the image more and more information will be eliminated. Note that if the images are captured with low resolution (weak camera) with soft lenses, we need to reduce the value of σ.

In order to filter out the edge-like features from the face images, we proceed as follows. We conducted an edge threshold tests with the given edge threshold values, T, such as T_1, T_2 … T_n (a range from 5, 10, 15 … 50) repeatedly. The keypoints that have a ratio between the principal curvatures greater than the value used are eliminated. The results showed that stable and accurate features are obtained at, $T = 10$. We noticed given the T beyond 10 retained the similar keypoints, which also increased the computational time (as seen in Fig. 3). Therefore, we can determine that T equal or less than 10 provide more reliable edge elimination on images.

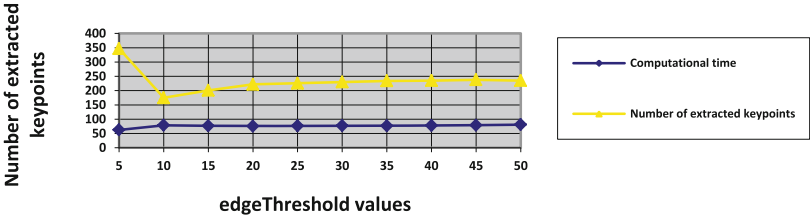


Fig. 3. A range of edge threshold values are tested corresponding with computational time on the extracted keypoints

As a result, the experiments in SIFT obtained a range of 200 to 2278 candidate keypoints from the given parameters. A total of 124 out of 200 images detected keypoints on the distinct face regions. We observed that the number of extracted keypoints within the range of 200 to 600 keypoints is located on eyes, nose and mouth (as seen in Fig. 4a). In addition, the number of keypoints, which is more than 600 keypoints displayed an over-extracted feature, where mostly hair and shirt are marked as keypoints (as seen in Fig. 4b).

4.2 Evaluating the Parameters of SURF

In SURF, we tested on a range of minimum Hessian (minHessian) values until the acceptance level of keypoints. The threshold determines how large the output of the Hessian filters to be in order for a point chosen as a keypoint. In practice, the higher the minHessian value will return a fewer keypoints, but with a more repetitive and informative keypoints on distinctive face regions. In contrast, the lower the minHessian value will return more keypoints but they may contain less information (such as the detection on cheeks, chin, hair and so on). Therefore, we have to determine minHessian to sort the resulting keypoints by its Hessian value then remove the least persistent ones. Generally, a value of minHessian between 400 and 800 works well. In this work, we extended the minHessian parameter to 1000 in order to choose the best candidate feature locations. As a result, it is found that the total extracted keypoints which is lower than 500 are mainly located on eyes, nose tip, nostril and mouth, otherwise, it is considered as excessive extraction of keypoints.

Fig. 4. The extracted keypoints by applying SIFT. 4(a). The candidate keypoints are located mostly on the distinct face features; 4(b). The 'over-extracted' features are located on shirt and hair.

4.3 Comparison the Results of SIFT and SURF

In addition, we compared the experimental results of both SIFT and SURF algorithm on the average number of extracted keypoints and the average computational speed in millisecond (ms) (as seen in Table 1). We averaged all the obtained keypoints within 200 images. We found that 655.82 and 229.65 keypoints are the average number of descriptors extracted by SIFT and SURF respectively. Both extracted descriptors are located in similar regions (as seen in Fig. 5). The number of the extracted keypoints in SIFT of each image is more than SURF. However, SIFT located numerous unwanted keypoints on the hair and shirt, whereby the keypoint descriptors from SURF are located mainly on the distinct anatomical features such as eyes, nose tip, nostrils and corners of the mouth (as seen in Fig. 6). The detection on eyes, nose and mouth are clustered as the potential landmark points about face features landmark candidate.

In previous studies [7, 8], the computation time in SURF outperforms SIFT. However, in our experiment, SIFT shows lesser computational time is used. This is because we applied a higher threshold value in Hessian detector in SURF to extract more salient keypoints using the selected data.

Table 1. Average numbers of extracted keypoints and average of computational speed

	SIFT	SURF
Avg. no. of extracted keypoints	655.82	229.65
Avg. computational time (ms)	0.856005	0.989419

Once the keypoints are extracted, we manually landmark the eye regions. Each of the distance in between both of the eyes is calculated. We compared the distance of each image on the same dataset. The results have shown that once the eyes are located reliably, it provides the constraints on the location of the nose where face images are frontal and semi-profile. Therefore, a template of face extraction of distinct face regions could be generated.

Moreover, we also observed that the edge of the foreground and the background of the image are extracted. In such, we are able to segment the area of the face and the background. This is a potential region that we are able to display as a useful region.

Fig. 5. The extracted keypoint descriptors from SIFT (row a) and SURF (row b) ('X' are labelled on the selected interest points such as eyes, nose tip, nostrils and mouth)

Fig. 6. The comparison of extracted keypoints on non-face regions by SIFT (row a) and SURF (row b)

5 Conclusion and Future Work

This paper has presented a work on SIFT and SURF algorithm on 2D face feature extraction. Based on the experimental results, we note that SIFT performed with a faster computational speed based on our experiments. Additionally, SIFT has detected more number of keypoints compared to SURF. However, those candidate keypoints extracted by SURF located mainly at the distinct facial features. This work is to provide a good start for our further work in face landmarking process. Further studies will include the detection and location of facial features by using a geometric model of face or triangular features only on the face region automatically. We will also conduct experiments on automatic 3D feature extractions.

Acknowledgement. Thanks to Dana Principal Investigator (DPI) grant.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91 (2004)
2. Bay, H., Tuytelaars, T., Gool, L.V.: SURF: speeded up robust features. In: 9th European Conference on Computer Vision, pp. 404–417 (2006)
3. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 79–116 (1998)
4. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: 7th European Conference on Computer Vision (2002)
5. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face recognition algorithms. Image Vis. Comput. J. **16**(5), 295–306 (1998)
6. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1090–1104 (2000)
7. Panchal, P.M., Panchal, S.R., Shah, S.K.: A comparison of SIFT and SURF. Int. J. Innovative Res. Comput. Commun. Eng. **1**(2), 323–327 (2013)
8. Sheena, S., Sheena, M.: A comparison of SIFT and SURF algorithm for the recognition of an efficient iris biometric system. Int. J. Adv. Res. Comput. Commun. Eng. **5**(1), 37–42 (2016)
9. Brown, M., Lowe, D.: Invariant features from interest point groups. In: BMVC (2002)
10. Evans, C.: Notes on the OpenSURF library. University of Bristol, Technical report CSTR-09-001 (2009)

Optimizing Congestion Control for Non Safety Messages in VANETs Using Taguchi Method

Mohamad Yusof Darus(✉), Mohd Salehuddin Zainal Abidin, Shamsul Jamel Elias, and Zarina Zainol

Faculty of Computer and Mathematical Sciences, University Technology of MARA, 40450 Shah Alam, Selangor, Malaysia
{yusof,zarina}@tmsk.uitm.edu.my, mrtelecaster.zas@gmail.com, shamsulje@kedah.uitm.edu.my

Abstract. VANETs are the next uprising technology in automotive industry to promote safety and valuable information to make a better driving experience. This technology allows vehicle to communicate directly to the next vehicle and exchange data. The main purpose of VANETs is to improve traffic safety and to provide efficient traffic management on road via safety and non-safety message. Safety and non-safety messages are disseminated instantly among the vehicles through broadcasting protocol. By broadcasting, message can be delivered to multiple vehicles at the same time within vicinity. When too many vehicles broadcasting in dense traffic, VANETs suffer network congestion due to excessive amount of broadcast messages consume the bandwidth of communication channel. When broadcast storm occurs, message could not be delivered properly due to packet loss in the transmission. To address this issue, congestion control mechanism was proposed to alleviate the congestion in the communication channel during broadcast storm. In this research, the congestion control is applied to Service Channel communication (SCH) for non-safety message. The proposed congestion control was also tested using Taguchi method to optimize packet loss reduction in the network. The experiment was conducted with and without Taguchi method in urban area. The obtained results from the experiment show packet loss was greatly reduced when applying congestion control with Taguchi method. The results have proven that integrating Taguchi method in congestion control mechanism could improve packet loss reduction when broadcast storm take place in high-density traffic.

Keywords: VANETs · Broadcast storm · Taguchi method · Non-safety messages Congestion control

1 Introduction

Vehicle Ad-Hoc Networks (VANETs) has become an important research area to improve traffic awareness by using wireless connection. VANETs communicate wirelessly via embedded On Board Unit (OBUs) wireless device. The OBUs act as an interface which allows vehicle to form a short range of wireless ad-hoc networks with the capability of broadcasting data and application to vehicular networks or transportation

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 74–87, 2018.
https://doi.org/10.1007/978-981-10-8276-4_8

authorities. The OBUs possess indefinite power transmission due to onboard power supply which makes it different from any other form of ad hoc network like MANET. When vehicular network is formed between nodes equipped with OBU, it is called vehicle-to-vehicle (V2V) communication, whereas when nodes communicates with the road-side-unit (RSU) and form the network, it is known as vehicle-to-infrastructure (V2I) communication.

In 1999, The Federal Communication Commission (FCC) has allocated 75 MHz in the 5.9 GHz band for Dedicated Short Range Communication (DSRC) as a standard for VANETs wireless band to communicate among the vehicles. In VANETs, there are two types of messages transmitted by DSRC, namely safety message and non-safety message. DSRC provides one Control Channel (CCH) for safety messages and six Service Channels (SCHs) for non-safety messages with 5 MHz of guard band respectively. Safety message in VANETs is critical information generated by safety application to predict and to prevent accident on road. Examples of safety applications are cooperative collision avoidance and emergency brake detection. Non-safety message on the other hand is a non-critical information message generated by non-safety application to comfort driver and passengers in the vehicle.

In VANETs, broadcast routing disseminates the safety and non-safety messages. Among all routing protocol, broadcast based routing is mostly used to disseminate data due to fast propagation towards all vehicles in vicinity without constructing dedicated path to reach the next vehicle [1]. Broadcast routing allows safety and non-safety to be disseminated to the vehicle beyond the transmissions range. Every vehicle received the message will re-broadcast the message to the next hop. This flooding technique allows message to be transmitted to the furthest vehicle in the area until all vehicles received the same information. Even though flooding is the simplest technique to be used and works fine in spare network but in a dense network, it produces collision, contention and redundant information [2]. Frequent contention and an excessive numbers of broadcasted packets will lead to collision, thus causing broadcast storm [3]. This significantly contributes to packet loss in the network [4].

Network congestion and packet loss will severely impact the performance of service applications in VANETs [5]. If the service application is related to safety message such as collision detection, the warning message sent by the source node may not successfully delivered to the neighboring nodes due to packet loss and it might increase the risk of accident. For non-safety message like routing discovery or traffic forecast, packet loss may reduce the accuracy of the information and user might be provided with false information too. Congestion problem can be alleviated with congestion control mechanism. Appropriate congestion control mechanism is essential to ensure an efficient operation of a network. Though a lot of research has been done on congestion control mechanism yet its focus was mostly on safety-messages. There is ample scope to deal with congestion control mechanism for non-safety messages since various reactions from drivers will generate multiple non-safety messages on mobile network environments. Two categories of congestion control mechanism are recognized [6], which include reactive and proactive congestion control.

In this research, three congestion control approaches were compared to be adopted in SCH for non-safety message. Although mostly congestion control are applied to CCH, it

also applicable to control congestion in SCH too. These three congestion control mechanisms are Cooperative scheme for service channel reservation (CRaSCH) [6], and Vehicular enhanced Multi-Channel MAC (VEMMAC) [7]. The following Tables 1 and 2 summarize the capabilities of the above-mentioned congestion control approaches with their schemes.

Table 1. Summary of VANETs congestion control

Approach	Packet rate	Utility function	Access priority	Carrier sense	Smart broadcast
CrasSCH	NO	YES	NO	YES	NO
MBDS	YES	YES	NO	YES	YES
VEMMAC	YES	YES	NO	NO	YES

Table 2. Summary of congestion control scheme solution area

Approach	Throughput	Delay	Packet loss
CrasSCH	YES	NO	NO
MBDS	YES	YES	NO
VEMMAC	YES	YES	NO

The summary clearly indicates that MBSD has more advantages for congestion control compared over VEMMAC and CRaSCH. The scheme will be adopted to control congestion in SCH especially in high density traffic.

2 Optimization

Any existing systems, programming, algorithm to solve or to serves specific function can be optimized to obtain the best output by modifying the constraint into the given process. A routing can be optimized by identifying the fastest route to reach destination. Network optimization is one way to maintain the existing protocols, mechanism or other network parameter rather than costly efforts to design and implementing new improved mechanism. Response Surface Methodology (RSM) and Taguchi Method are two most known system optimizations in manufacturing and engineering field. Both methodologies aim to achieve the best output and improve the manufacturing and engineering productivity. Taguchi provides very precise optimization but to follow the process is difficult as it requires great deal of accuracy in manufacturing. RSM on the other hand is easy to apply but the method is outdated and the results are not excellent in terms of quality. The design of the mobile ad-hoc network (MANET) with the implementation of Taguchi parameter design has been discussed [8]. The Taguchi method was chosen due to its ability to determine and choose the optimal combination of parameters. In addition, by using the loss function, the variance of the intended output can be reduced. This in turn reduced the loss and attenuation, and improved the connectivity, routing capability and bandwidth utilization.

3 Experiment Phase

3.1 Urban Map Set-up

The map in Fig. 1 was designed in a square shape with nine co-ordinate point for vehicles to move from starting point until the end of the map. Using a simple square map could simulate an urban scenario where roads are divided into few sections with multiple junction, thus could help to simulate broadcast storm when vehicle traffic is high. The RSUs also have been placed in the map at four areas. Vehicles also will exchanges broadcast message with the RSU as it moves along the road. The designed map was based on the parameter in Table 3.

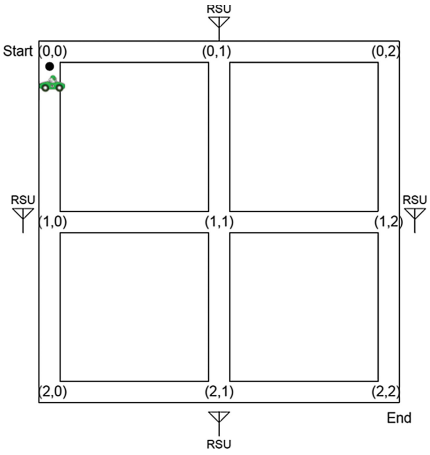


Fig. 1. Simple map of urban scenario

Table 3. Map parameters

Parameter	Value
Simulation area	1000 m × 1000 m/ 2000 m × 80 m
Number of vehicles	10–60
Number of RSU	4
Packet type	UDP
Network interface	IEEE 802.11
Vehicle speed	40–50 kph
Transmission range	250 m intervals
Simulation time	250 s
Data packet size	512–1024
CBR rate	0.5–5 packet per second
MAC protocol	IEEE 802.11p
Vehicle distance	4 m–5 m

Based on the parameters above, vehicles movement will be set based on assigned coordinate. There are four routes established in this experiment to simulate the vehicle movement in urban scenario which are:

Route 1: Co-ordinate (0, 0), (0, 1), (0, 2), (1, 2), (2, 2)
Route 2: Co-ordinate (0, 0), (0, 1), (1, 1), (2, 1), (2, 2)
Route 3: Co-ordinate (0, 0), (1, 0), (2, 0), (2, 1), (2, 2)
Route 4: Co-ordinate (0, 0), (1, 0), (1, 1), (1, 2), (2, 2)

3.2 Taguchi Method

Technically, Taguchi method consists of three phases, Planning Phase, Experiment Phase and Analysis Phase.

3.2.1 Planning Phase

Control and noise factors were to be selected in order to optimize congestion control for non-safety message in dense network to achieve packet loss reduction. Control factors are controllable parameters in designing a congestion control while the noise factors are factors that may influence the selected quality characteristic and cannot be adjusted in the design as shown in Fig. 2.

M Signal factors: non-safety data packet
X Noise factors: send interval, packet size, simulation time, bit rate, number of vehicle
Z Control factors: wifiPreambleMode, slotTime, rtsThresholdBytes, minSuccessThreshold, successCoef
Y Response/Output: Packet received

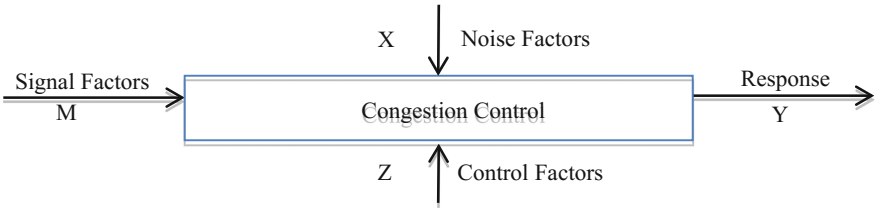


Fig. 2. Parameter diagram by Taguchi

Basically, in congestion control mechanism the dissemination packet to neighboring nodes is important especially during high traffic. When network becomes congested, the service application is not reliable anymore when the data failed to reach the destination due to packet loss. Therefore, for optimal performance, the smaller-the-better performance metric for packet loss reduction must be applied in non-safety message.

3.2.2 Experiment Phase

The experiments were conducted based on combination of control factors and noise factors. All protocols such routing and congestion control were based on INETMANET

modules implantation. The experiment was conducted based on L8 Taguchi orthogonal array design as shown in Table 4. Figure 3 shows a movement of vehicles during experiment using OMNET++ simulator.

Table 4. Orthogonal array L8 control parameter

Experiment	wifiPreambleMode	slotTime	rtsThresBytes	minSuccessThres	SuccessCoef
E1	L	L	L	L	L
E2	L	L	L	H	H
E3	L	H	H	L	L
E4	L	H	H	H	H
E5	H	L	H	L	H
E6	H	L	H	H	L
E7	H	H	L	L	H
E8	H	H	L	H	L

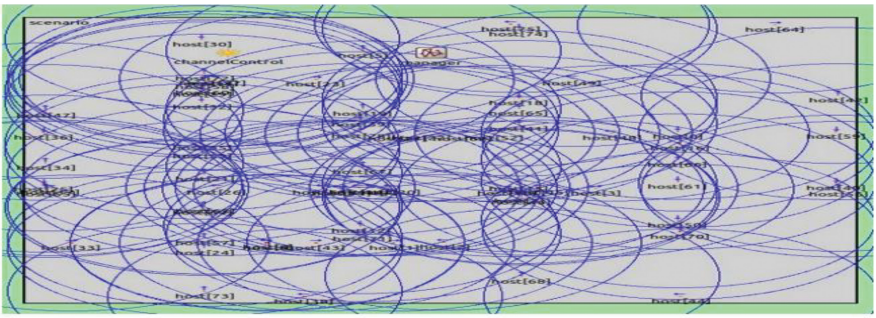


Fig. 3. Snapshot of vehicles movement during the experiment

3.2.3 Experiment Phase

The aim of this research is to optimize congestion control for non-safety message by using Taguchi method. The experiment conducted to improve packet delivery by minimizing packet loss during data exchange in high density traffic. After the experiments have been carried out, and results are obtained to experiment design, the loss function for each target will be calculated. The loss function to be used for this experiment will be the smaller-the-better. The performance metrics to measure the output from the experiment would be as follow:

$$Packet\ loss\ =\ Packet\ sent - Packet\ received$$

Each parameters experiment results will evaluate how many packet loss reductions for each variable. The output then will be recorded and will be analyzed.

4 Experimental Analysis

The proposed congestion control mechanism was tested with and without Taguchi parameter. The simulation was done to prove if Taguchi method could optimize the congestion control by minimizing packet loss during the transmission.

In Taguchi method, if one of the SCH communication channels is congested, it will launch the optimized congestion control algorithm immediately. During the simulation, the noise parameters were changed accordingly.

When simulation started, congestion control will start to monitor SCH channel. Vehicles move based on co-ordinate assigned and started to broadcast non-safety message to the RSUs which have been assigned in four places in the simulation map. Then the flow proceeds to control parameter assignment. When congestion occurs in SCH, congestion control is applied with all the parameters assigned. The output broadcast packet received by RSU will be measured.

5 Result and Findings

The following section summarizes the results for noise parameters that have been conducted.

5.1 Bit Rate Parameter

Result of packet loss reduction based on bit rate parameter is tabulated in Table 5.

Table 5. Packet loss reduction based on bit rate

Bit rate	3	6	9	12
Original	94335	90909	99814	109452
Optimized	91384	85888	94591	106810
Different	2951	5021	5223	2641

Figure 4 shows that the optimized technique parameter value is lower than the original technique parameter value as per Taguchi method for packet loss (the lower the better). The red line in the graph represents the performance of the optimized congestion control for non-safety packet while the dotted blue-line represent the original technique parameter.

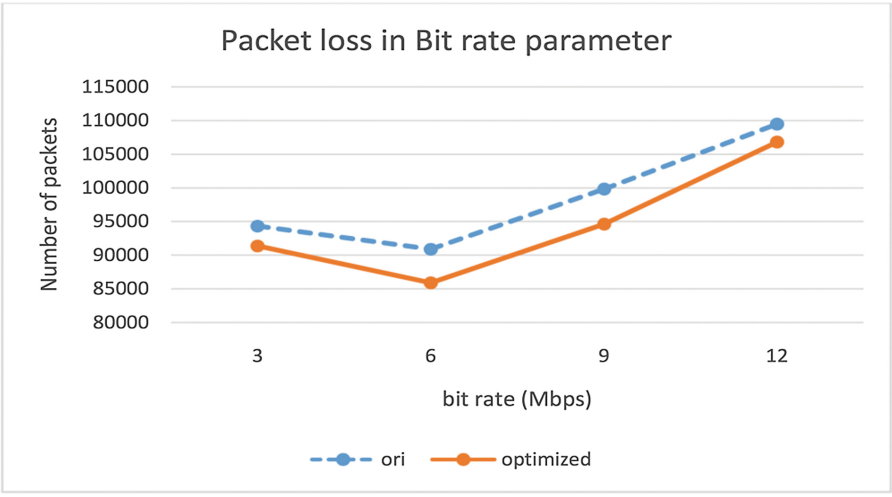


Fig. 4. Packet loss in bit rate parameter

5.2 Number of Vehicles Parameter

Result of packet loss reduction based on number of vehicle is tabulated in Table 6:

Table 6. Packet loss reduction based on number of vehicle

Num. of vehicle	10	20	30	40	50
Original	3611	6295	14333	23508	36883
Optimized	4541	5153	10125	19288	35526
Different	–930	1142	4208	4220	1357

Figure 5 shows the differences between the proposed optimized techniques with the original congestion control based on number of packet value 10, 20, 30, 40 and 50. Figure 5 shows that the optimized technique parameter value is lower than the original technique parameter value as per Taguchi method for packet loss (the lower the better). The packet loss reduction was optimum when vehicles at 40 and could be the ideal number of vehicle to implement Taguchi method in congestion control when broadcast storm occurs.

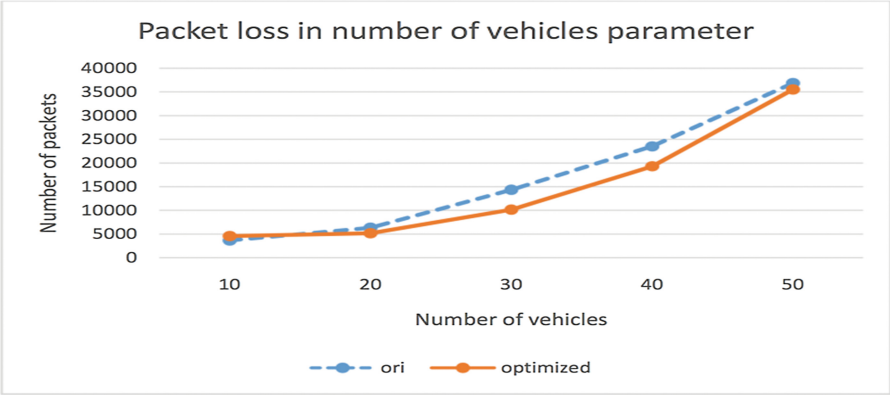


Fig. 5. Packet loss in number of vehicles

5.3 Packet Size Parameter

Result of packet loss reduction based on packet size is tabulated in Table 7:

Table 7. Packet loss reduction based on packet size

Packet size	512	640	768	896	1024
Original	92471	99234	104102	106935	110372
Optimized	92369	99056	103061	106104	109102
Different	102	178	1041	831	1270

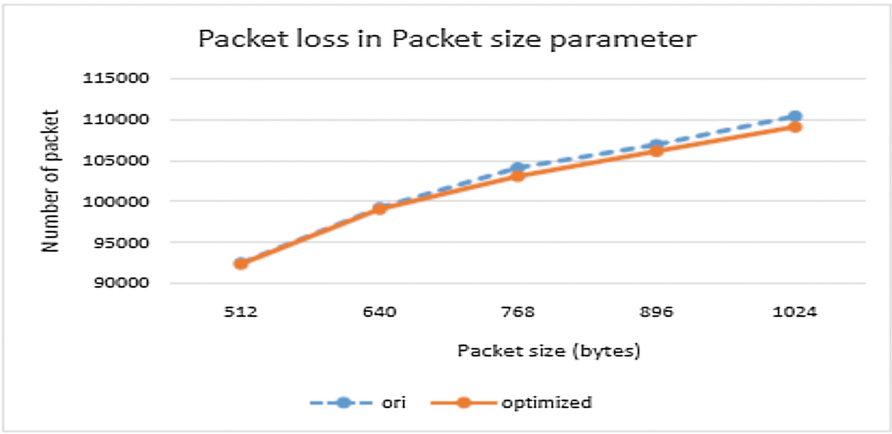


Fig. 6. Packet loss in packet size parameter

From the results tabulated in Table 7 and Fig. 6 shows that the highest packet loss reduction was captured when parameter 1024 bytes of packet size was assigned into transmission protocol. The lowest reduction was captured when 512 bytes packet size

was assigned. The result has proven when larger packet size was assigned, lesser time taken to complete the packet transmission and thus decrease the chance of the packet loss. While with smaller packet size, the time taken to complete the round trip message was longer and increases the chance of losing the packet during transmission.

5.4 Send Interval Parameter

Result of packet loss reduction based on send interval is tabulated in Table 8. Figure 7 shows the differences between the proposed optimized techniques with the original congestion control based on send interval value 0.5, 0.02, 0.01, 0.007 and 0.005. The purpose of this experiment is to determine the impacts of the optimized technique in different send interval time. The red line in the graph represents the performance of the optimize congestion control for non-safety packet while the dotted blue-line represent the original technique parameter. Figure 7 shows that the optimize technique parameter value is slightly lower than the original technique parameter value as per Taguchi method for packet loss (the lower the better).

Table 8. Packet loss reduction based on send interval (ms)

Send interval	0.5	0.02	0.01	0.007	0.005
Original	19551	618296	1239933	1775772	2492140
Optimized	15452	607743	1232502	1768069	2484240
Different	4099	10552	7430	7703	7900

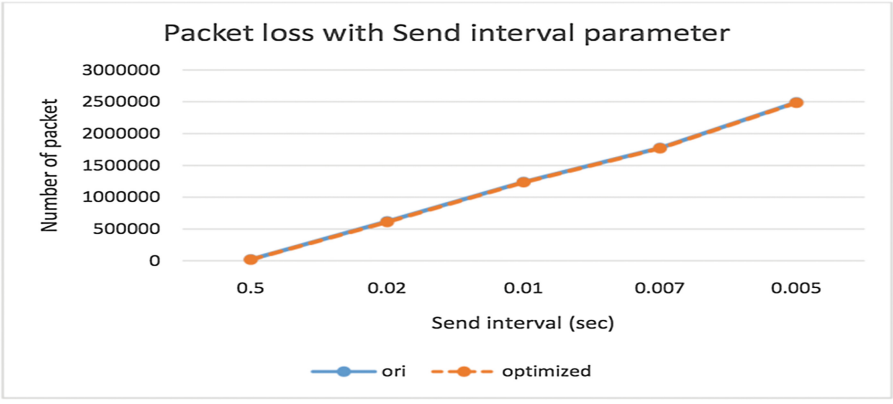


Fig. 7. Packet loss with send interval parameter

5.5 Simulation Time Parameter

Result of packet loss reduction based on simulation time is tabulated in Table 9. In this experiment, each vehicle started to leave the starting point with the given simulation time after the first vehicle which were 50 s, 100 s, 150 s, 200 s and 250 s. This will create a gap between two vehicles when approaching RSU. The results tabulated in Table 8

shows as simulation time getting wider, the number of packet loss significantly decreased. When the gap was close as 50 s gap, total number of packet received by RSU was high and when the gap was wider as 250 s, total number of packet received by RSU was low. The simulation time applied gave time to SCH channel slot in RSU to be released and received incoming broadcast message from the next vehicle. The highest packet loss reduction was captured when simulation time assigned was 100 s and the lowest packet loss reduction was at 50 s. However in broadcast storm vehicles are moving closer and the gap between vehicles might be less than one seconds.

Table 9. Packet loss reduction based on simulation time

Simulation time (sec)	50	100	150	200	250
Original	95495	47261	30798	23314	19438
Optimized	93343	37174	24751	18866	15670
Different	2151	10087	6047	4448	3767

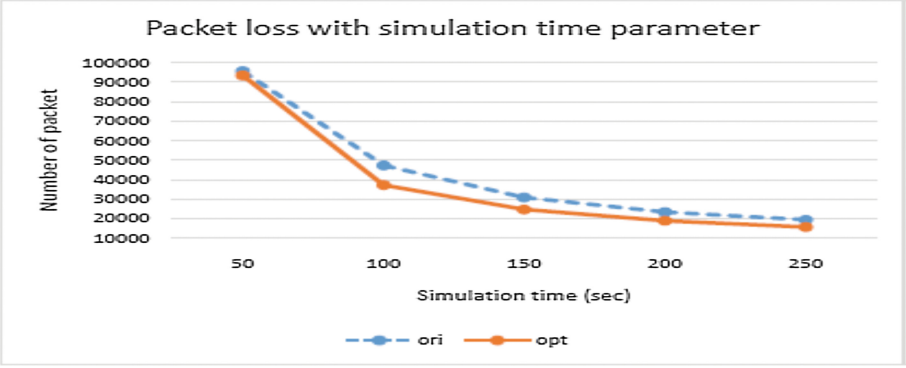


Fig. 8. Packet loss with simulation time parameter

Figure 8 shows the differences between the proposed optimized techniques with the original congestion control executed on different simulation time on value 50 s, 100 s, 150 s, 200 s and 250 s. The results tabulated in Table 9 shows as simulation time getting wider, the number of packet loss significantly decreased. When the gap was close as 50 s gap, total number of packet received by RSU was high and when the gap was wider as 250 s, total number of packet received by RSU was low. The simulation time applied gave time to SCH channel slot in RSU to be released and received incoming broadcast message from the next vehicle. The highest packet loss reduction was captured when simulation time assigned was 100 s and the lowest packet loss reduction was at 50 s. However, in broadcast storm vehicles are moving closer and the gap between vehicles might be less than one seconds.

6 Performance Comparison of All Tested Parameters

There are five parameters tested in each experiment to optimize the congestion control mechanism. From the results tabulated in Fig. 9 and Table 10, the output of packet loss reduction was categorized into two categories (a) the highest packet loss reduction and (b) the lowest packet reduction.

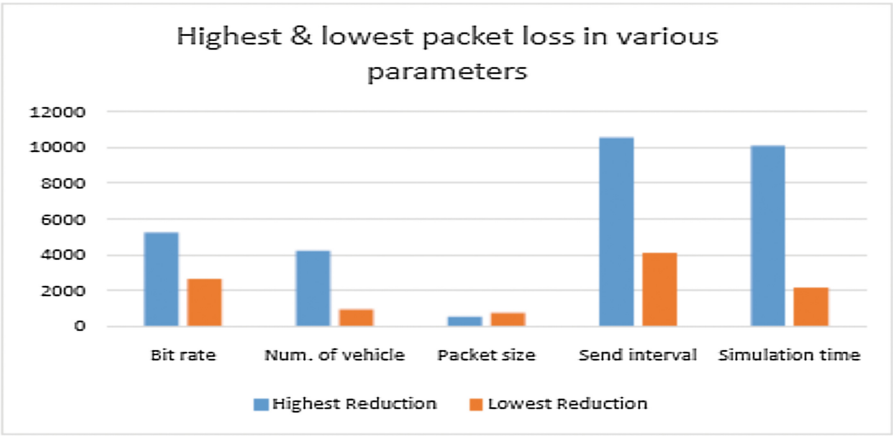


Fig. 9. Highest and lowest packet loss reduction

Table 10. Highest and lowest packet loss reduction

Parameters	Highest reduction	Lowest reduction	Different (%)
Bit rate	5223	2641	97
Num. of vehicle	4220	929	354
Packet size	1269	101	1156
Send interval	10552	4099	157
Simulation time	10087	2151	368

The highest reduction of packet loss was obtained by selecting the highest packet loss output among the parameter attributes. The lowest packet loss reduction on the other hand was obtained by selecting the lowest packet loss output among the parameter attributes. The highest and lowest results were compared to find the difference and the results were shown in the Table 11. There were huge gap between highest and lowest packet loss reduction in send interval and in simulation time.

The experiment results shows that after Taguchi parameters applied to the congestion control, the number of packet loss has decrease significantly. All noise parameter applied shown an improvement in reducing the packet loss when congestion control is triggered in broadcast storm. From the experiment, number of vehicle parameter has the highest percentage of improvement in reducing packet loss with the congestion control

compared to other tested factors while packet size variable has the lowest percentage of packet loss reduction.

Table 11. Average percentage of packet loss reduction

Parameters	1	2	3	4	5	Average (%)
Bit rate	3.13	5.52	5.23	2.41	0	4.07
Num. of vehicle	25.74	18.13	29.36	17.95	3.68	18.97
Packet size	0.32	0.18	0.73	0.50	0.30	0.41
Send interval	20.97	1.71	0.60	0.43	0.32	4.80
Simulation time	2.25	21.54	19.63	19.08	19.38	16.34

7 Conclusion

In a dense network especially in urban area, the numbers of vehicles are higher than usual. When there are too many vehicles in a certain area, the performance of services and application in VANETs are significantly degraded due too many data collision and congestion which also known as broadcast storm. The experiment is conducted to find out if Taguchi method could improve the congestion control by reducing packet loss in broadcasting non-safety message in dense network. With Taguchi method implemented during congestion control, number of packet loss has been greatly reduced. The experiment shows when the adopted congestion control runs with number of vehicle parameter it has the most packet loss reduced among the other variables. As the number of vehicle increase from 50 to 250 vehicles, the optimize congestion control has reduced more of packet loss. This result is significantly prove that when broadcast storm occurs when vehicles in traffic are increasing, Taguchi methods managed to optimize the congestion control and reduce the packet loss.

Acknowledgment. The authors would like to thank the Universiti Teknologi Mara (UiTM) for sponsoring this research under the ARAS Grant (600-IRMI/DANA 5/3/ARAS (0177/2016)) of Research Management Centre, Universiti Teknologi Mara (UiTM).

References

1. Chelha, I.O., Rakrak, S.: Best nodes approach for alert message dissemination in VANET (BNAMDV). In: Proceeding Third International Workshop on RFID and Adaptive Wireless Sensor Networks, pp. 82–85 (2015)
2. Zemouri, S., Djahel, S., Murphy, J.: A fast, reliable and lightweight distributed dissemination protocol for safety messages in Urban Vehicular Networks. J. Ad Hoc Netw. **27**, 26–43 (2015)
3. Kolte, S.R., Madankar, M.S.: Adaptive congestion control for transmission of safety messages in VANET. In: Proceeding International Conference for Convergence of Technology (I2CT 2014), pp. 1–5 (2014)
4. Patel, A., Gharge, A., Trivedi, P., Potdar, M.: Congestion control scheme for Vehicular Ad-Hoc Network (VANET). Int. J. Comput. Appl. **4**(1), 2014–2017 (2016)

5. Campolo, C., Cortese, A., Molinaro, A.: CRaSCH: a cooperative scheme for service channel reservation in 802.11p/WAVE vehicular ad hoc networks. In: Proceeding International Conference on Ultra Modern Telecommunications and Workshops (2009)
6. Hong, C.S.: An enhanced multi-channel MAC for vehicular ad hoc networks. In: Proceeding Wireless Communications and Networking Conference (WCNC 2013), pp. 351–355 (2013). http://doi.org/10.1109/WCNC.2013.6554589
7. Elshaikh, M.: Energy consumption optimization with Ichi Taguchi method for wireless sensor networks. In: Proceeding 2nd International Conference on Electronic Design (ICED 2014), pp. 493–498 (2014)
8. Mohamed, H., Lee, M.L., Sarahintu, M., Salleh, S., Sanugi, B.: The use of Taguchi method to determine factors affecting the performance of destination sequence distance vector routing protocol in mobile ad hoc networks. J. Math. Stat. **4**(4), 194–198 (2008)

An Authentication Technique: Behavioral Data Profiling on Smart Phones

Salmah Mousbah Zeed Mohammed(✉), Azizul Rahman Mohd Shariff, and Manmeet Mahinderjit Singh

School of Computer Sciences, Universiti Sains Malaysia (USM), George Town, Pulau Pinang, Malaysia
salmah.mousbah@gmail.com, {azizulrahman,manmeet}@usm.my

Abstract. Mobile devices have become an indispensable component in modern society. Many of these devices rely on personal identification numbers (PIN) as a form of user authentication. One of the main concerns in the use of mobile devices is the possibility of a breach in security and privacy if the device is seized by an outside party. Threats can possibly come from friends as well as strangers. Smart devices can be easily lost due to their small size, thereby exposing details of users' private lives. User behavior authentication is designed to overcome this problem by utilizing user behavioral techniques to continuously assess user identity. This study proposed a behavioral data profiling technique that utilizes data collected from the user behavior application to verify the identity of the user in a continuous manner. By utilizing a combination of analytical hierarchy process and correlation coefficient method, the best experimental results were obtained by verifying the identity of six types of user behaviors to determine the different behaviors. Based on the results, this study proposes a new authentication technique that enables verification of a user's identity through their application usage in a transparent manner. Behavioral data profiling is designed in a modular manner that will not reject user access based on a single application activity but on several consecutive abnormal application usages to balance the trade-off between security and usability. The proposed framework is evaluated using a PIN-based technique and achieved an overall 95% confidence level. Behavioral data profiling provides a significant improvement in the security afforded to the device and user convenience.

Keywords: Security of smartphone · User behavior · Decision-making
Correlation coefficient

1 Introduction

The rise in the usage of smartphones over the past few years has been a story of technological triumph. The latest expansion in mobile technologies has produced a new kind of device, a programmable mobile phone known as the smartphone. Generally, smartphone users can program any application tailored for their needs. Furthermore, smartphone users can share these applications in the online market. Therefore, smartphone and its applications are currently the most prevalent key-words in mobile technology [1].

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 88–98, 2018.
https://doi.org/10.1007/978-981-10-8276-4_9

However, a smartphone needs confidential information to provide these customized services, thus causing security weaknesses. All smartphones are preferred targets of attacks. Authentication, which can only be maintained by proper identification of the end users, is the primary step to safeguard the integrity and confidentiality of an infrastructure. Authentication and authorization controls help protect unapproved access to mobile devices and their stored data. Smartphone security [2] authentication is vital for our assets, which include individual data, corporate intellectual property, classified information, financial assets, device and service availability and functionality, and personal and political reputation. Moreover, authentication helps prevent data loss in the case of mobile device theft or damage. Numerous authentication techniques through which we can enhance security to prevent intruder breach have been proposed.

A smartphone is a vital source of information. However, the availability of this information has initiated a growth in cyberattacks. The cybersecurity risks to unauthorized data access is principally the same for smartphones [3] as well as tablets, laptops, or any other mobile device operating outside of an organization's physical office. Owing to an increasing number of people who use their smartphones to run their entire lives, hackers and other intruders will center their efforts on obtaining the information they want from these devices. Regrettably, such intention poses considerable challenges in terms of security for organizations with employees who use such devices in their daily work. Data security is the main concern not only for enterprises and small business [4] but also for everyday users. Extensive data breaches reveal everything, from customer login credentials to credit card information and personal health records. Smartphones store information on your calls, your location, what you have search on the Internet, and passwords to social networks. Thus, grave consequences are expected if your phone ends up in the wrong hands [5].

In this paper, we introduce a new concept to secure the mobile device based on the user's behavior and models. It utilizes a user's behavior pattern to improve existing services and use the models to detect significant variations in the user's behavior and anomaly detection. This is important as a mobile device (smartphone) is a near depiction of the users' themselves. As mobile device gets a very personalize in the further, the mobile device is actually a personal representation extension and embodiment of the user's behavior in the connected world. The user and the mobile device can be said to have a symbiotic relationship. This paper is organized as follows. Section 1 Authentication Techniques. Then, a novel framework is described for user behavior on mobile devices in Sect. 2. Section 3 outlines the evaluation and finally, Sect. 4 concludes this paper with some recommendation of future works.

2 A New Behavioral Data Profiling Framework Based on the User Behavior

Authentication is the process used to validate the true user of a system. In the context of security, authentication considers three primary implementation strategies. In this study, we propose the fourth strategy regarding authentication, as shown in Fig. 1.

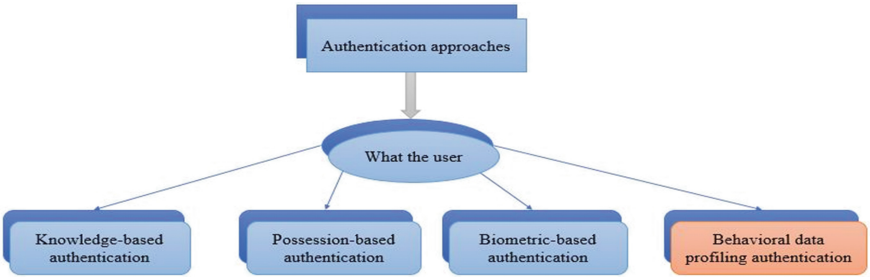


Fig. 1. Authentication strategies for interaction between users and smart devices

Authentication can be active or passive. Active authentication requires dealing with a device and inputting one or more pieces of valid information or answers to questions. If required by individual applications, then this kind of authentication can become tedious and frustrating because most individuals use many applications or services. Consequently, many individuals prefer to use their applications with few security impediments each time they decide to access their devices [6]. Current smartphones are based on entry-point authentication, which can be either a personal identification number (PIN) or a secret pattern. A user is usually required to pick four digits for validation to use the PIN. The user must correctly input this code; otherwise, he/she cannot pass the entry point to his/her device. Another current authentication method is the use of a secret gesture. A secret gesture is defined by moving a finger over the screen to create a certain pattern. This pattern can be used as an authentication to grant the user entry into the device.

The user behavior framework provides an enhanced security to provide adequate protection for the mobile device. The user behavior framework operates in numerous process engines, and a security manager is devised to achieve these objectives (as illustrated in Fig. 2).

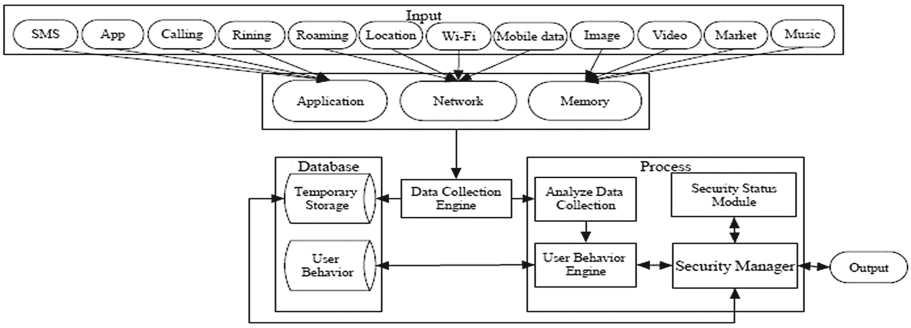


Fig. 2. A novel user behavior framework

As previously mentioned, the user behavior framework verifies the identity of a user based on their application activities. The entire verification procedure is implemented through the cooperation of the process engines and the security manager. First, the data

collection engine gathers and transforms the application behavior of a user. For example, transforming the utilized storage on a smartphone into various behaviors input samples. Then, the user behavior engine performs the verification process by comparing the input sample(s) with the suitable behaviors, which are generated by the analyzed data collection engine. Once the verification process is completed, the verification results will be appropriately processed by the security manager according to the mode in which the framework operates. The security manager handles the verification results by itself and carries out any necessary corresponding responses, such as associating appropriate labels for the verified input data (whether legitimate or illegitimate) and updating the SS level. The security manager simply makes any final decisions accordingly. A detailed description of this process is thoroughly explained throughout the subsequent sections.

2.1 Data Collection Engine

The main function of the data collection engine is to capture the behavior of the user based on application usage, utilized network, and storage on a smartphone. When an application is utilized by a user, the data collection engine (as illustrated in Fig. 3) automatically gathers the information associated with that application in the background of a mobile device OS.

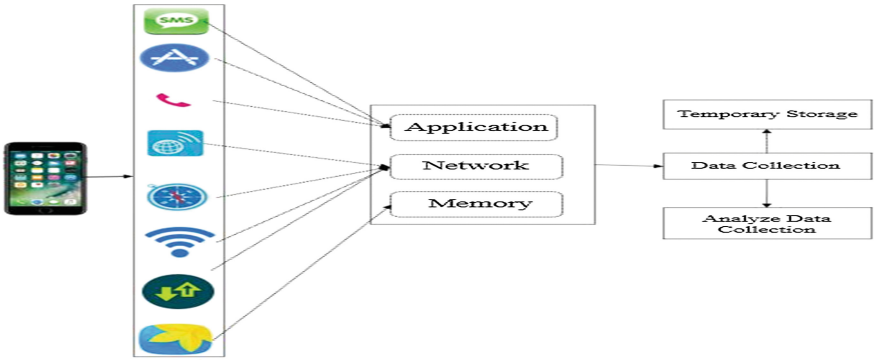


Fig. 3. Data collection engine

Once all the necessary applications, networks, and memory features are extracted, the data collection engine then proceeds to the subsequent phase, which is, preprocessing these applications, networks, and memory into behavior samples. The data collection engine sends the behavior samples to the temporary storage for further processing. The actual size of the temporary storage varies depending on the way the analyze data collection engine performs the verification process (as discussed in Sect. 2.2). Once the verification process is performed, the data are sent to the user behavior engine, and the data stored in the temporary storage will be removed accordingly. Nonetheless, the temporary storage should have the same structure for each mobile device regardless of hardware configurations. The temporary storage table contains several data fields, such as date, time, and name of the applications, networks, and memory. When verification

is required, the temporary storage forwards all input data to the security manager. In this way, a significant amount of data can be stored/saved. Moreover, the process speed will be improved when the data are required by any subsequent usages.

2.2 Analyze Data Collection

The primary function of the analyze data collection engine is to generate various user behavior templates, as illustrated in Fig. 4. This function is achieved by utilizing analytical hierarchy processes (AHPs) for the data entered from the data collection engine. The hierarchy is important because it attempts to realistically represent user behavior by providing the most used apps, networks, and storage on their smartphone and their relationships [7]. A hierarchy is an abstraction from the inputs to analyze the user behavior as well as the characteristics and patterns of the total system [8]. These hierarchies represent the basic human way of behaving in terms of separating reality into sets and subsets. Moreover, once the hierarchy has been established, the behavior weights of each usage on the phone can be determined to rate the decision-making alternatives based on user characteristic. Meanwhile, the analyses of the hierarchy are based on three categories, namely application, network, and memory. The details of application usage include SMS, apps, calling, and ringing. All criteria and alternatives in the first category are illustrated in Fig. 5. The networks utilized by a user involve roaming, location, Wi-Fi, and mobile data, as shown in Fig. 6. Therefore, the storage space in the mobile device is based on image, video, market, and music, as clearly depicted in Fig. 7.

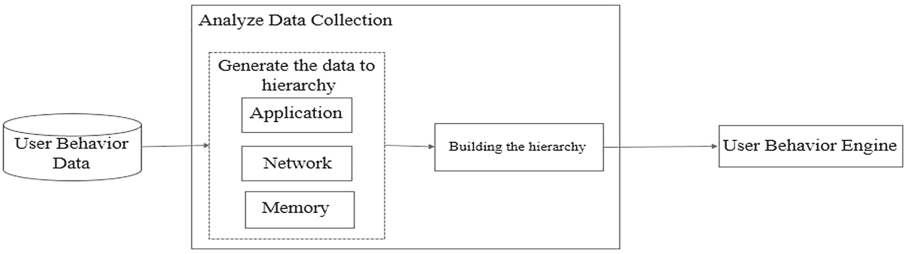


Fig. 4. Analyze data collection

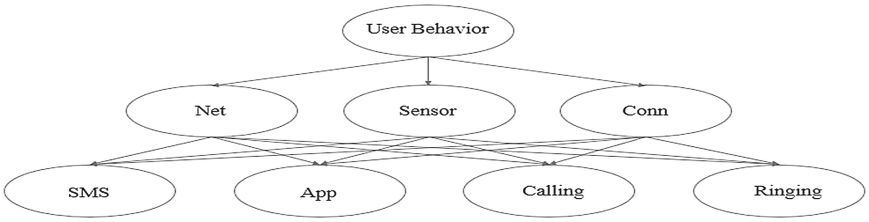


Fig. 5. Analyze data collection for application

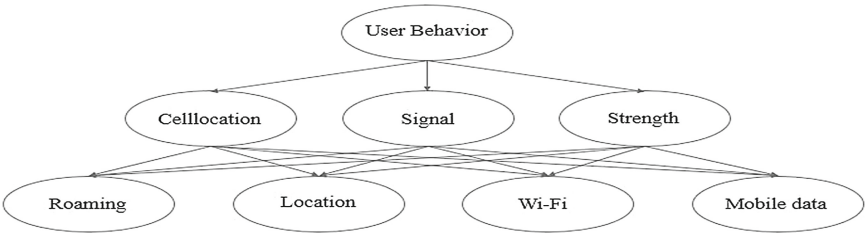


Fig. 6. Analyze data collection for network

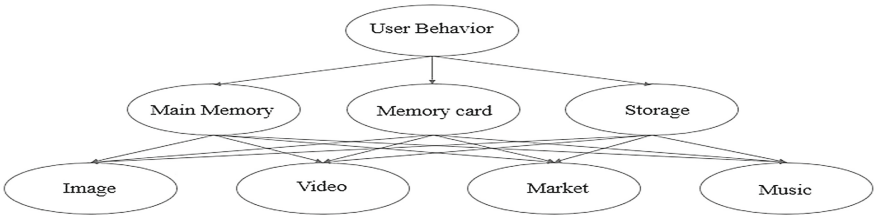


Fig. 7. Analyze data collection for memory

2.3 Use Behavior Engine

The User Behavior Engine provides the main functionality for the verification process. Figure 8 shows that when a verification requirement is met, the user behavior engine calculates the ordering value for all application activity input (obtained from the User Behavioral Analyze Data Collection) by utilizing their behavior. The value of the ordering for all categories will be compared with the predefined average usage for each category: within the average usage, the activity will be assumed as legitimate; if exceeding the average usage, the activity will be classified as illegitimate. Then, the User Behavior Engine checks the correlation and sends the verification result to the Security Manager which will make a response based upon the operation modes of the framework.

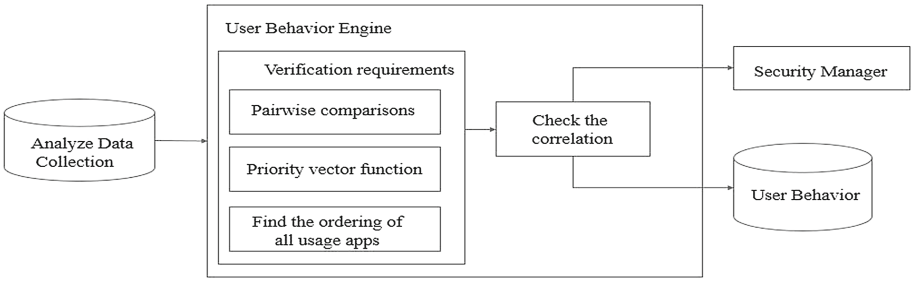


Fig. 8. User behavior engine

2.4 Security Manager

The security manager is the brain of the framework because it controls all other elements. The security manager has a variety of roles based on the working role of the framework. These roles are fully discussed in the following section. The key task of the security manager is to monitor the current SS level and make subsequent decisions according to the user behavior. The process algorithm is the core security component of the proposed framework. This algorithm contains three main verification stages, which require the correlation between user behavior and provider to enable further use of the device before it is locked down. These verification stages were selected to provide an elevated level of user convenience and improved security. The process algorithm also employs the AHP method to verify the identity of a user. Most legitimate users will experience transparent phases. Meanwhile, the order verification challenges of user behavior are utilized to ensure the legitimacy of a user in the event of access requirement to the mobile device. However, the SS level is below the set security requirements.

3 Evaluation

Two aspects of the framework, namely the influence on the processing power and the effective authentication of the user, should be examined to understand the effect of the framework on the overall performance. Regarding consumption of the processing power, previous research showed that a complicated multimodal biometric authentication system (i.e., TAS) was prototyped within the mobile environment, and users were satisfied with the performance [9]. The framework for the authentication performance was evaluated within the JavaScript environment and the process algorithm was employed. The behaviors of each user were divided into three scenarios containing the application usage, network, and memory space storage.

As discussed previously, the performance of the framework can be influenced by the following three key parameters: verification of user behavior, correlation among user behaviors, and SS level. Therefore, the evaluation sought to analyze the effect of these parameters on the performance. As such, three scenarios were set up to verify the user behavior, as illustrated in Tables 1, 2, and 3.

Table 1. Verify user behavior the most usage for the first category

Category 1	User1	User2	User3	User4	User5	User6
SMS	0.089	0.147	0.085	0.088	0.236	0.088
App	0.471	0.421	0.469	0.534	0.458	0.231
Calling	0.184	0.213	0.163	0.083	0.0181	0.186
Ringing	0.256	0.219	0.283	0.295	0.125	0.0494

Table 2. Verify user behavior the most usage for the second category

Category 2	User1	User2	User3	User4	User5	User6
Roaming	0.446	0.249	0.159	0.299	0.457	0.273
Location	0.055	0.08	0.126	0.282	0.199	0.428
Wi-Fi	0.395	0.587	0.524	0.097	0.119	0.104
Mobile data	0.091	0.083	0.19	0.321	0.224	0.194

Table 3. Verify user behavior the most usage for the third category

Category 3	User1	User2	User3	User4	User5	User6
Image	0.214	0.471	0.246	0.152	0.175	0.294
Video	0.183	0.179	0.259	0.089	0.291	0.464
Market	0.173	0.277	0.285	0.325	0.094	0.127
Music	0.429	0.073	0.211	0.435	0.411	0.115

3.1 Experiment Results

The verification requirement of the user behavior will be verified as soon as one application is utilized. Therefore, the results used to verify the behaviors of users based on the AHP for the three categories are presented. The most usage for the first, second, and third category is presented in Tables 1, 2, and 3, respectively.

Compared with the AHP results of category 1, usage behaviors of users are slightly similar among all the users, as illustrated in Table 1. Thus, if the framework operates depending only on the first category, then such framework will not be sufficiently accurate to detect the behavioral change of users. Therefore, usage of the user behavior application, network, and memory should be secure; more categories provide more security to the mobile device. Tables 2 and 3 demonstrate that the user behavior is totally different.

The second step is to check the relationship among the behavior of users. The security level of the correlation among the behaviors of users should be less than 0.06 to permit automatic access to legitimate users. Otherwise, the framework will verify the behavior of users based on the SS level. Based on the verification setup of the correlation among behaviors of users, the method of Pearson is utilized to determine the relationship among six users for all categories. Meanwhile, the results for each category show r, the correlation between the users and the security level. If the relationship between users increased, then the security level decreased. The second step employed the method of Pearson to detect the relationship among six users. This step allows the framework to detect similar behaviors for the most usage application. Accordingly, the application of this method is investigated. Figure 9 shows the correlation among the behaviors of users for the first category.

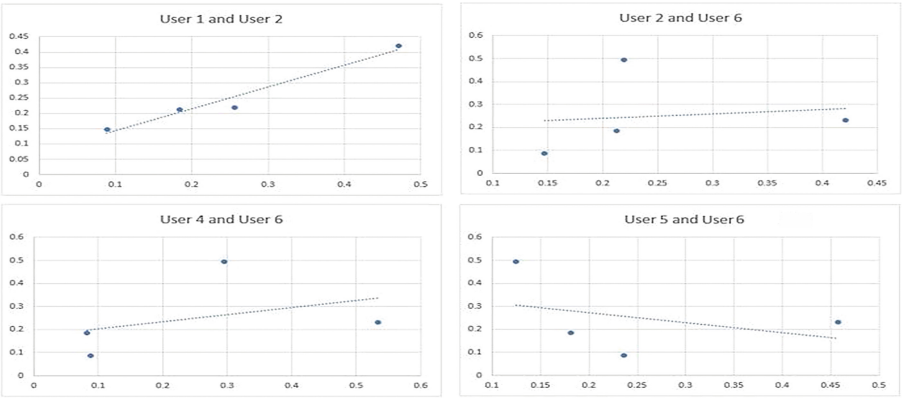


Fig. 9. The relationship between users in the first category

The correlation of behavior among users is a number between −1 and 1, which determines whether the users have a similar behavior. A number close to 1 indicates a positive linear correlation while a number close to −1 indicates a negative linear correlation. When the correlation between users is close to zero, no evidence of any relationship is observed between them.

Figure 9 shows a zero relationship between User1 and User6 in terms of their behaviors. The behaviors between User2 and User6, User3 and User5, and User5 and User6 are clearly different. Therefore, we can conclude that user identity varies from one user to another based on the data collected from their mobile devices. What can be confidently claimed is that the mobile data collected from the users do exhibit varying user behaviors.

Figure 10 illustrates that the performance of user behavior is totally different from one user to another. As clearly shown in Fig. 10, although the relationship between User2 and User6 is slightly similar, no relationship exists among all users. By contrast, Fig. 11 shows a totally different behavior. Thus, the main idea of using data profiling is

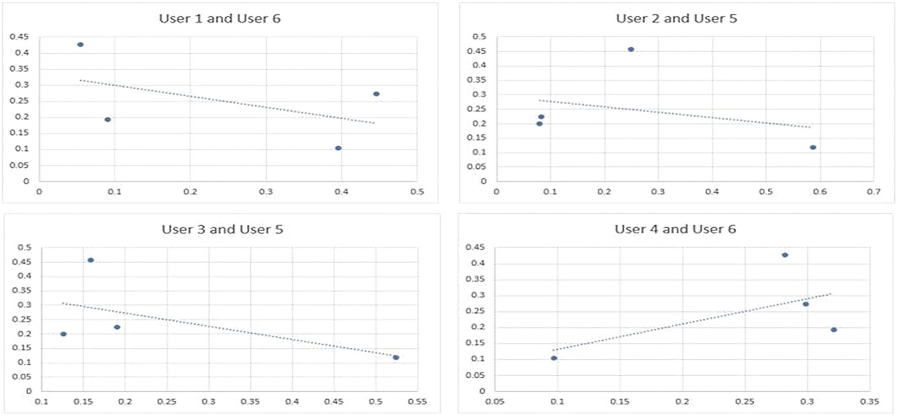


Fig. 10. The relationship between users in the second category

that failure in one category still provides the user with two other categories to authenticate the true user. Behavioral data profiling is sensitive to the tests and strongly detects the correlation among user behaviors.

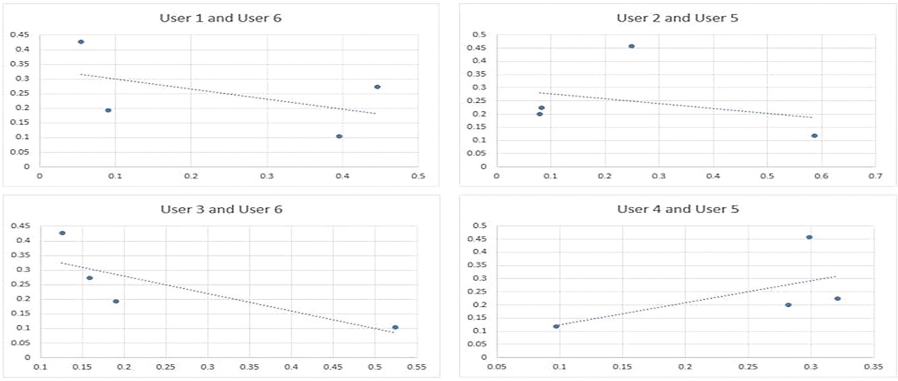


Fig. 11. The relationship between users in the third category

4 Conclusion

Behavioral data profiling authentication is specific of the data itself. Meanwhile, biometric authentication identifies the user based on the shape or behavior of the human body, where it is directly related to the device, such as face recognition, fingerprint, and gait. Therefore, our proposed technique is indirect, that is, the data do not directly come from the device but from different user behaviors. The experimental results show that the user behaves differently based on the data collected. Moreover, some data are useful while some are not. Using behavioral data to design an authentication model, which is advanced, continuous, and secure, is reasonable. Behavior is unique from one user to another in terms of using mobile devices. Previous authentication approaches, where the PIN login can be copied, are not sufficiently secure; thus, a secure mechanism is necessary. Behavioral data profiling has multiple data sources that check user identity from various sources based on the user behavior provides the system with accurate data, thereby ensuring the security of the mobile device. The main idea of using data profiling is that failing in one source/category can still afford the user with other sources/categories to authenticate the true user.

Acknowledgment. This work is funded by a research grant under the title 'Research on Big Data in Heterogeneous Fixed Wireless Network (304/PKOMP/650804/M151).

References

1. Guo, C., Wang, H.J., Zhu, W., et al.: Smart-phone attacks and defenses. In: hotnets III, San Diego, CA (2004)
2. Leavitt, N.: Malicious code moves to mobile devices. IEEE Comput. **33**(12), 16–19 (2000)
3. Dagon, D., Martin, T., Starner, T.: Mobile phones as computing devices: the viruses are coming! IEEE Pervasive Comput. **3**(4), 11–15 (2004)
4. Li, Q., Clark, G.: Mobile security: a look ahead. IEEE Secur. Priv. **11**(1), 78–81 (2013)
5. La Polla, M., Martinelli, F., Sgandurra, D.: A survey on security for mobile devices. IEEE Commun. Surv. Tutor. **15**(1), 446–471 (2013)
6. Draffin, B., Zhu, J., Zhang, J.: Keysens: passive user authentication through micro-behavior modeling of soft keyboard interaction. In: International Conference on Mobile Computing, Applications, and Services, pp. 184–201. Springer (2013)
7. Saaty, T.L., Vargas, L.G.: Hierarchical analysis of behavior in competition: prediction in chess. Syst. Res. Behav. Sci. **25**(3), 180–191 (1980)
8. Saaty, T.L.: The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In: Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 345–405 (2005)
9. Clarke, N., Karatzouni, S., Furnell, S.: Flexible and transparent user authentication for mobile devices. In: IFIP International Information Security Conference, pp. 1–12. Springer (2009)

An Efficient ElGamal Encryption Scheme Based on Polynomial Modular Arithmetic in $\mathbf{F_2^n}$

Tan Soo Fun[1(✉)] and Azman Samsudin[2]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
soofun@ums.edu.my

[2] School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Penang, Malaysia
azman@cs.usm.my

Abstract. The ElGamal cryptosystem was originally proposed by Taher ElGamal in 1985, in which its security level is based on the Discrete Logarithm Problem (DLP). ElGamal cryptosystem is relatively an expensive algorithm. For security guarantees, ElGamal cryptosystem requires modulo operation of large prime integer whose size range approximately from 1,024 to 4,096 bits. As a consequence of such requirement, the application of ElGamal cryptosystem is limited for securing only small messages such as secret keys. This paper aims to propose an efficient variant of ElGamal cryptosystem. The proposed scheme is designed based on quotient ring of polynomial, $Z_2[x]/<f(x)>$, where $f(x)$ is an irreducible polynomial. The decryption algorithm was further optimized with the use of the multiplicative inverse of the generator $g(x)$, which only generated once during the key generation algorithm, thus leading to a simpler and faster decryption process. The proposed scheme is as secure as the original ElGamal scheme, since both schemes are based on the DLP. The preliminary result shows that the proposed scheme minimizes complex arithmetic operations and achieves very practical performance compared to the classic ElGamal algorithm and its variants. The proposed F_2^n based ElGamal scheme outperforms the F_p based scheme by significantly reducing 69.74% of the numbers of required logic gates in the case study of VLSI implementation.

Keywords: ElGamal · Polynomial modular arithmetic · Binary fields

1 Introduction

In 1985, Taher ElGamal presented a new probabilistic public key cryptosystem [1] based on the concept of Diffie-Hellman Key Exchange. Currently, ElGamal public key scheme is being widely used in commercial open source applications such as Pretty Good Privacy (PGP) encryption (recent version PGP 3), free GNU Privacy Guard software, etc. [2]. However, one of the significant performance bottleneck of ElGamal cryptosystem is its expensive exponential calculation on big integers. Security of the ElGamal algorithm depends on the (presumed) difficulty of computing discrete logarithm in a large prime modulus, or so called as the Discrete Logarithm Problem (DLP). The DLP

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 99–107, 2018.
https://doi.org/10.1007/978-981-10-8276-4_10

requires modulus operation of large prime in the range of 1,024 bits to 4,096 bits in order to achieve sophisticated level of security protection. Therefore, the applicability of the ElGamal cryptosystem is limited to only securing small messages such as secret keys.

Recently, there are several variants of ElGamal cryptosystem [3–6, 18, 19] that had been proposed in order to enhance the performance of the ElGamal scheme. The finite fields of these variants can be classified into two categories: Prime Fields (F_p) and Binary Fields (F_2^n), which are further defined in the following:

Definition 1 (Prime Fields). *A field with finitely many prime elements is called a Prime Fields. We denote a prime field with p elements by F_p, where the p is a prime number.*

Definition 2 (Binary Fields). *A field of characteristic two F_2^n or GF(2^n) where n is an integer number greater than 0. In standard basis, an element of F_2^n can be represented as polynomial: $a(x) = a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1x^1 + a_0$ of degree $n - 1$ with coefficients in F_2.*

The original ElGamal scheme was designed based on F_p, where p is usually large enough to promise adequate security of the scheme against the possible cryptanalysis attacks such as chosen-plaintext attack (CPA) and non-adaptive chosen-ciphertext attacks (CCA1). Thus, leading to an expensive computation in finding a large prime p. To solve this problem, several variants of ElGamal scheme [3–6] has been proposed. On 2001, El-Kassar et al. [3] modified the ElGamal scheme into the domain of Gaussian integers and further extended it in the multiplicative group of $Z_p[x]/<x^2>$. Their main contribution is that no additional efforts are required for finding the prime p since the cyclic group used in their proposed scheme has an order larger than the square as in the original ElGamal scheme [1]. On 2002, Hwang and Chang [4] proposed another variant of ElGamal scheme by breaking a plaintext M into t pieces of $M_1, M_2, \ldots M_t$, where each piece is a 512 bits block. As compared to the original ElGamal scheme that requires a computation of $2t$ times exponential operations and t times multiplication operations during the encryption and decryption process, Hwang and Chang [4] have reduced the computational complexity which requires four times exponentiation operations, t times exclusive-OR (*XOR*) operations, t times multiplication operations and t times square operations. After about a decade, Hu et al. [5, 6] further modified ElGamal scheme by placing the plaintext in the exponent by using the method as pointed out by [7] and the decryption process is enhanced by the Chinese Remainder Theorem (CRT). Unfortunately, the primitive design of these schemes [3–6] are based on F_p, therefore, the scheme seemed to pose rather inherent efficiency bottlenecks due to computational burden associated with the large prime p in F_p.

In general, there are two main advantages of F_2^n over F_p. Firstly, F_2^n in hardware implementation provides a resistance to tampering attacks as compared to F_p in software implementation [9, 10, 13, 14]. Secondly, the computation operation i.e. addition, subtraction, multiplication, division in F_2^n are more efficient than in F_p [8–12]. The bit addition and subtraction in F_2^n are executed over modulus *2*. Thus, leading a fast XOR

gates in hardware implementation. On the other hand, both addition and subtraction operations in F_p, in contrast, needs up to three times more operations: the actual addition, a comparison of the result with the prime p, and a modular reduction afterwards. Meanwhile, the bit multiplication over F_2^n is more efficient than F_p, which requires a combinational of bit-shift and XOR gates in hardware implementation. Satoh and Takano [15] shows that the F_2^n operations can be performed approximately six times faster than F_p. Wenger and Hutter [16] further investigated and compared both F_2^n and F_p in terms of their speed, power and energy. Their experiment result further proven the F_2^n based processor is 3.3 times faster, 20% smaller, uses 16% less power, and needs 3.9 times less energy compared to the F_p based processor.

In this paper, we proposed a new variant of ElGamal scheme by exploiting the efficient computation of F_2^n in order to cope with the inherent performance bottleneck of ElGamal. Besides that, the proposed new optimization of ElGamal algorithm is less complex compared to existing variants of ElGamal. The rest of this paper is organized as follows: In the next section, the original ElGamal cryptosystem and related ElGamal variants in F_2^n are visited. Section 3 introduces the proposed variant of ElGamal scheme based on F_2^n. Section 4 compares the proposed work with classic ElGamal and existing variants of ElGamal. Lastly, Sect. 5 concludes.

2 ElGamal Public Key Encryption Scheme and Related Works

Public key cryptosystem was first proposed by Diffie and Hellman [17] in 1976, in order to solve the key management problem. In public key cryptosystem, each party obtains a pair of keys, called the public key and the secret (private) key. The public key is published and widely distributed, while the secret key is never revealed. This paper denotes the definition of Public Key Encryption (PKE scheme) as following:

Definition 3 (Public Key Encryption (PKE) Scheme): *A Public Key Encryption (PKE) scheme is a triplet of polynomial-time algorithms (KeyGen, En, Dec), where:*

KeyGen:	*Key Generation Algorithm which take a security parameter, 1^{λ} and returns a public key pair (sk, pk)*
Enc:	*Encryption Algorithm which take a public key (pk) and a plaintext message M and return a ciphertext C*
Dec:	*Decryption Algorithm which take a secret key (sk) and a ciphertext C and returns either a plaintext message M or a symbol $\perp$ that indicates invalid ciphertext C*

In this section, we recall the original ElGamal scheme and several existing variants of ElGamal scheme that are based on F_2^n. The ElGamal cryptosystem is a well-known public key cryptosystem that extends the Diffie-Hellman Key Exchange concept into Public Key Encryption (PKE) scheme algorithm. The ElGamal PKE encryption scheme

is formally defined as a triple (*KeyGen, Enc, Dec*) of probabilistic polynomial-time algorithms defined as follows:

Definition 4 (ElGamal Public Key Encryption (PKE) Scheme): *The ElGamal public key encryption scheme is defined as following:*

KeyGen:	*Take a security parameter* 1^{λ} *and outputs the system parameters* (p, q, g)*, where the* (p, q, g) *is an instance of Discrete Logarithm Problem collections.* p *is a uniformly chosen prime of length* $\lvert p\rvert = n + \alpha$*, for a specified constant* α*;* g *is a uniformly chosen generator of the sub-group* G_q *of the prime order of* q *of* Z_p^**, where* $q = (p-1)/\beta$*, where* β *is a specified relatively small integer. Then, a public key is a quadruple of* (p, q, g, k) *and a private key consists of a quadruple of* (p, q, g, x)*, where* x *is a uniformly chosen element of* Z_q *and* $k \equiv g^x \bmod p$
Enc:	*With the input* (p, q, g, k)*, for a message* $m \in G_q$*, an element* r *is uniformly chosen of* Z_q *and outputs* $Enc\left[(p, q, g, k), m\right] = \left[g^r \bmod p, m \times k^r \bmod p\right]$ *as a tuple of* (C_1, C_2)
Dec:	*With the input* (p, q, g, x) *for a ciphertext* (C_1, C_2)*, outputs* $Dec\left[(p, q, g, x), (C_1, C_2)\right] = C_1^{-x} \times C2$

As noted here, the primitive design of ElGamal PKE scheme is heavily relies on modular exponentiation operation. In general, the modular exponentiation operation is a classical operation for scrambling data and it is widely used in several cryptosystems such as RSA, Diffie-Hellman Key Exchange scheme, etc. Although in ElGamal PKE scheme, the base generator g and the modulus prime p are known in advance, some pre-computation works (i.e. g powers of g can be further precomputed and saved) are allowed to confront performance issues; however, ElGamal PKE scheme still consumes computation resources as it requires four times modular exponentiation operation $\left(g^x, g^r, k^r, C_1^x\right)$. Besides that, the security of ElGamal PKE scheme is established based on the difficulty (assumed) of computing the Discrete Logarithm in finite fields, F_p. To achieve a reasonable security level, ElGamal PKE scheme requires the extension of degree n to about 1000 bits. This big integer calculation requirement further leads to relatively low performance on the ElGamal PKE and it is widely recognized as a major shortcoming in practical applications. For instance, for encrypting a plaintext with message size 256 bits, the size overhead of ElGamal PKE scheme is approximately ten times as the plaintext size [18]. Therefore, current applications of ElGamal PKE is limited to hybrid cryptosystem such as PGP, where the corresponding message is encrypted with lightweight symmetric key and ElGamal PKE scheme is used to encrypt that particular symmetric key.

In the past few years, several variants of ElGamal PKE schemes were proposed in order to solve the performance bottleneck problem by exploiting the efficient computation of finite fields, F_2^n. In 2005, Kassar and Haraty [19] had extended the ElGamal PKE scheme to the setting of quotient ring of polynomials over a finite field $Z_p[x]/<f(x)>$ where $f(x)$ is not necessarily irreducible. In their first attempt, they have set the $f(x)$ as reducible polynomial. In order for the group of units $U(of\ Z_2[x]/<f(x)>$

to be a cyclic group, where p is an odd prime, $f(x)$ must be a square power of only one linear irreducible polynomial. Thus, they proposed a variant of ElGamal PKE scheme over $Z_2[x]/<x^2>$. However, one of the significant drawback of this scheme is subjected to critical security flaws. The ElGamal PKE scheme over $Z_2[x]/<x^2>$ confines the range of private elements x and random number r can be chosen. Both x and r can be chosen only from the range of *2* to $\phi(x^2) - 1$, where $\phi(x^2) - 1$ is equal to five. Therefore, an adversary can simply conduct brute force attack on this scheme with only five attempts.

On the same year, Kassar and Haraty [19] proposed another extension of the ElGamal public key cryptosystem by employing the group of units of $Z_2[x]/<f(x)>$, where $f(x) = f_1(x) \times f_2(x) \ldots f_t(x)$ is a product of irreducible polynomials whose degrees are pairwise relatively prime. Inherently, this setting extends the security assumption of ElGamal PKE scheme to both Modulus Factorization Problem (MFP) and Discrete Logarithm Problem (DLP). However, the payoff cost is a much heavier computation as compared to the original ElGamal PKE scheme. For finding $f(x)$, the scheme consumes t times multiplication operation of irreducible polynomials. Finding $\phi(f(x))$ involves t times modular exponential operation, where $\phi(f(x))$ is computed as $\phi(f(x)) = (2^{d1} - 1)(2^{d2} - 1) \ldots (2^{dt} - 1)$. Furthermore, this scheme is still a scheme over prime fields F_p, which extra resources have to be allocated for finding pairwise relatively prime integers $d_1, d_2, \ldots, d_t$. Therefore, this scheme does not fully exploits the efficient computation over binary fields F_2^n.

In 2012, Haraty et al. [20] presented another extension of the ElGamal public key cryptosystem. In order to apply the ElGamal PKE scheme in the second group of units of $Z_2[x]/<f(x)>$, where $f(x)$ is an irreducible polynomial, they have combined both characterization of the prime fields F_p and the binary fields F_2^n. However, this combination does sacrifice the scheme performance as compared to their proposal on 2005 [19]. Firstly, the implication of prime fields F_p to generate a pair of public key and private key has burdening the computation. Secondly, the complexity of the key generation algorithm as compared to the original ElGamal PKE scheme does not lessening the performance bottleneck of ElGamal PKE scheme. Furthermore, their security level is considered weaker than the original ElGamal PKE scheme, since the encryption of message m is now solely depend on random integer r and without using the generated public key.

3 Proposed Scheme

In this section, a new scheme is proposed to overcome the performance bottleneck of ElGamal PKE. We proposed a new variant of ElGamal PKE scheme by fully exploiting the efficient computation over F_2^n. The proposed scheme is based on quotient ring of polynomial, $Z_2[x]/<f(x)>$ where $f(x)$ is an irreducible polynomial. To recover message m correctly, the scheme uses multiplicative inverse of generator $g(x)$, which only generated one time during the keyGen algorithm, instead of generated $\phi(f(x))$ twice during

the KeyGen and Dec algorithm as found in [18, 19]. Thus, the proposed scheme has a simple decryption process as compared to [1, 3–6, 18, 19]. We now present, in generic form, our proposed extension of ElGamal PKE scheme, with security parameter, λ is specified as follows.

KeyGen:	*Take a security parameter,* 1^{λ} *and outputs the system parameters* $[f(x), g(x), g^{-1}(x)]$, *where* $f(x)$ *is a uniformly chosen irreducible prime polynomial with a degree n and* $g(x)$ *is a uniformly chosen generator of primitive root of* $f(x)$ *and* $g^{-1}(x)$ *is an multiplicative inversion of a element* $g(x)$ *such that* $g(x) \times g^{-1}(x) \equiv 1$ *mod* $f(x)$. *Then, a public key is a tuple of* $[f(x), g(x), k(x)]$, *and private key is a tuple of* $[f(x), g(x), x]$, *where x is a uniformly chosen element of* 2^{n-1}, *and* $k(x) \equiv g^{-1}(x)^{x} mod\ f(x)$
Enc:	*With the input* $[f(x), g(x), k(x)]$, *for a message* $m(x) \in Z_2[x]/<f(x)>$, *an element r is uniformly chosen of element of* 2^{n-1} *and outputs* $Enc[(f(x), g(x), k(x)), m(x)] = [g(x)^{r} mod\ f(x), m(x) \times k(x)^{r} mod\ f(x)]$ *as a tuple of* $[C_1(x), C_2(x)]$
Dec:	*With the input* $[(f(x), g(x), x)]$, *for a ciphertext* $[C_1(x), C_2(x)]$, *outputs* $Dec[(f(x), g(x), x), (C_1(x), C_2(x))] = C_1(x)^{x} \times C_2(x)$

Numerical Example:

KeyGen:	For key generation, $f(x) = x^5 + x + 1$ is a uniformly chosen irreducible prime polynomial with a degree $n = 5$ and $g(x) = x^3 + x + 1$ is a uniformly chosen generator of primitive root of $f(x)$ and $g^{-1}(x) = x^4 + x^3$ is a multiplicative inversion of an element $g(x)$ such that $g(x) \times g^{-1}(x) \equiv 1 \mod f(x)$. Then, a public key is a tuple of $[(x^5 + x + 1), (x^3 + x + 1), (x^3 + x)]$, and private key is a tuple of $[(x^5 + x + 1), (x^3 + x + 1), 25]$, where $x = 25$ is a uniformly chosen element of 2^{5-1}, and $k(x) \equiv (x^4 + x^3)^{25} \mod x^5 + x + 1 \equiv x^3 + x$
Enc:	With the input $[(x^5 + x + 1), (x^3 + x + 1), (x^3 + x)]$, for a message $m(x) = x^4 + x^2 + x \in Z_2[x]/<f(x)>$, an element $r = 19$ is uniformly chosen of element of 2^{5-1} and outputs $Enc[(f(x), g(x), k(x)), m(x)] = [(x^3 + x + 1)^{19} \mod x^5 + x + 1, (x^4 + x^2 + x) \times (x^3 + x)^{19} \mod f(x)]$ as a tuple of $[(x^4 + x^3 + x^2 + x), (x^3 + x + 1)]$
Dec:	With the input $(x^5 + x + 1), (x^3 + x + 1), 25$, for a ciphertext $[(x^4 + x^3 + x^2 + x), (x^3 + x + 1)]$ outputs $Dec[((x^4 + x^3 + x^2 + x), (x^3 + x + 1), 25)), ((x^4 + x^3 + x^2 + x), (x^3 + x + 1)] = (x^4 + x^3 + x^2 + x)^{25}.(x^3 + x + 1) = x^4 + x^2 + x$

The correctness of the proposed scheme follows immediately from the correctness of the original ElGamal PKE scheme and the correctness of Euclid's Algorithm in finding multiplicative inverse of $g(x)$.

4 Preliminary Result

In this section, the proposed scheme is compared with other two variants of ElGamal PKE Scheme [19, 20] that are based on F_2^n. The schemes' performance are measured in terms of the number of operations for each algorithm in PKE scheme and the result is further summarized in Table 1 as follows.

Table 1. Computational performance of ElGamal variants over F_2^n

Computation operations		Kassar and Haraty [19]	Haraty et al. [20]	Proposed scheme
Key-Gen	Modular addition/subtraction	$1 + t$	1	1
	Modular multiplication/division	$2t$	$2t$	–
	Modular exponentiation	$1 + t$	$2t + 1$	2
	Modular inversion	–	–	1
Enc	Modular addition/subtraction	–	–	1
	Modular multiplication/division	1	1	t
	Modular exponentiation	2	4	$2 + t$
	Modular inversion			
Dec	Modular addition/subtraction	$1 + t$	$1 + t$	–
	Modular multiplication/division	$1 + t$	t	1
	Modular exponentiation	$1 + t$	$2 + t$	1
	Modular inversion	–	–	–

(t denotes a t times modular operations, e.g.: $f(x) = f_1(x) \times f_2(x) \ldots f_t(x)$ involves t times modular multiplication)

The result of the comparison on computation operations revealed that the proposed scheme has a minimal computation operations compared to [19, 20]. In the key generation algorithm, the proposed scheme has less operations as compared to [18, 19], which suffers from finding pairwise relatively prime integers and the totient for the polynomial $f(x)$. To decrypt message m correctly, the proposed scheme applies the multiplicative inverse of generator $g(x)$ instead of using $\phi(f(x))$. Therefore, there is an extra work to compute the multiplicative inverse of generator $g(x)$ in the proposed scheme. However, this modular inversion only computed once during the KeyGen algorithm, instead of computing $\phi(f(x))$ twice during the KeyGen and Dec algorithm as [19, 20]. Since finding $\phi(f(x))$, where $\phi(f(x)) = \left(2^{d1} - 1\right)\left(2^{d2} - 1\right) \ldots \left(2^{dt} - 1\right)$, involves t times of multiplications, t times exponentiation, and t times subtractions, thus leading to heavier computations in KeyGen and Dec algorithms for both schemes [19, 20]. In the nutshell, the proposed scheme has theoretically lessening the performance bottleneck of ElGamal PKE scheme by fully exploiting the efficient binary fields.

5 Conclusion

We have proposed an efficient new variant of ElGamal cryptosystem, fid for real-world applications. By taking the advantages of fast, efficient and better security of the binary fields, F_2^n, the proposed scheme is designed based on quotient ring of polynomial, $Z_2[x]/<f(x)>$ where $f(x)$ is an irreducible polynomial. We have also shown that with the new optimization algorithm, the proposed scheme is computationally less complex and enjoys a simple decryption process than the other variants of ElGamal. The proposed scheme was proven secure as the original ElGamal scheme as it is based on the DLP. In future work, we will conduct experiment testing and evaluation of our scheme in real-world applications.

References

1. Elgamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inf. Theor. **31**(4), 469–472 (1985)
2. Callas, J., Donnerhacke, L., Finney, H., Thayer, R.: OpenPGP message format (RFC4880). Zhurnal Eksp. i Teor. Fiz., 1–90 (1998)
3. El-Kassar, A.N., Rizk, M., Mirza, N.M., Awad, Y.A.: El-Gamal public-key cryptosystem in the domain of Gaussian integers. Int. J. Appl. Math. **7**(4), 405–412 (2001)
4. Hwang, M., Chang, C.: An ElGamal-like cryptosystem for enciphering large messages. IEEE Trans. Knowl. Data Eng. **14**(2), 445–446 (2002)
5. Hu, Y., Martin, W., Sunar, B.: Enhanced flexibility for homomorphic encryption schemes via CRT. Appl. Crypt. Netw. (2012)
6. Hu, Y.: Improving the Efficiency of Homomorphic Encryption Schemes. Ph.D. thesis, Worcester Polytechnic Institute (2013)
7. Cramer, R., Shoup, V.: Universal hash proofs and a paradigm for adaptive chosen ciphertext secure public-key encryption. In: Advanced Cryptology—Eurocrypt 2002. LNCS, vol. 2332, pp. 45–64. Springer, Heidelberg (2002)
8. Wu, H.: Efficient computations in finite fields with cryptographic significance. Ph.D. thesis, University of Waterloo (1998)
9. Alkalbani, A.: Comparison between RSA hardware and software implementation for WSNs security schemes. In: Proceedings of International Conference on Information and Communication Technology for the Muslim World, pp. 84–89 (2010)
10. Agarwal, K., Basu, S.. Venkateswarlu, V.: Optimized architecture of low power, high performance multiplier for crypto chips. Int. J. Comput. Appl. Eng. Sci. **1**(Special Issue), 282–285 (2011)
11. Abhijith, P., Srivastava, M.: High performance hardware implementation of AES using minimal resources. In: Proceedings of International Conference on Intelligent Systems and Signal Processing (ISSP), pp. 338–343 (2013)
12. Stallings, W.: Cryptography and Network Security: Principles and Practice, 2nd edn. Prentice Hall, New Jersey (1999)
13. Fiaz, F., Masud, S.: Design and implementation of a hardware divider in finite field. Natl. Conf. Emerg. Technol. **6**, 167–170 (2004)

14. Tawalbeh, L.A., Tenca, A.F.: An algorithm and hardware architecture for integrated modular division and multiplication in GF(p) and GF(2^n). In: Proceedings of 15th IEEE International Conference on Application-Specific Systems, Architectures and Processors, pp. 247–257 (2014)
15. Satoh, A., Takano, K.: A scalable dual-field elliptic curve cryptographic processor. IEEE Trans. Comput. **52**(4), 449–460 (2003)
16. Wenger, E., Hutter, M.: Exploring the design space of prime field vs. binary field ECC-hardware implementations. Inf. Secur. Technol. Appl. **7161**, 256–271 (2012)
17. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inf. Theor. **22**(6), 644–654 (1976)
18. Kim, M., Kim, J., Cheon, J.: Compress multiple ciphertexts using ElGamal encryption schemes. J. Korean Math. Soc. **50**(2), 361–377 (2013)
19. El-Kassar, A.N., Haraty, R.: ElGamal public-key cryptosystem in multiplicative groups of quotient rings of polynomials over finite fields. Comput. Sci. Inf. Syst. **2**(1), 63–77 (2005)
20. Haraty, R., Kassar, A.N., Fanous, S.: Hardening the ElGamal cryptosystem in the setting of the second group of units. Int. Arab J. Inf. Technol. (2012)
21. Stern, J.: Evaluation report on the discrete logarithm problem over finite fields (2001)
22. Tsiounis, Y., Yung, M.: On the security of EIGamal based encryption. In: Public Key Cryptography. LNCS, vol. 1431, pp. 117–134 (1998)
23. Shannon, C.E.: Communication theory of secrecy systems. MD Comput. **15**(1), 57–64 (1948). 1945
24. Katz, J., Lindell, Y.: Introduction to Modern Cryptography, 1st edn., pp. 1–498. CRC Press, Washington (2007)
25. Loftus, J., May, A., Smart, N., Vercauteren, F.: On CCA-secure somewhat homomorphic encryption. In: Selected Areas on Cryptography. LNCS, vol. 7118, pp. 55–72 (2012)
26. Alfred, A., Menezes, J., Oorschot, P.C.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)
27. Fournaris, A.P., Koufopavlou, O.: GF (2K) multipliers based on montgomery multiplication algorithm. In: Proceedings of International Symposium on Circuits and Systems, vol. 2, pp. 849–852 (2004)
28. Guajardo, J., Güneysu, T., Kumar, S.S., Paar, C., Pelzl, J.: Efficient hardware implementation of finite fields with applications to cryptography. Acta Appl. Math. **93**(1), 75–118 (2006)

Proposed DAD-match Mechanism for Securing Duplicate Address Detection Process in IPv6 Link-Local Network Based on Symmetric-Key Algorithm

Ahmed K. Al-Ani[1(✉)], Mohammed Anbar[1(✉)], Selvakumar Manickam[1], Ayman Al-Ani[1], and Yu-Beng Leau[2]

[1] National Advanced IPv6 Center, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia
{ahmedkhallel91,anbar,selva}@nav6.usm.my, ayymn@yahoo.com
[2] Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
lybeng@ums.edu.my

Abstract. Duplicate address detection (DAD) is an essential procedure of neighbor discovery protocol (NDP). Further, DAD process decides in case an IP address is in conflict with other nodes. In usual DAD process, the target address to be identified is multicast via the network, which provides an ability for malicious nodes to attack. A malicious node can send a spoofing reply to prevent the address configuration of a normal node, and thus, a denial of service (DoS) attack is launched. This study proposes a new mechanism to hide the target address in DAD, which prevents an attack node from reaching target node. If the address of a normal node is identical to the detection address, then its IP address should be able to decrypt the random word and compare the decryption with decryption in "DADmatch" tag. Consequently, DAD can be successfully completed. This process is called DAD-match. We expect DAD-match will provide a lightweight security resolution and less complexity as well as fully prevent of DoS attacks during DAD process in IPv6 link-local network.

Keywords: Duplicate address detection · DoS attack · IPv6 security

1 Introduction

Internet Protocol Version 6 (IPv6) [1] has been enhanced to overcome the shortfalls of IPv4 address space exhaustion limitation. Furthermore, IPv6 intentionally made to substitute IPv4 which still hold the massive majority of web traffic in 2017. In April 2017, the proportion of clients reached Google services over IPv6 exceeded16% [2]. IPv6 was engineered and designed to make the network more proficient, quicker and furthermore to extend its reach to a bigger area by using a simpler and less complex header format and by producing new mechanisms i.e. address auto-configuration and neighbor discovery (NDP) concept which is a new protocol implemented in IPv6 and provide several of tasks.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 108–118, 2018.
https://doi.org/10.1007/978-981-10-8276-4_11

The IPv6 execution is yet standing up to challenges especially on its security architecture. Notwithstanding the convention was planned and refined with compulsory IP Security (IPSec) [3]. Nevertheless, IPv6 still has some susceptibilities that can be utilized by malicious nodes to expose the other nodes as well as the network. Besides, numbers of susceptibilities and attacks in IPv6 [4] revealed that the security objective, reliability services, and availability have not been reached and that the general risk situation is getting worse. Thus, the most major challenge to the Internet today is to develop and improve the security level of the network [5]. Researcher in [6] has presented their studies on the vulnerability of NDP. An attacker can abuse vulnerabilities in ND to launch a DoS attack, hijack traffic or decrease network performance.

The paper proposes a security mechanism based on an encryption algorithm to protect DAD procedure which is one of NDP tasks in IPv6 link-local network, this security mechanism called DAD-match. The rest of this paper is structured as follows. Section 2 provides a background of NDP, SLACC and DAD with its security issues. Related works on securing DAD as explained in Sect. 3. Section 4 shows an overview of the DAD-match proposal, followed by Sect. 5 which is a conclusion.

2 Background

NDP is a number of messages and procedures which control the connections between neighbors i.e. routers and hosts in the same link-local network [7]. In addition, NDP offers new functionalities that are utilized for generating an IP addresses through using Stateless Address Auto-configuration (SLAAC) [8], confirming the uniqueness of the produced IP address using DAD, resolving IP addresses to their link-layer addresses, keeping reachability track with neighbors around using Neighbor Unreachability Detection (NUD) and utilizing Redirect Message (RM) for advertising a superior next-hop [9]. NDP employ five ICMPv6 messages [10], they are; Router Solicitation (RS), Router Advertisement (RA), Neighbor Solicitation (NS), Neighbor Advertisement (NA) and Redirect Message (RM) messages, where each message used for a particular purpose.

Furthermore, NDP plays such an important protocol in IPv6 and facing many attacks. Most common NDP attacks are; Router Advertisement Spoofing Attack, NS and NA spoofing, Malicious Last-Hop Router Attack, Spoofed Redirect Message Attack, Replay Attacks, ND Flooding DoS Attack and lastly DoS on Duplicate Address Detection which is the target of this paper.

2.1 Stateless Address Auto-Configuration (SLAAC)

SLAAC is a new feature of IPv6 use for generating an IP address automatically for the host [11]. Once the node joins a network, it configures its IP address without need any human intervention. SLAAC works in a “Plug and Play” fashion. Furthermore, there are different ways to generate IP addresses such as; EUI-64 method offered by IEEE (Internet Engineering Task Force) [11]. EUI-64 is an address format, which is done by reformatting the Ethernet Interface “unique 48-bit MAC” address to match the EUI-64 specification. However, since this method generates same IID whenever a node joins a

network, so it makes intruders easy to track the node. In addition, Privacy extension method is an another method to generate an IP address randomly [8]. By using this method, the address will keep change over time, thus it makes difficult intruders to identify the target host address [12]. For the newly proposed mechanism which is called DAD-match, it is going to use privacy extension method for generating an IP address throughout DAD procedure in IPv6 link-local network as long as this method provide more security compared with the EUI-64 method.

2.2 Duplicate Address Detection (DAD) with Its Security Issues

DAD is a new process in IPv6 network, which make sure that all nodes able to configure their interface identifier (IID) "IP Address" are unique in the same link, every host located on the same link must execute DAD procedure before configuring its IP address as aforesaid in [13]. Where there is a new node joins IPv6 link-local or perhaps an existing node placed on the same link planned to generate a new IP address. Moreover, the uniqueness of the tentative IPv6 address at this point is still questionable, the target node should make sure there is no node on the same link has the same IP address, this is executed through multicasting (NS) messages which are carrying the tentative IP address to all neighboring nodes located on the same link. In case of the tentative address is has been already assigned by another neighboring host, that specific neighboring node should send a (NA) message as a response to NS message [14]. While, if the new node has not received any response NA message from the neighboring nodes, it means the newly generated address is unique and there is no any other neighboring node used it. Therefore, a node now is able to use it and communicate with other neighboring nodes.

Since NDP messages are insecure by design [15]. Thus, any potential encroaching host can exploit this phenomenon and can abuse the DAD procedure by controlling NS and NA messages. Therefore, when another host performs DAD procedure an assailant can interrupt the verification process by transmitting fake messages in response. Researches [16] have demonstrated that DAD procedure is uncovered to Denial of Service (DoS) attacks. DoS attack on DAD procedure, an attacker deliberate to make the target host is unable to obtain an IP address by claiming the existence of tentative

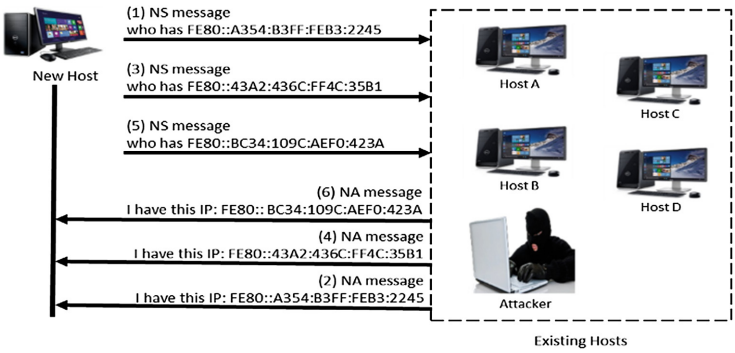


Fig. 1. DoS attack on DAD process

IP address by sending a fake NA messages in reply to its NS messages. Therefore, victim host will be unable to verify the uniqueness of tentative IP address. Thus, the IPv6 host cannot obtain an IP address due to DAD procedure failure [18] (Fig. 1).

3 Related Work

Many researchers proposed different mechanisms to secure DAD procedure in IPv6 link-local network. This section highlights the most common related works with their limitations.

Secure Neighbor Discovery (SeND) [19] introduce a new options add to NDP to secure the IPv6 network, namely; Cryptographically Generated Address (CGA) option to verify CGA senders. SEND uses CGAs to ensure the ownership of the claimed IPv6 address as defined in RFC 3972 [21], Rivest-Shamir-Adleman cryptosystem (RSA) signature option to attach a public key-based signature and Times stamp [20] and in order to avoid replay attacks Nonce option was used. SeND also come out with two new ICMPv6 messages types like Certificate Path Advertisement (CPA) and Certificate Path Solicitation (CPS) [19]. SeND tried to prevent a number of IPv6 attacks such as NS and NA spoofing attacks, DoS-on-DAD attacks, NUD failure attacks, RS and RA attacks and NDP DoS attacks [15].

However, many studies [17, 22] have indicated that SeND mechanism is not recommended to implement as an NDP extension in IPv6 link-local network because of its security options like RSA options CGA option. In addition, CGA failed to differentiate between legitimate and illegitimate node. Therefore, a malicious node able to capture the NDP messages i.e. NS and NA messages and manipulate the CGA parameter, as the result the attacker will exploit the target node during NDP processes. Furthermore, other drawback of SeND implementations with RSA signature is that it required a significant amount of process due to its designed technique [23]. Therefore, due to those drawbacks attacker can abuse SeND mechanism by engaging IPv6 hosts in neighbor discovery messages verification process which can prompt DoS attack on the particular host within DAD procedure in IPv6 link-local network.

In the other hand, Praptodiyono [24] proposed a new mechanism called Trust-ND. According to this research, the proposed mechanism is viewed as lightweight because of its design. Trust-ND utilizes SHA-1 hash function to fulfill the security necessities [26]. Trust-ND mechanism presents another security choice option known as Trust-Option. Based on study [24] shows the Trust-Option is attached to every ND message to guarantee the protected correspondence between nodes within DAD procedure in IPv6 link-local network. Since, the Trust-ND mechanism depends on SHA-1 hash function [27], it performs address verification speedier contrasted with the SeND mechanism as cited in [15]. Researchers have demonstrated that Trust-ND is a lightweight mechanism for IPv6 DAD procedure. However, research [28] has indicated that SHA-1 hash functions are vulnerable to hash collision attacks. Based on these studies, any malicious host can create hash collision attacks against SHA-1. Since the Trust-ND mechanism depends upon SHA-1 cryptographic hash function to fulfil its security necessities. This makes it susceptible to collision attacks that can cause DoS attacks on DAD procedure

in IPv6 link-local communication. Therefore, because of its architecture Trust-ND mechanism cannot be an appropriate security system for DAD procedure in IPv6 network [29].

Recent research on 2016 aims to secure IPv6 DAD procedure in link-local network by extended the ND header called Secure-DAD [29]. Secure-DAD mechanism propose a new notion called Secure-tag option which append to NS and NA messages exchange among the hosts within DAD procedures in IPv6 the link-local network. This Secure-tag contains "message authentication code" (MAC) to identify the legitimate from the illegitimate messages. Secure-DAD mechanism redesigned the NS and NA message via add the Secure-tag option to become Secured NS and Secured NA messages respectively. The study in [29] considers a comparison between Secure-DAD and Trust-ND, the result show that the Secure-DAD can protect DAD procedure in IPv6 link-local network better than Trust-ND in terms of processing time. However, sending the NS message to all-nodes on the same network will cause network overhead. In addition, Secure-DAD consumes high computational cost and adds more processing overhead in verification process due to its design. According to [30] Secure-DAD mechanism not easy to deploy.

To conclude, the most related works were suffering from the complexity due to their designs. Besides, added more processing for the verification process, as well as increase the network overhead bandwidth utilization. With these limitations, this study attempts to propose a new mechanism can provide a better security for DAD procedure in IPv6 link-local network as explained in the next section.

4 Proposed DAD-match Mechanism

This section shows the main goals of DAD-match mechanism and the NS/NA message format that are going to use throughout DAD procedure in IPv6 link-local network. Furthermore, the main stages, as well as the workflow of the DAD-match mechanism are presented and supported with an extensive example.

4.1 Design Goal of DAD-match Mechanism

Considering the weaknesses of the current proposed mechanisms to secure DAD procedure in IPv6 communication, this paper aims to propose a new mechanism by redesigning the DAD procedure to overcome the existing mechanisms limitations that are highlighted in the previous section and get a sufficient secure for DAD procedure in IPv6 link-local network.

The absence of integrity authentication is the main issue of the current DAD mechanism. Due the receiver has no way to identify between a fake and the legitimate message. Additional limitation of the related proposed works is the intricacy of the message creation, message authentication as well as the necessary resources to procedure the message. Further, reveal the tentative IP address throughout DAD procedure in IPv6 network leads to the inherent vulnerability of DAD like allow the malicious nodes of launch targeted attacks. As a result, if the goal address detection can be secured,

then DoS attacks can be efficiently prevented. To accomplish the main function of DAD, the plan goals of DAD-match are the following:

- It does not reveal the tentative IP address within DAD procedure.
- It should avert DoS attack on DAD.

DAD-match proposed relies on encryption "Symmetric Algorithms" also called "Secret-Key" or "Private-Kay" to secure DAD procedure in IPv6 link-local. The following subsections explain how DAD-match work.

4.2 Message Format of DAD-match

The traditional NDP message is composed of three parts, namely, Ethernet Header, IPv6 Header and ICMPv6 part. The values of the "Type" field for NS and NA are 135 and 136, respectively. The "Option" field has different functions based on the types of the ICMPv6 messages. Which usually keep the MAC address of the host in NDP. DAD-match mechanism propose a new message format through redesign the NS and NA messages to become *NS-match and NA-match*, and its type field values are 200 and 201 respectively since DAD-match under experimentation [31]. Compared with NDP message, DAD-match adds a new field called "DAD*match*" tag which stores the random word (w) and the word encryption $\left(E_w\right)$ through using the first 40 bit of host portion as a secret key.

4.3 The Stages of DAD-match Mechanism

DAD-match mechanism proposal based on five main stages to secure DAD procedure in IPv6 link-local network as illustrated below:

- **Generation Stage (Stage One):** In the first stage, the target host will generate an IP address as a tentative IP through using a privacy extension method, as mentioned above privacy extension method can generate a random IP address and that will make it harder for the attacker to guess the IP address. In addition, pick the first 40 bit of target host IP portion to use for encryption and decryption processes.
- **Encryption Stage (Stage Two):** Herein, at the both sides sender NS and the receiver NA will use this stage, which is generating a random word (w) and the first 40 bit of target IP portion will use as a secret key for encryption $E(w)$ and decryption D(w) during verification process. Then, insert the random word (w) and the random word encryption $E(w)$ in DADmatch tag.
- **Craft Stage (Stage Three):** Subsequently, NS-match/NA-match should create in this stage and append the DADmatch tag to NS-match/NA-match messages with an unspecified source address (::) and Solicited Multicast address (SNMA) will be the destination address.
- **Sending Stage (Stage Four):** NS-match/NA-match packet messages will send to (SNMA) FF02::1:FFSS:SSSS/104, where S represents the last 24-bit of target IP address, as a result, only the existing hosts on the same link which has the same SNMA address will receive the message, this stage called Sending Stage.

- **Messages Authentication Stage (Stage Five):** Message filtering stage will verify the NS-match and NA-match messages to detect whether it comes from a fake or legitimate host.
  - For NS-match verification, the existing host who will receive the NS-match message, firstly must check the existing of DADmatch tag, if it does not exist, NS-match message should discard. Otherwise, the process will proceed.
  - For NA-match message verification, at the first, each NA-match message without DADmatch tag will discard, moreover, the target host decryption must match the random word in the DADmatch tag.

Furthermore, if the target host did not receive any NA-match message as a response to NS-match message or 3 s left and no response from the other existing hosts, the tentative IP will consider as a unique IP and the target host will configure its IP. Message filtering stage consider as a core stage for DAD-match mechanism.

4.4 Workflow of DAD-match Mechanism

When there is a host recently joined the IPv6 link-local network or an existing host placed on the same link intents to generate a new IP address for its own purpose. First of all, it must make sure that no other node on the same link uses that address already. This is executed by performing DAD process through multicast a number of Neighbor Solicitation (NS) messages on the link. Before sending any NS message, target host will generate an IP address as a tentative IP through using privacy extension method. Then, generate a random word $(w)_S$ and encrypt this random word $E(w)_S$ via using the first 40 bit of tentative IP portion as a secret key for encryption and decryption later during the verification process.

Subsequently, the random word $(w)_S$ and the encryption word $E(w)_S$ will insert into DAD*match* tag, then, DAD*match* must append to NS-match message. The target host should send the NS-match message to Solicited -node multicast group (SNMA) FF02::1:FFSS:SSSS based on the last 24 bit of the tentative IP address, Fig. 2 shows NS-match/NA-match message format.

Part	Description	Value
Ethernet Header	Dest MAC	33:33:FF:SS:SS:SS
	Scr MAC	Sender/Receiver MAC
	Type	0x0806
IPv6 Header	Scr IP	:: (unspecified address)
	Dest IP	FF02:1:FFSS:SSSS
	Next header	0x3a
ICMPv6	Type	200 for NS-match 201 for NA-match
	DAD-match	*E(w)* | *(w)*

Fig. 2. NS-match/NA-match message format

All the existing hosts on the same link that joined the same address of (SNMA) will receive the NS-match message. Firstly, receiving hosts will use the first 40 bits of its IP portion attempting to decrypt the encryption word, then compare the decryption word $D\left(E(w)_S\right)_R$ with the random word $(w)_S$:

$$D\left(E(w)_S, IP_R\right)_R = (w)_S \tag{1}$$

If the result matched then duplicate IP address occurred and the receiving host must send NA-match message to inform the target host about the duplication IP. Before sending the NA-match message, should generate a random word $(w)_R$ and encrypt this word through the first 40 bits of its IP portion $E(w)_R$ and insert them to *DADmatch* tag. Thereafter, append the tag to NA-match message and send it to SNMA address and the source address will place empty. Hiding the IP during the whole DAD process in NS and NA will completely prevent any attacker to catch the IP and perform any kind of attacks i.e. DoS or Man-in-the-Middle.

Recent hosts that join the SNMA address are going to receive the NA-match message, as the result, the target host will verify the NA-match message while the other hosts should discard it. For NA-match message verification, first step must contain the *DADmatch* tag otherwise the target host will discard it and consider the message came from illegitimate host. In case *DADmatch* exist, the verification process will proceed through decrypt the encryption word, then compare the decryption word $D\left(E(w)_R\right)_S$ with the random word $(w)_R$:

$$D\left(E(w)_R, IP_S\right)_S = (w)_R \tag{2}$$

If the result matched then duplicate IP address occurred and NA-match message comes from legitimate host, and the target host must regenerate a new tentative IP and redo DAD-match process. However, if the result was not match, target host must discard the NA-match message. Furthermore, DAD-match mechanism will consider the tentative IP unique if no NA-match message send back within 3 s, and the target host can use it as a permanent address.

In this way, an efficient DAD process can be executed in IPv6 link-local network and fully preventing for any DoS attack on DAD procedure. Since, the target host can detect the uniqueness of tentative IP address with existing hosts. Thus, target host will be capable to join IPv6 network and share the data with the other neighboring hosts locate on the same link. DAD-match mechanism procedure is illustrated in Fig. 3, when new host referred to (Sender) and existing host referred to (Receiver) perform DAD procedure in IPv6 link-local network.

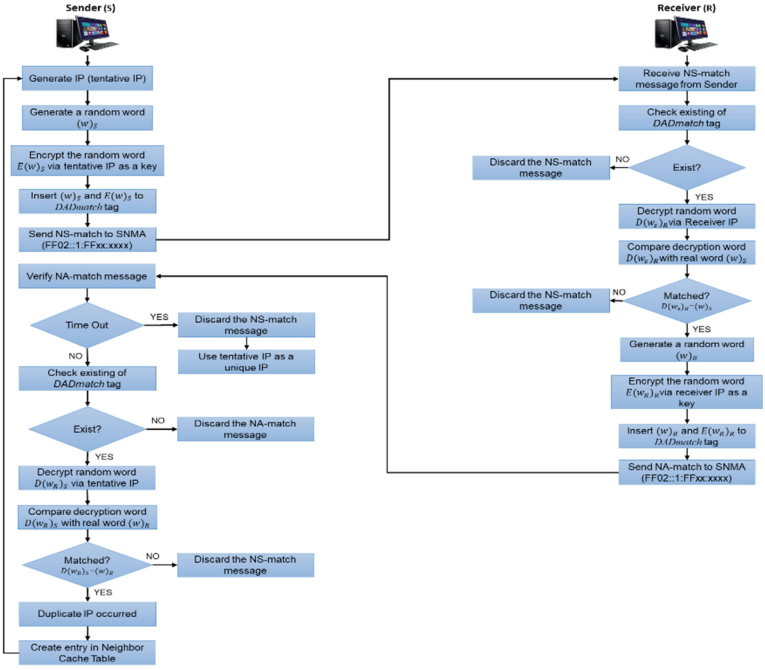


Fig. 3. Workflow of DAD-match mechanism

5 Conclusion

With an increasing number of network nodes and the extensive use of IPv6, DAD attacks pose a serious risk to link-local network security due to growing networking nodes and the wide use of IPv6 in the near future. In standard DAD procedure, which lets all connected existing hosts to identify the new IP address used by the target host, and as a result, malicious nodes can forge responses to launch DoS attack. The aim of DAD-match mechanism is to hide the tentative IP within DAD procedure through utilizes symmetric-key algorithm which uses the same key for both encryption and decryption. The proposed mechanism generate a random word and encryption as well decryption via using the first 40 bits of host IP portion as a key, this procedure will be at the both sides to provide a sufficient verification and prevent the malicious node for forge reply in order to secure IPv6 link-local network from any DoS attacks. For the next step is to choose the encryption algorithm and implement the proposed mechanism and evaluate it with current existing techniques.

References

1. Deering, S.E.: Internet Protocol, Version 6 (IPv6) Specification (1998)
2. IPv6 – Google Statistics (2017). https://www.google.com/intl/en/ipv6/statistics.html. Accessed 13 Apr 2017
3. Stockebrand, B.: IP security (IPsec). IPv6 Practice. A Unixer's Guide to Next Generation Internet, pp. 311–317 (2007)
4. Atay, S., Masera, M.: Challenges for the security analysis of next generation networks. Inf. Secur. Tech. Rep. **16**(1), 3–11 (2011)
5. Huston, G.: A rough guide to address exhaustion. Internet Protoc. J. **14**(1), 2–11 (2011)
6. Arkko, J., Aura, T., Kempf, J., Mäntylä, V.-M., Nikander, P., Roe, M.: Securing IPv6 neighbor and router discovery. In: Proceedings of the 1st ACM Workshop on Wireless Security, pp. 77–86 (2002)
7. Narten, T., Simpson, W.A., Nordmark, E., Soliman, H.: Neighbor Discovery for IP Version 6 (IPv6) (2007)
8. Narten, T., Draves, R., Krishnan, S.: Privacy extensions for stateless address autoconfiguration in IPv6 (2007)
9. Elejla, O.E., Anbar, M., Belaton, B.: ICMPv6-based DoS and DDoS attacks and defense mechanisms: review. IETE Tech. Rev. **4602**, 1–18 (2016)
10. Elejla, O.E., Belaton, B., Anbar, M., Alnajjar, A.: Intrusion detection systems of ICMPv6-based DDoS attacks. Neural Comput. Appl. 1–12 (2016)
11. Narten, T., Thomson, S., Jinmei, T.: IPv6 stateless address autoconfiguration (2007)
12. Tayal, P.: IPV6 SLAAC related security issues and removal of those security issues. Int. J. Eng. Comput. Sci. **3**(9), 4 (2014)
13. Thomson, S: IPv6 Stateless Address Autoconfiguration (1998)
14. Arkko, J.: Secure Neighbor Discovery (SEND), pp. 1–56 (2005)
15. AlSa'deh, A., Meinel, C.: Secure neighbor discovery: review, challenges, perspectives, and recommendations. IEEE Secur. Priv. **10**(4), 26–34 (2012)
16. Supriyanto, Hasbullah, I.H., Murugesan, R.K., Ramadass, S.: Survey of internet protocol version 6 link local communication security vulnerability and mitigation methods. IETE Tech. Rev. **30**(1), 64–71 (2013)
17. Elejla, O.E., Anbar, M., Belaton, B.: ICMPv6-based Dos and DDOS attacks and defense mechanisms: review. IETE Tech. Rev. **34**, 1–18 (2016)
18. Caicedo, C.E., Joshi, J.B.D., Tuladhar, S.R.: IPv6 security challenges. Comput. (Long. Beach. Calif) **42**(2), 36–42 (2009)
19. Arkko, J., Kempf, J., Zill, B., Nikander, P.: Secure Neighbor Discovery (SEND) (2005)
20. Smart, N.P.: Public key encryption and signature algorithms. In: Cryptography Made Simple, pp. 313–347. Springer, Cham (2016)
21. Kukec, A., Bagnulo, M., Mikuc, M.: SEND-based source address validation for IPv6. In: 10th International Conference on Telecommunications 2009, ConTEL 2009, pp. 199–204 (2009)
22. Kukec, A., Krishnan, S., Jiang, S.: The Secure Neighbor Discovery (SEND) Hash Threat Analysis (2011)
23. Gagneja, K., Singh, J.: Survey and analysis of security issues on RSA algorithm for digital video data. J. Discret. Math. Sci. Cryptogr. **19**(1), 39–55 (2016)
24. Praptodiyono, S., Hasbullah, I.H., Kadhum, M.M., Wey, C.Y., Murugesan, R.K., Osman, A.: Securing duplicate address detection on IPv6 using distributed trust mechanism. Int. J. Simul. Syst. Sci. Technol. **17**(26) (2016)
25. Guo, J., Peyrin, T., Poschmann, A.: The PHOTON family of lightweight hash functions. In: Annual Cryptology Conference, pp. 222–239 (2011)

26. Aumasson, J.-P., Henzen, L., Meier, W., Naya-Plasencia, M.: Quark: a lightweight hash. In: International Workshop on Cryptographic Hardware and Embedded Systems, pp. 1–15 (2010)
27. Wang, S., Liu, G.: File encryption and decryption system based on RSA algorithm. In: 2011 International Conference on Computational and Information Sciences (ICCIS), pp. 797–800 (2011)
28. Turner, S., Chen, L.: Updated security considerations for the MD5 message-digest and the HMAC-MD5 algorithms (2011)
29. Rehman, S.U., Manickam, S.: Improved mechanism to prevent denial of service attack in IPv6 duplicate address detection process. Int. J. Adv. Comput. Sci. Appl. **8**(2), 63–70 (2017)
30. Ahmed, A.S., Hassan, R., Othman, N.E.: IPv6 neighbor discovery protocol specifications, threats and countermeasures: a survey. IEEE Access **5**, 18187–18210 (2017)
31. Fenner, B.: Experimental Values in IPv4, IPv6, ICMPv4, ICMPv6, UDP, and TCP Headers (2006)

Image-Based Technique for Turbulent Flow Segmentation

A. B. Osman[1(✉)], Mark Ovinis[1(✉)], I. Faye[2], and F. M. Hashim[1]

[1] Mechanical Engineering Department, Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia
anotood@yahoo.com, mark_ovinis@utp.edu.my
[2] Fundamental and Applied Sciences Department, Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia

Abstract. Turbulent flow segmentation from image data is a challenging problem. This is due to the un-defined edge and the complex flow nature of turbulence. In this paper, an image-based technique is proposed for turbulent flow segmentation from image. The proposed technique segments the flow region based on enhancing the input image intensity at flow edges and by defining a thresholding value to differentiate between flow region and image background. To test the image-based segmentation technique, a turbulent buoyant jet was experimentally simulated at different nozzle flow rates which have a Reynolds numbers of 960, 1560, and 3210. Then, a video camera was used to record the jet flow data. Then, the image-based technique was applied to segment the flow region and estimate the jet penetration area. As a result, the turbulent flow region was segmented well for all cases of nozzle flow rates. Moreover, application of the image-based technique for jet penetration estimation showed a good agreement with the previous work, in which the jet propagated linearly over time.

Keywords: Turbulent jet · Penetration area · Turbulent flow

1 Introduction

A turbulent buoyant jet is referred to the flow type in which two fluids with different densities are mixed. An example of turbulent buoyant jet flow is underwater oil spill in which the spilled oil injects into seawater. As well as underwater black smoker in which hydrocarbon leaks into seawater. Study of such flow type is an important task to understand its flow dynamics. This can helps to quantify the amount of spilled oil/hydrocarbons into the environment. Then, appropriate response such as clean-up process can be taken to minimize their negative effects on the environment. Moreover, the importance of studying of jet flow arises in numerous fields either in industrial applications or natural flows such as, waste disposal systems, and smoke plumes from a chimney. Several researchers studied the jet flow behavior in order to build a relationship between the penetration of turbulent buoyant jet and its initial conditions including inlet diameter, flow rate and nozzle geometry. Most of past researchers [1] conducted to investigate modeling of jet penetration by considering the stream-wise and cross penetration separately.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 119–129, 2018.
https://doi.org/10.1007/978-981-10-8276-4_12

Philippe et al. [2] was experimentally investigated the jet penetration into liquid, while Ivan [3] proposed a mathematical model for jet penetration. Faris [4] investigated the entertainment of an asymmetrical turbulent jet flow. Although the extensively investigated the turbulent jet flow behaviour. In their works, the penetration models were developed for single direction including axial and radial directions separately. However, measuring the jet penetration based on single direction is uncertain. This is due to the flow variability at the jet boundary and complicated nature of flow in two dimension space. It hypothesized that tracking the jet penetration based on the overall space would represent the actual penetration as compared to single direction.

On the other hand, several researchers have developed image segmentation algorithms for various applications such as cancer detection [5], color image segmentation using k-mean classifier [6], and image segmentation for computer vision applications [7]. By segmenting the jet flow region and tracking its propagation over the time, one can develop an experimental model to predict the jet entrainment as well as estimating the inlet flow rate. However, these methods are limited to use for rigid object segmentation and have not been applied for turbulent flow segmentation. This is because the boundary for turbulent is not well defined.

This work proposed an image-based technique for turbulent flow boundary segmentation. The proposed method is based on separating the jet flow region from image background by defining a threshold value. First, turbulent buoyant jet was experimentally simulated at various nozzle flow rates. Second, video camera was used to track the jet flow propagation. Finally, the jet penetration area, as well as the jet penetration in both stream-wise and radial direction was investigated. The rest of the paper is organized as, Sect. 2, describes the general methodology including the experimental work and the proposed image-based approach, Sect. 3 results and discussion followed the conclusion.

2 Method

This section describes the overall methodology including description of the proposed image-based method used for flow region segmentation as well as the experimental works done to simulate the turbulent buoyant jet.

2.1 Image-Based Segmentation

An image-based technique is proposed to segment the turbulent buoyant jet region from the image sequence. Figure 1 shows the general flow chart for jet flow segmentation and area estimation. First, some preprocessing steps were conducted to the original video sequence such as converting video to image sequence in which a method proposed in [8] was used. The jet flow region is cropped in order to remove unnecessary image boundary. For easy jet flow region segmentation, the input images were tested to select the best image channels (i.e. R, G, and B channels) as well as the gray-level image was tested. This is by presenting and comparing the histogram of these channels. We found that the gray-level image has a good quality that can be used segmentation process. Moreover, two problems were observed in the original video includes light variation

over time and animated background due to the movement of water in the main tank during video recording. Then, the contrast of all image sequence was adjusted and the image background was subtracted.

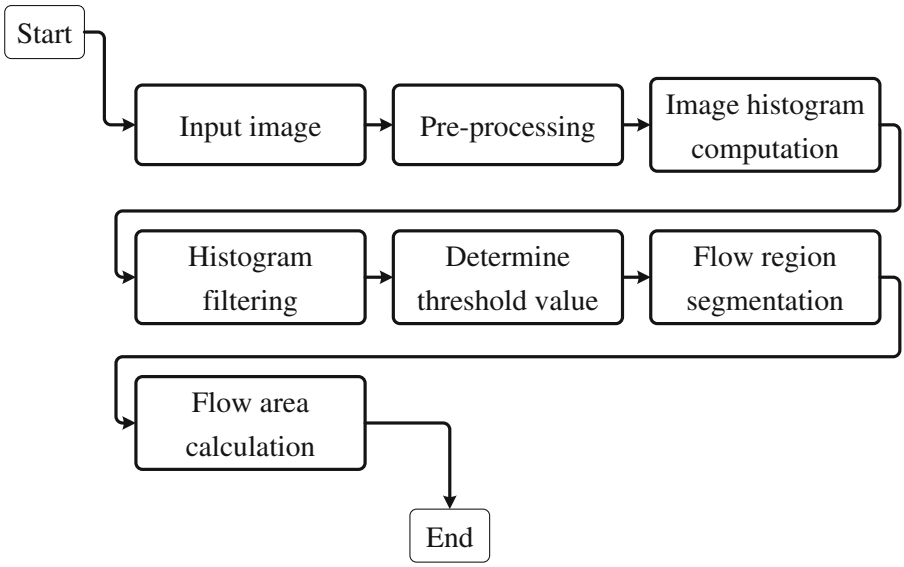


Fig. 1. Image-based technique used for turbulent flow region segmentation.

Then, the jet flow region is segmented from the processed image sequence by testing the image histogram. This approach has been proposed by Otsu [9]. Otsu method segments image based on differentiating between the jet flow region (foreground) and image background. By calculating the optimum threshold value, the two classes can easy differentiate. Figure 2 illustrates a description of flow region segmentation using image histogram. The histogram here shows, the position of the optimum threshold value that used to differentiate between the image background and jet flow region (i.e. foreground).

A Matlab code was developed for processing the image sequence in order to segment the jet region and find jet penetration by using these steps:

1. Convert the video data to image sequence.
2. Save the image sequence in 3D-matrix for easy processing.
3. Enhance the quality of images by adjusting their contrast, in which the quality of the image tested using image histogram.
4. Rectangular cropping of images to include jet penetration area.
5. Jet flow boundary detection and segmentation using Otsu's method.
6. Estimate a linear penetration of the jet from the segmented images in flow direction.
7. Estimate the jet flow penetration area from binary images.

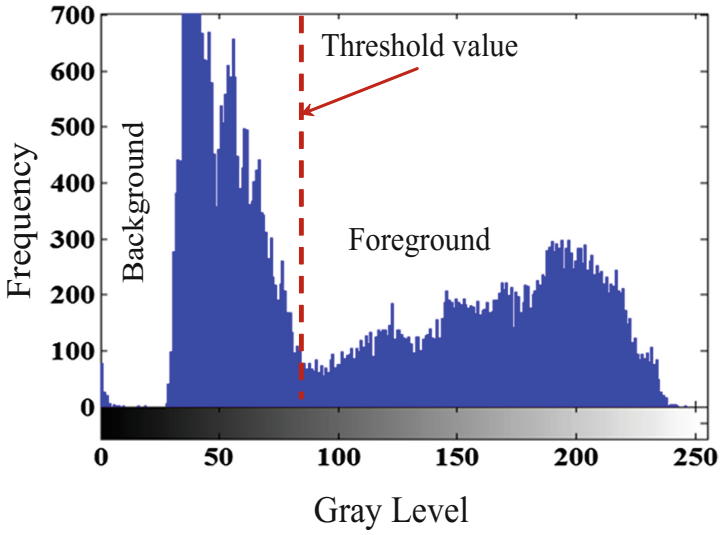


Fig. 2. Jet flow segmentation using Otsu's thresholding algorithm

2.2 Experimental Facility

Figure 3 shows the physical components of experimental set-up used to simulate turbulent buoyant jet flow. Two fluids were used to simulate the jet flow water and water mixed by 5% salt to change the density with adding graphite for visualization as introduced by Crone et al. [10]. This set-up consists of, supply tank (80 gallons) which filled by the mixed fluid. By using a submersible pump, the mixed fluid delivered to the upper tank in which the fluid level is constant during each experiments runs, and any extra fluid will return back to the supply tank. By opening the flow control, the mixed fluid mix with the tap water of main tank (has a size of $900 \times 900 \times 2000$ mm) to simulate the required turbulent buoyant jet.

Jet nozzle flow rate was measured by calibrating the opening of control valve. This by measuring the time required to pass one liter of the mixed fluid through the control valve. Three cases of nozzle flow rates have Reynolds numbers of 960, 1560, and 3120 were considered. Then, the initial nozzle parameters were calculated such as flow rate, Reynold's number, the average velocity at the nozzle exit, with calculating the measurement uncertainty. Table 1 summarized the measured time, calculated flow rates, exist velocity and the corresponding Reynold's numbers.

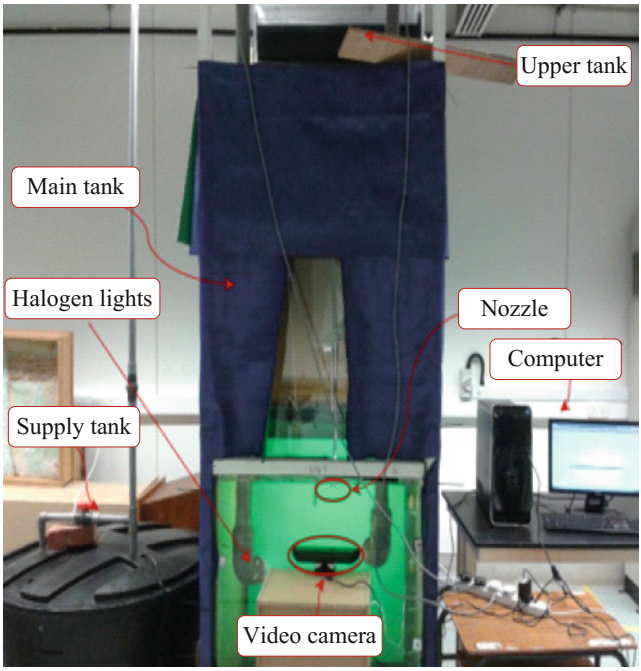


Fig. 3. Experimental rig set-up for simulating turbulent jet flow

Then, the jet propagation for each case of nozzle flow rates was recorded by using a video camera. A 15 s video was recorded with a frame rate of 18 frames per second to get a total of 270 frames. To improve the quality of the recorded video, the flow region was illuminated with two halogen lights, while black background was used.

Table 1. Initial parameters estimation (*Re, U, Q*)

Condition	Time measured (sec)	Flow rate (Liter/sec)	Exit velocity U (m/sec)	Average of U (m/sec)	Re
1	88	0.0114	0.0113	0.064	960
	90	0.0111			
	87	0.0115			
2	53	0.0189	0.0184	0.104	1560
	54	0.0185			
	56	0.0179			
3	26	0.0385	0.0385	0.208	3120

3 Results and Discussion

This section discusses the results of image-based technique used to analysis the penetration of turbulent buoyant jet.

3.1 Flow Visualization

Figure 4 shows typical consecutive images for the simulated turbulent jet flow which extracted from the recorded video at various cases of nozzle flow rate. Different gray levels was observed by changing the nozzle flow rate, in which darker images were observed in case of high flow rate, while gray level increases with reducing the flow rates. This due to the small amount of graphite used for low flow rate cases as compared to the large amount in case of high flow rate. Uses of light with constant intensity for all cases of nozzle flow rate lead to reduce the opacity of the jet flow with reducing the flow rate. However, this observation was not affected the accuracy of proposed technique in flow region segmentation since the flow region can be differentiated from image background for all cases. Moreover, one can investigate the effect of using different light intensity which is not considered in this study.

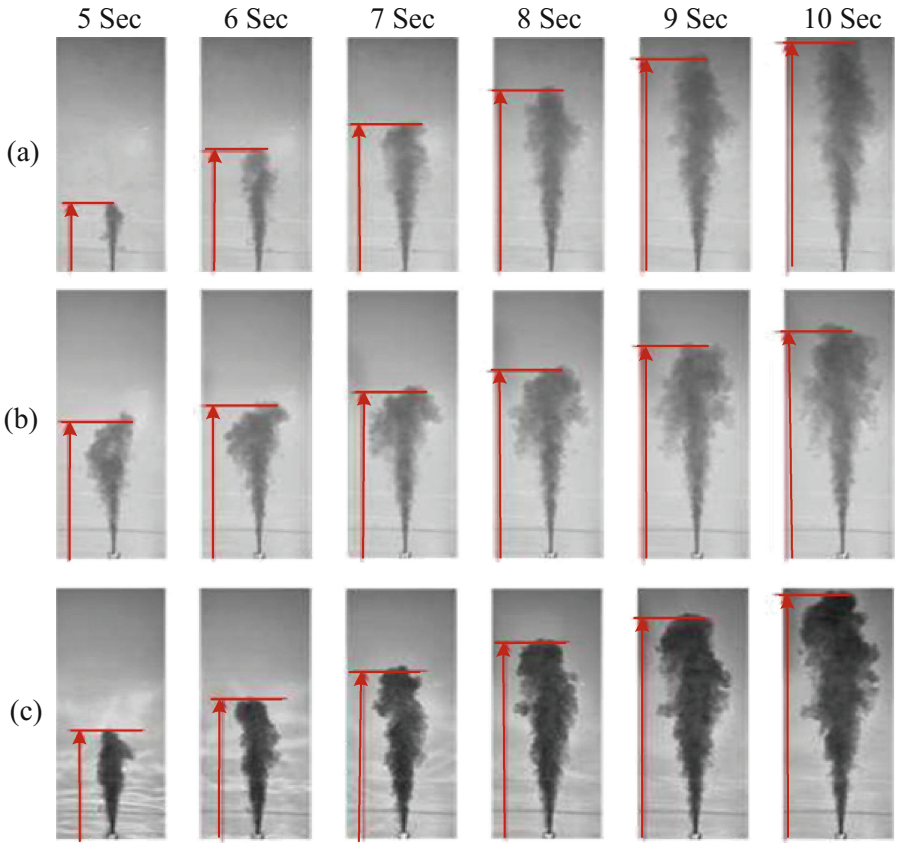


Fig. 4. Visualization of turbulent buoyant jet flow for different nozzle flow rate cases includes: (a) Re = 960, (b) Re = 1560, (c) Re = 3120.

As observed the flow structure for higher flow rate penetrated smoothly with observation of larger coherent structures at jet the boundary (see Fig. 3(a)). By decreasing the nozzle flow rate, more vortices was noted at the far-field region and the flow will be very chaotic. This is true from the fundamental bases of the turbulent jet theory due to the mixing process of the two fluids. Large scale structures usually can be observed in the case of low flow rate as compared to higher flow rate when the jet flow simulated for the same nozzle geometry and diameter.

3.2 Flow Penetration Segmentation

Jet flow penetration was measured by direct measuring the stream-wise penetration distance from nozzle up to the maximum position of jet penetration as done by Sangras et al. [1]. As well as the jet penetration area this was measured by defining and segmenting the jet flow region in the image sequence. In the following sub-sections, we discussed the results of both penetration measures.

3.3 Stream-Wise Penetration

Linear stream-wise penetration of turbulent jet was measured manually by measuring the distance from nozzle to the maximum point of flow penetration in flow direction as shown by a red color arrow in Fig. 4. Figure 5 shows the estimated jet penetration in stream-wise direction for the three cases of nozzle flow rates. Therefore, strong linear relationships were obtained for the jet flow penetration distance in stream-wise direction over time. These linear models were obtained with high-value R-square as a measure of linearity.

3.4 Estimation of Jet Area Propagation

Figure 6 shows the segmented jet boundary using the image-based technique overlapped over the original image. Then, the jet area was estimated using the method proposed by [11]. In their work, a flame surface area was estimated using the converted binary images. The image was divided into horizontal segments (x), and then the total area calculated by sum-up the segmented areas using the following equation:

$$A_f = \sum_{j=1}^{h^r} A\left(h_j\right) \tag{1}$$

where x is the total height of the jet flow region, *'a'* is the area of the small segment, j is the number of segment ranged from 1 to the total height of jet h_T. Table 2 summarizes the area calculated for cases of Reynolds number over time (from 0–3.5 s).

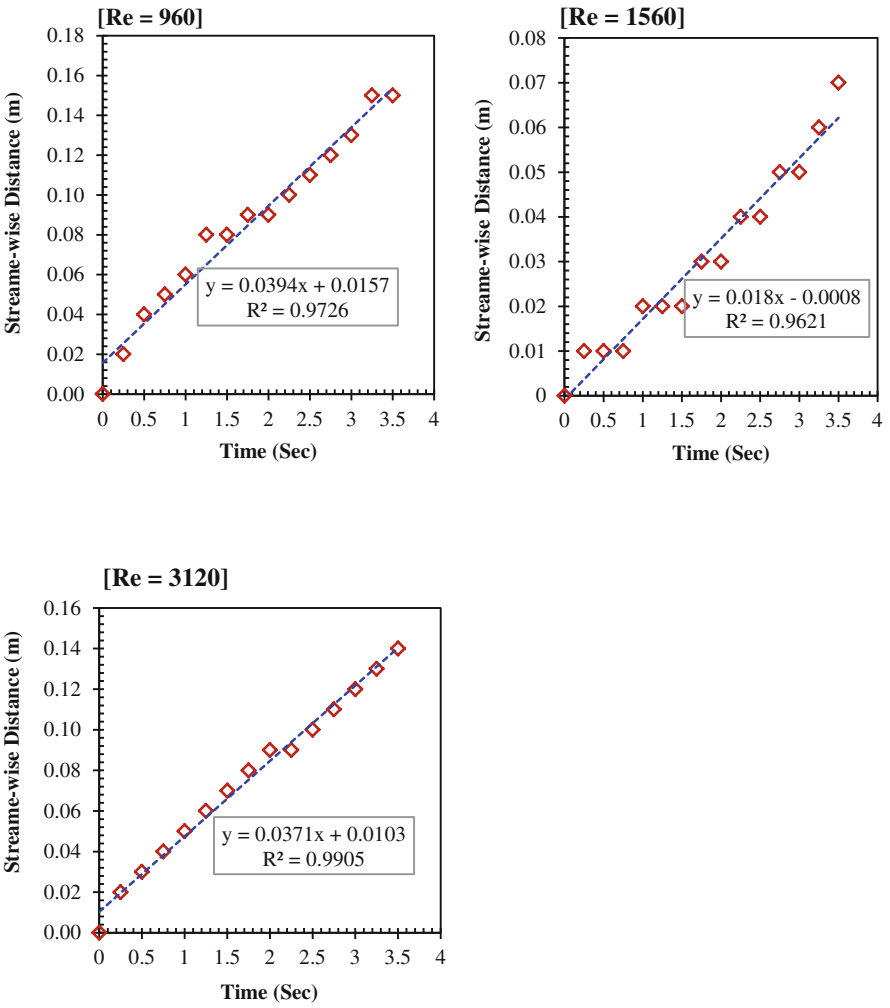


Fig. 5. Stream-wise penetration of turbulent jet flow for three cases of nozzle flow rates includes, (a) Re = 960, (b) Re = 1560 and (c) Re = 3120.

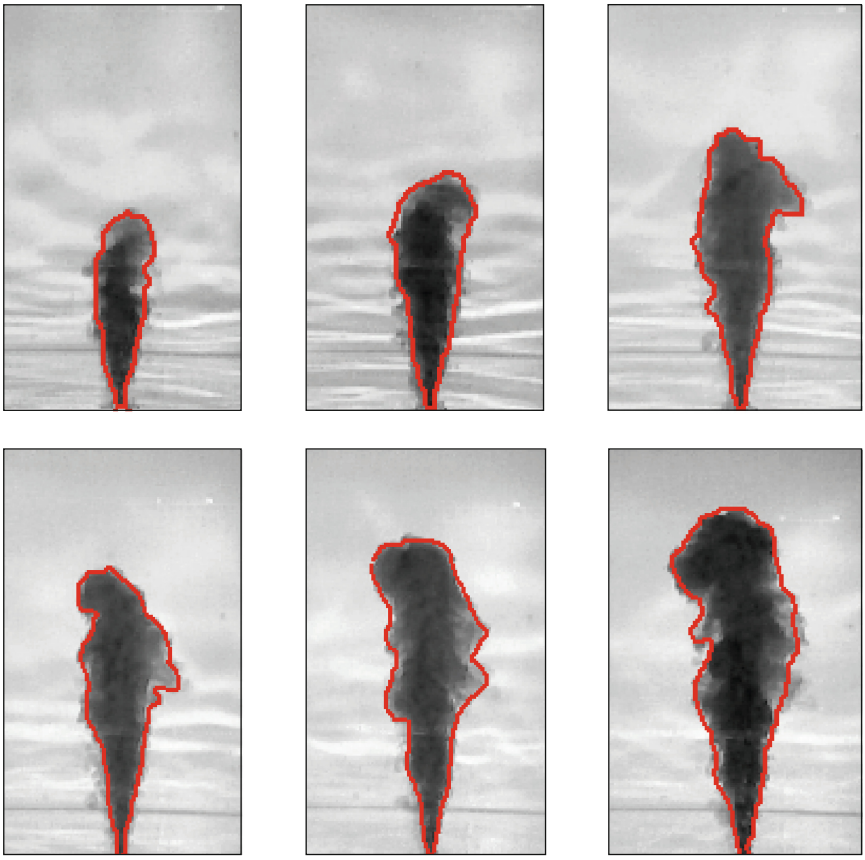


Fig. 6. Segmentation of turbulent buoyant jet penetration area

Figure 7 shows the estimated jet penetration area for 15 consecutive images for the three cases of nozzle flow rate. Strong linear relationships were obtained between the value of jet flow penetration area and the propagation time. Therefore, by knowing the flow area over time we can infer the total volumetric flow rate at the nozzle.

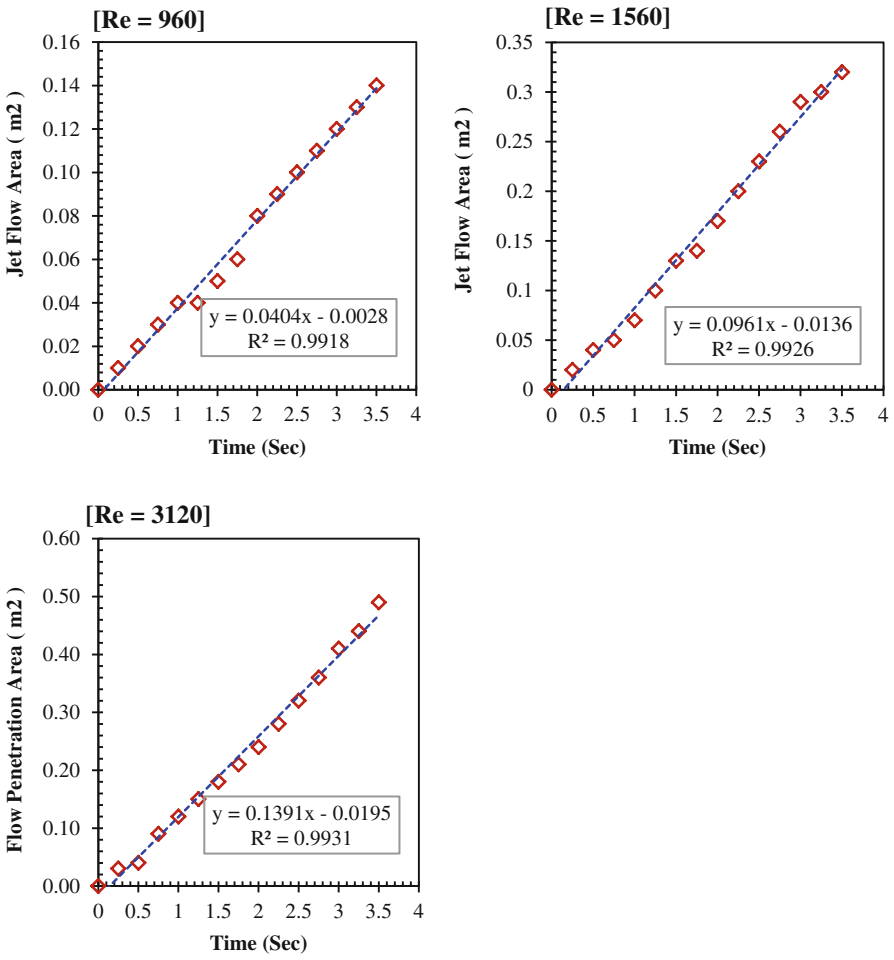


Fig. 7. Turbulent jet penetration area estimated overtime for various cases of reynolds numbers.

4 Conclusion

In this work, image-based technique is proposed for segmentation of turbulent flow from image sequence. An experimental work is conducted to study the turbulent buoyant jet flow penetration in which three nozzle flow rates were considered which has Reynolds number (960, 1560, and 3120). Therefore, a linear penetration was observed when the technique used for jet penetration area estimation. This technique has advantage of determining the jet penetration area directly from image sequence, while the previous work such as the work of Sangras et al. [1] focused on studying the jet penetration in axial and radial direction individually. Due to the error that expected with using single direction the area-based method is more appropriate to describe the propagation of turbulent jet. Moreover, the approach can be extended in order to estimate the volumetric

flow rate of the jet nozzle. A detailed investigation will be done using various nozzle diameters and shapes include both round and plane nozzle. Then, a volumetric approach will be proposed for directly estimating the volumetric flow rate of the turbulent jet.

Acknowledgments. The authors would like to express their appreciation to Universiti Teknologi PETRONAS for supporting this work under YUTP 0153AA-E85.

References

1. Sangras, R., et al.: Buoyant turbulent jets and plumes: III. Round turbulent non-buoyant starting jets and puffs and buoyant starting plumes and thermals (1999)
2. Philippe, P., et al.: Penetration of a negatively buoyant jet in a miscible liquid. Phys. Fluids **5**, 053601 (2005)
3. Kazachkov, I.: The mathematical models for penetration of a liquid jets into a pool. WSEAS Trans. Fluid Mech. **6**, 1–22 (2011)
4. Faris, G.N.: Some Entrainment Properties of a Turbulent Axi-Symmetric Jet. Mississippi state univ Mississippi state dept of aerophysics (1963)
5. Ruppert-Felsot, J., Farge, M., Petitjeans, P.: Wavelet tools to study intermittency: application to vortex bursting. J. Fluid Mech. **636**, 427–453 (2009)
6. Samma, A.S., Salam, R.A.: Adaptation of k-means algorithm for image segmentation. World Acad. Sci. Eng. Technol. **1**(50), 58–62 (2009)
7. Liu, D., Soran, B., Petrie, G., Hapiro, L.: A review of computer vision segmentation algorithms. Lecture Notes, vol. 53 (2012)
8. (2017). https://www.mathworks.com/help/matlab/examples/convert-between-image-sequences-and-video.html
9. Otsu, N.: A threshold selection method from gray-level histograms. Automatic **11**(285–296), 23–27 (1975)
10. Crone, T.J., McDuff, R.E., Wilcock, W.S.: Optical plume velocimetry: A new flow measurement technique for use in seafloor hydrothermal systems. Exper. Fluid **45**(5), 899–915 (2008)
11. Hu, L., et al.: Flame size and volumetric heat release rate of turbulent buoyant jet diffusion flames in normal-and a sub-atmospheric pressure. Fuel **150**, 278–287 (2015)

Optimization of Remaining Energy and Error Rates for Wireless Sensor Network

Samirah Razali, Kamaruddin Mamat(✉), and Nor Shahniza Kamal Bashah

Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia
samirah.razali@gmail.com, {kamar,shahniza}@tmsk.uitm.edu.my

Abstract. Wireless Sensor Network has become one of the crucial and vital technology in environmental monitoring and tracking. The advancement of such technology had evolved WSN to transmit a heavy data as well as handled the high number of traffics which had increased the demand for studies and further research on the aspects of error control protocols. Based on the previous work, the existing error control techniques were not able to combat the issue of interference levels and excessive overhead properly. Thus, the problem regarding unnecessary overhead, interferences and error rates in changing conditions of the WSN becomes our motivation to propose a new method for extending the capability of HARQ error control algorithm in CDMA WSN. This paper evaluates the proposed HARQ based Multiple Error Correction in terms of Bit Length and Node Densities. This paper has demonstrated that the proposed methods show promising results in optimizing the energy consumption and error rates.

Keywords: WSN · CDMA · Error control protocol · BER · HARQ
Energy efficiency · Remaining energy

1 Introduction

Many researchers over a past few years had been studying the method to solely enhance the energy efficiency of Wireless Sensor Network (WSN). The factors such as error rates, overhead and interference present needed to be taken into consideration and cannot be cast aside. The Coded Division Multiple Access (CDMA) had been deployed in WSN as the technology can improve many attributes such as latency, energy efficiency and as the mechanism of fault-tolerance rather than conventional WSN [1]. However, Multiple Access Interference (MAI) problem in CDMA become significant when the power of the user increases [2]. The SNR of one's network changes as well due to certain factors such as attenuation and the present of interferers. The problem of extra overhead during congestion and might be able to be mitigated in considering the link quality into the solution. HARQ is said to be able to outperform the ability of ARQ or FEC in terms of optimizing the energy consumption in CDMA WSN and also help to lower the Bit Error Rates (BER). However, the implementation of existing HARQ in CDMA WSN foster additional overhead if the improper usage of Error Correcting Codes (ECC) were not studied and assigned carefully. As for the problems stated above, we proposed the extension to the HARQ error control scheme based on the network condition estimation

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 130–140, 2018.
https://doi.org/10.1007/978-981-10-8276-4_13

to optimize the energy consumption as well as error rates. This paper demonstrated the proposed Multiple Error Correction (MEC) algorithm that acts as extensions to the existing HARQ scheme which aims to reduce the energy consumption of the network without neglecting the effect of error rates, overhead that rises due to the computation of error correcting codes and interferences as well. This paper also evaluates the performance of CDMA WSN network that implements the proposed MEC algorithm. For this paper, we test the MEC implementation in terms of Remaining Energy and BER. The error correction schemes will be changed according to the change in network condition. The network condition will be estimated using Kalman Filter and the estimation value is in the form of SNR will be used to decide the error correction schemes that will be initiated by the sender as well as error correcting capabilities. The experiment conducted will be evaluated based on different Bits Length and Number of Nodes. The rest of this paper is organized as follows: In Sect. 2, the related works are stated and analyzed. The methods involving MEC design, simulation and measurement models are illustrated in Sect. 3. In Sect. 4, the results and findings from the simulation were presented. The discussion is explained in Sect. 5. Lastly, In Sect. 6, conclusion and future work are stated out.

2 Related Works

Early studies of this research have been demonstrated in [3]. In this paper, the existing HARQ algorithm was proposed to be modified according to the channel state and RSSI. This paper follows the continuation of [3] with different sets of testing and results of different sets of scenarios and error control environment. There are also a few previous studies of HARQ in CDMA WSN by [4, 5]. In these papers, the author had implement HARQ with BCH in multi-hop communication. The authors also test the HARQ-BCH with different node densities and message length. Not to mention, they also evaluate the energy efficiency and include the correlations among interferers. Based on [6], these authors had stated that MAI that presents in the CDMA cannot be controlled by transmitting power due to the absence of central base station. However, on some other paper that implements Transmission Power Control (TPC), by controlling the TPC, the lifetime of the network could be boosted. Not to mention, the paper by [4] also stated that the transmit power can aim to minimize the interference.

Incorporating transmit power in the process of HARQ can be advantageous in lowering the effect of MAI. However, the improper assignment of transmitting power in a network such as CDMA can further increase the MAI or can just bring down the BER for the whole network. The latest work on TPC had implemented the estimation using Kalman Filter (KF). The power will be adjusted according to environment condition that was predicted by KF. From their work, they also had proof that the KF can track the channel state properly [7] besides, minimizes a quadratic function of estimation error for a linear dynamic system with white measurement and disturbance noise [8, 9]. The usage of KF in estimating channel conditions can be significant as SNR of the network can change over time due to various factors such as the attenuation which causes by refractions of wireless signal or collisions during transmissions.

3 Methods

3.1 Multiple Error Correction (MEC) Scheme

Multiple Error Correction (MEC) Algorithm was based on the existing process of HARQ. The MEC algorithm was divided into three phases. We merged the algorithm that includes phase 1 estimation, phase 2 error correction and phase 3 update module into one algorithm of MEC to provide flexibility and reduce complexity. Phase 1 includes initialization and estimation in which the standardized Kalman Filter was used to measure the network condition. From the estimation value, we divide the error correction schemes into five different categories of error correction mode. The Kalman Filter equation that was used in this research was shown as follows [10, 11]:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \tag{1}$$

$$z_k = Hx_k + v_k \tag{2}$$

Where x_k is denoted as the current estimator, Kalman Filter consists of two phases which are the time update (prediction) and measurement update (correction). The time update projects the current state estimate ahead of time. The measurement update adjusts the projected estimate by an actual measurement at that time [12]. The phase 2 of MEC algorithm started when the estimated SNR was obtained from the estimation process in Phase 1. During this phase, the sender will decide the error correction mode that will be initiated before the encoding process occurred. If the network is not congested, the sender node will recognize the error correction mode as Lowest and instead of using complex error correcting codes such as BCH, the sender will only encode the data using CRC-4 and prepare for retransmissions. However, if the network is recognized as highly congested or when the SNR dropped to around 0 to 23, the sender will initiate the Highest Error Correction Mode and will encode the data with the high error correcting capability of BCH. As WSN suffered from the resource-constraint issue, we identified the overhead factor as one of the important aspects to be lowered down in between transmissions. The use of simple CRC-4 and low error correcting capability such that in the range of t = 3 to t = 10 was to reduce this complexity of codes and computational overhead that might rises due to encoding and decoding of error-correcting codes. The transmit power will also be considered and changes according to the SNR. We had classified different error correction schemes with different error correcting capabilities and different levels of transmit power corresponding to the changes in network condition as the network condition gradually changes over time due to the refraction and reflection of the signal, fading and attenuation. We follow the indication of SNR that determines the network condition or link quality of wireless technology as according to [13]. However, we had adjusted the SNR values to suit our WSN environment models. Phase 3 involves update module in which the KF is being used at the receiver to estimate the SNR for upcoming transmission from the sender. The SNR value estimated information will be appended to the acknowledgment message (ACK) as a feedback mechanism to the sender. The information appended have a size of 3 bits which we considered to have insignificant effects on the performance as the bits added is very small. We merged the algorithm that

includes phase 1, phase 2 and phase 3 into one algorithm of MEC to provide flexibility and reduce complexity. As WSN suffered from the resource-constraint issue, we identified the overhead factor as one of the important aspects to be lowered down in between transmissions. The use of simple CRC-4 and low error correcting capability such that in the range of t = 3 to t = 10 was to reduce this complexity of codes and computational overhead that might rises due to encoding and decoding of error-correcting codes.

3.2 System Model and Parameters Defined

In this paper, we apply CDMA WSN topology which involves layered node architecture. The uniform node deployment of CDMA WSN follows the architecture from the work of [14]. Figure 1 shows the uniform node deployment in CDMA WSN. The data will be passed from any specific location, for example, the nodes "n5" will be passing the data to "n6" until reaching the sink. The simulation and testing were carried out in MATLAB environment. We have done the initial testing in which initial testing only simulates the network aided by the existing error correction scheme such as error detection using CRC-4 and error correction using common error correcting codes. The scenarios were defined in Table 1 below.

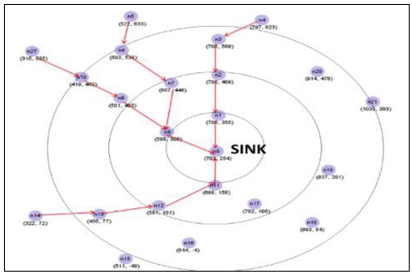


Fig. 1. The topology for uniform node deployment [14, 15].

The scenarios were classified into two groups such that scenario 1 and scenario 2 will test the Uniform Node deployment CDMA WSN while scenario 3 and scenario 4 will be testing the non-uniform topology. For Scenario 1.1, 1.2, 3.1, and 3.2, the performance of CDMA WSN was analyzed with the same value of (n, k) for both BCH codes and RS codes such that BCH (63, 45) RS (63, 45) regardless of error correcting capabilities, t. The bit length and node densities will be changed throughout the experiments. While, for Scenario 2.1, 2.2, 4.1 and 4.2, the performance of CDMA WSN was analyzed with the changing in error correcting capabilities, t such that t = 3, 6 and 10. Table 2 shows the parameters and variables used throughout the experiment. We test the defined parameters to measure the remaining Energy and average BER for the whole network in order to get the effect of the different error correcting codes with its error correcting capability towards the redundancy added to the data. The redundancy added have a very significant effect on the network that causes the rise in overhead due to the error correcting codes and error correcting capability, not to mention, the corruption of bits that cause by interference present, and architecture of the network itself.

Table 1. The scenarios generated for simulation

Scenarios		Parameter	Description
1	1.1	Bit length	Uniform node deployment
	1.2	Node densities	MEC algorithm ECC with the same value of (n, k) such that BCH (63, 45) RS (63, 45) BCH (15, 5) and RS (15, 5)
2	2.1	Bit length	Uniform Node Deployment
	2.2	Node densities	MEC algorithm ECC with different error correcting capabilities, t where t = 3, 6, 10
3	3.1	Bit length	Non-uniform node deployment
	3.2	Node densities	MEC algorithm ECC with the same value of (n, k) such that BCH (63, 45) RS (63, 45) BCH (15, 5) and RS (15, 5)
4	4.1	Bit length	Non-uniform node deployment
	4.2	Node densities	MEC algorithm ECC with different error correcting capabilities, t where t = 3, 6, 10

Table 2. Variables and parameters used throughout the experiments

Variables	Values
Channel model	Multi-Carrier CDMA (MC-CDMA) with Rayleigh Fading
Modulation	BPSK
Noise	AWGN
Path loss exponent	3.5
Error detection	CRC-4
Error correction codes	Reed-Solomon and BCH
Number of nodes	16, 32, 80
Bit length (Bits)	10000, 20000, 30000
Additional bits of information on ACK message	3 bits
Minimum distance between nodes (meter)	5 m
Monitoring area (meter)	1000 m × 1000 m

The CRC-4 will be appended to the input binary data before modulation if the error correcting codes were not implemented, then spreading will occurred and passed the encoded data through the channel. During the despreading at the receiver, the Fast Fourier Transform (FFT) was carried out. The CRC decoder will identify the corrupted bits. While, RS decoder and BCH decoder will decode then correct the errors that occurred in the receiving data. We have conducted the simulation using Matlab under Window 10 of 64-bit environment. The simulations follow the defined scenarios in Table 1. We assume as the SNR gets higher, the network condition is less congested.

The estimation module of MEC algorithm has been explained before will be initiated first before any transmission from sender to receiver. The sender will decide on the correction mode and appended the required information to the generated data. The encoding and spreading process remain to follow the existing HARQ process until the data were passed through the channel. After the data were decoded, the receiver sent the sender with acknowledgment message with the updated KF estimation while checking on the estimation errors. The updated KF estimation and BER information were appended to the acknowledgment message to prepare the sender for next transmissions.

3.3 Measurement Models

We present several formulas that had been used to calculate the performance of the network throughout the experiments. We formulate the BER calculation, expression of Received Signal and Remaining Energy to measure the performance of CDMA WSN and the effectiveness of MEC algorithm when implemented on the CDMA WSN. The formula of BER for BPSK can be calculated using the equation in [10]. Thus, the formula of BER for BPSK in Rayleigh Fading can be derived as shown in Eq. (3) [16] below:

$$\mathrm{BER} = \frac{1}{2}\left(1 - \sqrt{\frac{\frac{E_b}{N_0}}{\frac{E_b}{N_0} + 1}}\right) \tag{3}$$

While, the received signal power, Y follows free space path loss which added the path loss exponent and the distance between transmitter and receiver. Given the received signal, Y in Additive White Gaussian Noise (AWGN) which includes Rayleigh Fading can be formulated as in Eq. (4) [17]:

$$\mathrm{Y} = \mathrm{h}\left(\mathrm{d}^{\frac{-\alpha}{2}}\right) \tag{4}$$

Where α, the path loss exponent is denoted that $2 \leq \alpha \leq 6$ while d is defined as the distance between sender and receiver. Based on [18], each transmitted bit consume 1 unit of energy and received bit consume 0.75 units of energy. The remaining energy for the network considering also the decoding energy of error correcting codes as follows:

$$E_{ECC} = H \times N_p \times \left(N_{bits} + (N_{bits} \times 0.75\right) + E_{DEC} \tag{5}$$

Where H denoted the number of hops, N_p is the number of packets, and N_{bits} is the total number of bits in which include the header and payload [19]. Based on Eq. (6) below, the decoding energy can be calculated as follows [20]:

$$E_{DEC} = \left(2nt + 2t^2\right)\left(E_{addition} + E_{multiplication}\right) \tag{6}$$

Where n is block length for the error correcting codes and t is error correcting capability. While, $E_{addition} + E_{multiplication}$ is the energy consumed in addition and multiplication that occurred in encoding and decoding of error-correcting codes.

4 Results and Analysis

The results of the implementation of MEC algorithm are to analyze the effectiveness of MEC towards reducing the remaining energy as well as BER. The initial testing performed to study the effects of different error correcting capabilities of BCH and RS in MEC towards CDMA WSN performances in terms of the increasing number of nodes and bit length. The results obtained can be the reference to further enhance the MEC algorithm in the aspects of different sensor's environment and WSN architectures. Figure 2 shows the remaining energy between the different error correcting capability of t = 3 and t = 10 and the increasing number of nodes of 16, 32 and 80 nodes. For t = 3, we had used MEC BCH (31, 16) RS (31, 25) while for t = 10, we applied MEC BCH (31, 6) RS (31, 17). From the graph, it was shown that the different error correcting capabilities give the different performance in the changing number of nodes. The highest remaining energy can be seen when the lower error correcting capability, t = 3 was used for the network with 16 nodes. While, for t = 10 for the same network with 16 nodes, the remaining energy degraded slightly. Same goes to the network running 32 nodes for t = 3 and t = 10. However, for the network that has 80 nodes, the remaining energy almost the same for both t = 3 and t = 10. Even the remaining energy of network with 16 nodes is at its highest due to lower congestion and deployment of nodes, the BER was higher than the network deploy the same number of nodes but with different t = 10.

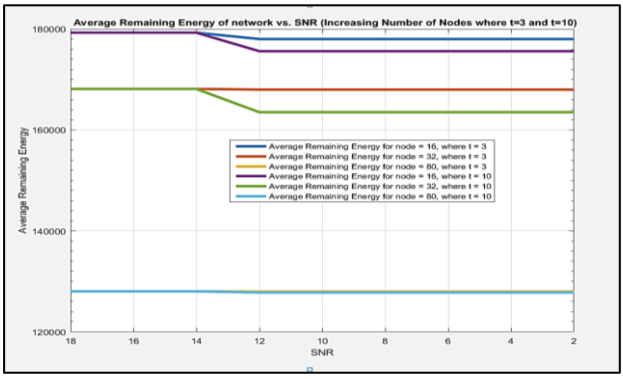


Fig. 2. The graph of comparison of average remaining energy for increasing number of nodes with different error correcting capabilities of t = 3 and t = 10

The graph of Fig. 3 shows the comparison of the average remaining energy between the proposed MEC BCH (63, 45) RS (63, 45) and existing BCH (63, 45). Here, the existing BCH (63, 45) throughout the experiment regardless of network condition or SNR. While, the MEC change the error correcting codes corresponding the changing in

SNR. As the SNR increased, the error correcting codes of BCH change to RS and at the highest SNR, retransmission was initiated instead of using error correcting codes to minimize the possible overhead. Based on the graph, as the SNR is at highest between 14 to 18, the energy also remains at the highest level compared to existing schemes. Figure 4 shows the average BER between proposed MEC and existing BCH. BCH (63, 45) has the error correcting capability of t = 3. For MEC, we also used BCH (63, 45) RS (63, 57) where t = 3. RS (63, 57) has the error correcting capability of t = 3 even it is different in data symbol size, k.

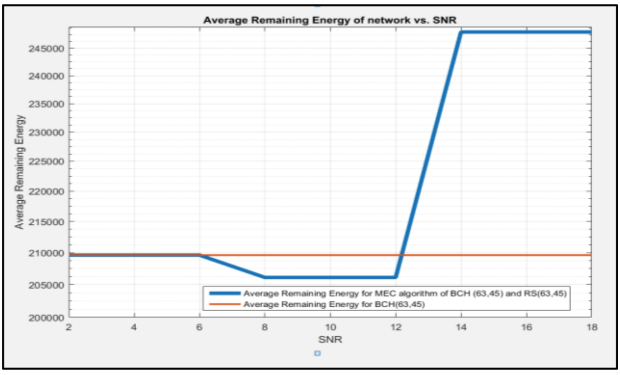


Fig. 3. The graph of comparison of average remaining energy for MEC BCH (63, 45) RS (63, 45) with existing BCH (63, 45)

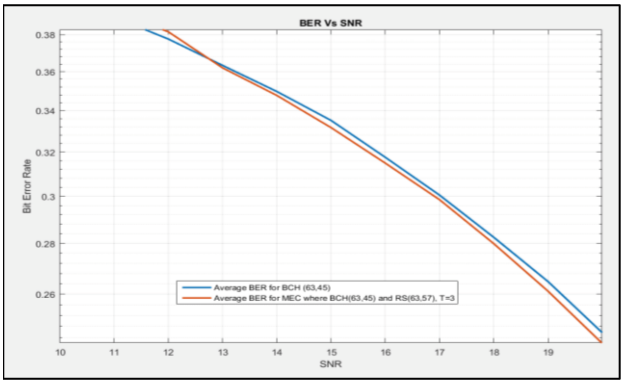


Fig. 4. The graph of comparison of average BER for MEC BCH (63, 45) RS (63, 57) of t = 3 with existing BCH (63, 45) t = 3

Based on the graph in Fig. 4, the used of same error correcting capability in MEC where BCH and RS have the same t = 3 which had reduced the inconsistency in the value of BER and outperform the existing BCH (63, 45). While the graph on Fig. 5 shows the comparison between the average BER between MEC and BCH (63, 18) for t = 10. From this graph, the BER for proposed MEC and existing BCH almost the same

except when SNR going up to 14 to 15, there was a slight increment in BER of MEC. However, the BER of MEC gradually reduced as the SNR increased.

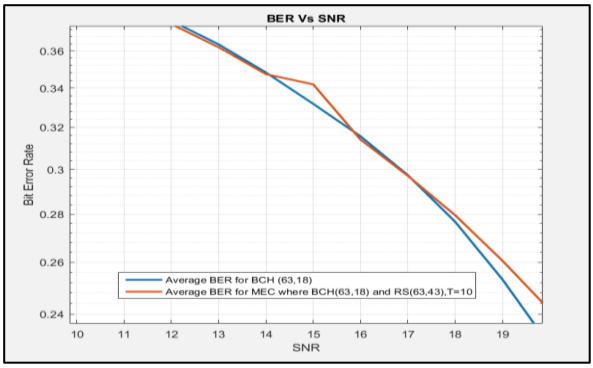


Fig. 5. The graph of comparison of average BER for MEC BCH (63, 18) RS (63, 43) of t = 10 with existing BCH (63, 18) t = 10

5 Discussion

From the graph that compares the existing BCH with the proposed MEC, it was observed that MEC can optimize the BER and remaining energy and slightly outperform the existing BCH. This is because MEC able to change between different error correcting codes such that between BCH and RS and also change the error correcting capability, t to match the network condition. As for this matter, we proposed MEC to reduce the unnecessary overhead that was present from the unnecessary usage of highest error correcting capability code when not needed. MEC also able to maintain the remaining energy as the SNR changes as we had assumed that as the SNR increased, less probability of congestion would occur. Thus, the highest error correcting capability might not be necessary when retransmission might have been necessary to solve the corrupted packets. During this research, some inconsistencies have been recorded. This inconsistency recorded during the SNR = 6 until SNR = 12 as the RS code is used instead of BCH. Thus, the used of RS (63, 45) seems to use more energy than expected because RS (63, 45) have high error correcting capability and from Fig. 7, it helps to reduce BER and make MEC outperform the existing BCH. Figure 6 shows that BER during the RS code being applied had increased slightly compared to the existing BCH. This is because, lower error correcting capability of RS code was used rather than that higher error correcting capability of BCH. This saves energy but does not lower the BER. However, the usage of same error correcting capability between both of BCH and RS code for MEC had outperformed the inconsistency problem which also helps in optimized between the remaining energy and average BER.

6 Conclusion and Future Works

The MEC algorithm had shown some promising result to maximize lifetime by means of extending the remaining energy according to the channel condition. BER of MEC had slightly outperformed the existing technique for MEC BCH (15, 5) RS (15, 5) and MEC BCH (31, 16) RS (31, 15). The inconsistency of BER obtained from the initial result when the error correcting capability, $t = 10$ might due to the redundancy added during the transmission. The increasing number of nodes will gradually increase the BER and energy usage. However, by accurately assign the proper error correcting capability, t to the MEC error correcting codes can help in reducing and optimized the BER and remaining energy. The MEC needed to be further studied to ensure its reliability in congested networks. Fur future works, MEC can be tested with convolutional codes and its effectiveness towards the different environment of sensor nodes implemented.

Acknowledgement. This research is supported by the Research Management Institute, Universiti Teknologi MARA and this publication is funded by Ministry of Higher Education (MOHE). (Registered under the Fundamental Research Grant Scheme (FRGS) #600-RMI/FRGS 5/3 (46/2015)).

References

1. Benkic, K.: Proposed use of a CDMA technique in wireless sensor networks. In: 2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, pp. 343–348. IEEE (2007)
2. Gill, H.S., Gaba, G.S., Gupta, N.: Reduction of multiple access interference in CDMA by using improved minimum mean square error receiver. Int. J. Sci. Eng. **2**, 2–5 (2011)
3. Razali, S., Mamat, K., Bashah, N.K.: A new approach for wireless sensor network lifetime maximization and low overhead with hybrid ARQ (HARQ) error control protocol. Aust. J. Basic Appl. Sci. **9**, 36–41 (2015)
4. Gupta, P., Mohammed, H., Hashem, M.M.A.: Characterization of downlink transmit power control during soft handover in WCDMA systems. CoRR abs/1304.3, 0–4 (2013)
5. Razali, S.M., Mamat, K., Abdul-Basit, K., Ali, F.H.M.: Performance enhancement of wireless sensor network (WSN) with the implementation of Hybrid ARQ (HARQ) and Transmission Power Control (TPC). In: 2014 IEEE Conference on Wireless Sensors (ICWiSE), pp. 36–40. IEEE (2014)
6. Liu, B.H., Chou, C.T., Lipman, J., Jha, S.: Using frequency division to reduce MAI in DS-CDMA wireless sensor networks. In: IEEE Wireless Communications and Networking Conference 2005, pp. 657–663. IEEE (2005)
7. Masood, M.M.Y., Ahmed, G., Khan, N.M.: A Kalman Filter based adaptive on demand transmission power control (AODTPC) algorithm for wireless sensor networks. In: 2012 International Conference on Emerging Technologies, pp. 1–6. IEEE (2012)
8. Grewal, M.S., Andrews, A.P.: Kalman Filtering: Theory and Practice with MATLAB. Wiley, Hoboken (2014)
9. Haykin, S.: Kalman Filtering and Neural Networks. Wiley, Hoboken (2004)
10. Faragher, R.: Understanding the Basis of the Kalman Filter via a Simple and Intuitive Derivation. Important and Common Data Fusion Algorithms (2012)

11. Ribeiro, A., Schizas, I.D., Roumeliotis, S.I., Giannakis, G.B.: Kalman Filtering in Wireless Sensor Networks (2010)
12. Welch, G., Bishop, G.: An Introduction to the Kalman Filter, pp. 1–16 (2006)
13. Karafilis, L.: SNR margin, line attenuation, ADSL and a slow internet speed. https://www.giantstride.gr/snr-margin-adsl/
14. Datta, U., Kundu, S.: Performance of multi-hop CDMA wireless sensor networks with correlated interferers using different retransmission strategies and error control schemes. Int. J. Sens. Networks **15**, 40 (2014). https://doi.org/10.1504/IJSNET.2014.059988
15. Datta, U., Sen, S., Kundu, S.: Energy level performance of HARQ-II scheme in CDMA wireless sensor network with correlated interferers. In: Proceedings of 2011 Annual IEEE India Conference Engineering Sustainable Solutions, INDICON-2011, pp. 1–4 (2011). https://doi.org/10.1109/indcon.2011.6139478
16. Meghdadi, V.: BER calculation. Communications, 1–9 (2008)
17. Weber, S., Andrews, J.G., Jindal, N.: The effect of fading, channel inversion, and threshold scheduling on ad hoc networks. IEEE Trans. Inf. Theory **53**, 4127–4149 (2007). https://doi.org/10.1109/TIT.2007.907482
18. Lin, S., Costello, D.J.: Error Control Coding: Fundamentals and Applications. Pearson-Prentice Hall, Upper Saddle River (2004)
19. Kleinschmidt, J.H., da Cunha Borelli, W.: Adaptive error control using ARQ and BCH codes in sensor networks using coverage area information. In: 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1796–1800. IEEE (2009)
20. Akyildiz, I.F., Vuran, M.C.: Wireless Sensor Networks. Wiley, Hoboken (2010)

MYTextSum: A Malay Text Summarizer Model Using a Constrained Pattern-Growth Sentence Compression Technique

Suraya Alias[1(✉)], Siti Khotijah Mohammad[2], Keng Hoon Gan[2], and Tan Tien Ping[2]

[1] Knoweldge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
suealias@ums.edu.my

[2] School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Penang, Malaysia
{sitijah,tienping}@webmail.cs.usm.my, khgan@usm.my

Abstract. As more information becomes accessible online, users are faced with difficulties in digesting and selecting important information from longer text. A summary can serve as a condensed version of a text, where salient information can be presented. In order to improve a summary's quality, a special task in the area of Automatic Text Summarization known as Sentence Compression (SC) can be applied. Existing SC techniques are highly dependent on syntactic knowledge applied on individual word or phrases to decide on the compressions decision. In contrast, this study introduces a new constrained Pattern-Growth SC (PGSC) technique inspired by the "*divide and conquer*" strategy tailored to the Malay language. The basic idea is to divide the sentences into segments where unimportant segments are removed while the important ones are conquered iteratively. Using a Malay news dataset, the application of PGSC have shown promising results where the compressed summaries reported an F-Measure score of 0.5752 agreements when evaluated against manual human summaries and perform better than the Baselines (uncompressed) model. Manual human evaluation produced *readability* score of 4.31 out of 5 and 4.1 for *content responsiveness*, suggesting a better quality and readability of the compressed summaries produced by the MYTextSum model.

Keywords: Text Summarization · Sentence Compression
Sequential Pattern-Growth · Malay

1 Introduction

Automatic Text Summarization (ATS) is an automated process of creating a summary from a single or multiple document input sources. The output type of summary can either be generic, query-focused (based on specific user-topic), or sentiment-based such as summarizing user's opinion. The methods to produce an automated summary can be performed via extractive or abstractive methods [1, 2]. An extractive method selects and concatenates the most important sentences to produce a shorter version of a document. In contrast, an abstractive method constructs a summary by modifying, paraphrasing

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 141–150, 2018.
https://doi.org/10.1007/978-981-10-8276-4_14

and joining related information to form a new sentence, which is somewhat similar with the summary produced by humans.

However, due to the complexity in generating an abstractive based summary where extensive Natural Language Processing (NLP) knowledge is needed, the extractive summarization method has dominated the research area in the ATS field until today [3]. As a way to improve the quality of the extractive summary produced, literature in ATS focusing in Sentence Compression (SC) techniques has been an interest to researchers.

Researcher [4] defined SC as a problem in ATS where the goal is to remove unimportant constituents from a summary sentence while preserving the salient ones through keeping the sentence's grammar intact. SC can also be viewed as a scaled down version of summarization performed at a sentence level where the problem is typically formulated as a word deletion task [5]. Some known SC techniques such are *linguistics rule* based [4, 6, 7], *statistical* [5, 8, 9], *machine learning* [10, 11], *keyword-based* [12] and *Integer Linear Programming* (ILP) by [13] have been previously explored. Furthermore, some recent works in this area also include *Graph* related optimization by [14, 15] that has been applied in the area of Multi-Sentence Compression.

Most of the aforementioned SC techniques perform *extractive compression* and are highly dependent on syntactic knowledge (syntactic parser, dependency parser) applied on individual word or phrases to decide on the compressions decision. The main reason for this dependency is to avoid composing ungrammatical sentence after the compression process. Nonetheless, there is still trade-off that exists between the model's performance in balancing both *grammatical* and *informativeness* of the summary.

Alternatively, this study proposes a new constrained Pattern-Growth SC (PGSC) technique inspired by the *divide and conquer* strategy in Sequential Pattern Mining tailored to the Malay language. The underlying idea is to divide the sentences into segments where unimportant segments are removed while the important ones are conquered iteratively. A new pattern-based representation named Frequent Adjacent Sequential Pattern (*FASP*) with "*textual constraint*" discovered in this study serves as a feature to identify significant information from the text document. Meanwhile, from the Malay Text Summary corpus developed in this study, a set of human's compression pattern named Frequent Eliminated Pattern (*FASPe*) with conditional probability confidence *Conf* value was mined to indicate the constituents that are frequently removed by human practice. The removal decision in his PGSC technique is derived from the combination of both discovered textual patterns fulfilling the proposed "*removal constraints*".

2 Related Works

In creating an extractive summary, an extractive summarizer model usually perform three main tasks described in [16] that are consists of (1) creating intermediate representation of the input text, (2) sentence scoring and (3) sentence selection. The task of Sentence Compression can be applied at any level of the summarization process such as during pre-processing step or before the sentence scoring step.

A Linguistic Rule Based approach relies on training corpus of human summaries where heuristic linguistic rules or knowledge can be built upon understanding how human writes, removes and constructs certain phrases in a summary. Prior work in rule-based approach by [4] in their Automatic Sentence Reduction system has reported a success rate of 78.1% for the decision of either to remove or retain a clause. They added knowledge resources such as syntactic knowledge from WordNet, lexicon database, contextual and statistics information extracted from human summaries to assist the decision to remove a certain phrase from a sentence. However, the sentence parse tree in their system has a very high dependency on those additional resources in which large training and expensive corpus resources are needed.

Meanwhile, in keyword based approach, the deletion refers to list of keywords such as in [12]. They generate a list of "function" words to be trimmed that contains conjunctions such as "*As a matter of fact*" or "*however*", where it has yield better summary performance. Meanwhile, [7] applied sets of linguistic motivated rules to the compression process iteratively to each source sentences before moving to the sentence extraction module also known as the "*parse-and-trim*" approach.

Another common SC problem definition include viewing the sentence as a word deletion problem using Statistical Probabilistic and Decision Tree model [5]; where the Probabilistic model learns how likely each sentence is compressed using the Naïve Bayes Rules. Despite the good grammatical performance in individual sentence level, a study by [17] reported that the statistical approach has insignificant effect on the overall Text Summarization system performance. This is due that during the compression process; some important content might have been dropped (deleted). Following the drawbacks of this model, researchers have made some improvement by applying Machine Learning approach. For instance, [11] combines unsupervised Compression approach where additional rule and enforcement on the deletion constraint were added; while [10] added semantic information from WordNeT to the Decision Tree Model. This additional task has shown enhancement on the accuracy of the sentence reduction algorithm based on the words importance.

Another approach in solving the SC problem is where the task is viewed as an optimization problem using Integer Linear Programming (ILP) method by [13]. The presence of linguistically motivated constraints has gained better performance over models without constraint. However, the success of an ILP approach highly relies on finding the optimum set of linear constraints for the compression output which becomes a trade-off between summary's performance and optimization.

Currently, some works in this area has included Graph related optimization which has been applied in the Multi-Sentence Compression [14]. A Multi-Sentence Compression (MSC) is a task to produce a compressed single sentence summary from a cluster of related sentences. The idea is to eliminate redundant sentence by weighting the edge by i.e. link frequencies to search for the lightest and shortest path based on pre-defined minimum length. Nevertheless, the issue in graph representation due to the "*shortest path*" approach is missing of salient information which effect on the summary generation [18]. Due to this, some improvement has been experimented that includes adding heuristic knowledge to the link frequencies of the connected words and also re-ranking method based on the key phrase extraction by [15]. Their work shows some

improvement in the score of sentence informative value of the compressed summary, but a slight decreased on grammatical evaluation value.

To summarize, thus far this study has not encountered a pattern-growth based text representation model for the Malay language and for Sentence Compression Technique. Hence, this study would like to extend the interest in this area by combining pattern-growth based text representation with heuristic rules from human linguistic pattern to develop a Sentence Compression technique for Malay ATS.

3 MYTextSum Framework

The MYTextSum model workflow in Fig. 1 consists of: (1) Pre-processing, (2) Sentence Representation using *FASP*, (3) Pattern-Growth Sentence Compression technique, (4) Sentence Scoring, (5) Sentence Selection and (6) Summary Generation.

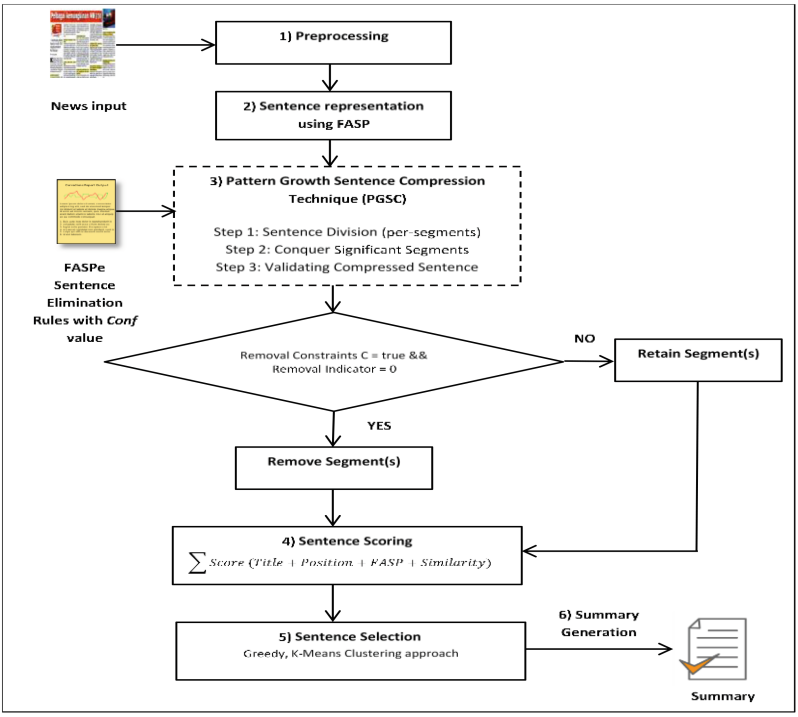


Fig. 1. The MYTextSum model with PGSC technique

3.1 Preprocessing

The pre-processing step involves general tasks such as tokenizing, removing Malay stop words [19], removing punctuation and converting terms to lowercase. Since the aim is

to preserve the sentence flow and linguistic pattern of the terms, no word stemming process was applied.

3.2 Sentence Representation Using *FASP*

The basis of Pattern-Growth method in Sequential Pattern Mining is to divide the database into smaller sets based on the current discovered frequent patterns, and then conquering the subsequent pattern, based on the local frequent pattern. Given pattern P is a sequence of terms, the support of pattern P is the frequency of a sequence.

***FASP* Generation and Discovery**

The discovered *FASP* is used as text features where it represents significant information for each sentence s in the article. Given user specified minimum support threshold denotes as *min_sup* or σ; The *gSupp* counts the overall frequency of the sequence being used against in the source collection. If the $gSupp(P) \geq \sigma$, then the sequence is considered as frequent or is a Frequent Pattern (FP). Given user specified minimum confidence threshold denotes as *min_conf* or β, Sequential Rules that meets the confidence value or *Conf* condition can be generated. In the initialization step, a set of frequent terms of length-1(one) was set as the *prefixTerms* list denoted as α. After initialization, the next step is to divide the search space of each document based on the *prefixTerms* α. Each term in the *prefixTerms* list is used as prefix to recursively project or conquer the word sequence list. This is done by joining the α initialized with the next adjacent sequence term $t_{(k+1)}$ denoted β from the list of sentences in the current document.

The generation of *FASP* for each document follows the equation in (1):

$$\boldsymbol{FASP}_{\boldsymbol{m}} = \boldsymbol{\alpha} \cup \boldsymbol{\beta} \tag{1}$$

However, before the joining process, this study introduced three textual constraints C known as (1) *item*, (2) *adjacent sequence order* and (3) *length* constraints. Table 1 shows an example of sentence representation using *FASP*-weight pairs using $gSupp \geq 2$.

Table 1. Sample sentence representation using *FASP*-weight pairs until 2-sequence using an excerpt from Article GBK009 with gSupp ≥ 2

s_i	Sentence	*FASP representation*
s_3	"Kenyataan ini tidak adil kepada anggota penyelamat dan malim gunung yang bertungkus-lumus membantu mangsa," katanya ketika dihubungi BH, semalam	s_3 = *{(anggota,2), (gunung,4), (kenyataan,2), (malim,3),* ***(malim gunung, 3)****, (mangsa,3), (membantu,2),* ***(membantu mangsa, 2)****, (penyelamat,2), (semalam,2)}*
s_4	Beliau berkata, operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancar, sekali gus mengelakkan musibah lain	s_4 = *{(berkata,3), (agensi,2), (sar,3)}*

3.3 A Constrained Pattern-Growth Sentence Compression (PGSC) Technique

The Divide and Conquer strategy

The first step is to *divide* the source sentence from the news article into smaller sets or segments to mark which constituent is eligible to be deleted without affecting the sentence's grammar and structure. Dividing the sentence into *segments* has the benefit of smaller text structure with reduced search space to find any frequent pattern or salient information in the inspected segment. Here, each candidate sentence is segmented into sequence of terms based on the comma (,) delimiters. From this study's observation, the panels frequently removed a group of adjacent words together since they complement each other semantically. The list of compression pattern named Frequent Eliminated Pattern or (*FASPe*) with conditional probability confidence *Conf* value discovered from the Malay Summary corpus is given in Table 2, where the discovery process can be referred in [20].

Table 2. Sample of human's Frequent Eliminated Pattern (*FASPe*) with *Conf* value discovered in the Malay Summary corpus

#	*FASPe*	*eSupp*	*gSupp*	*Conf*
1	mengulas	11	13	0.84
2	**mengulas → lanjut**	6	7	**0.85**
3	sehubungan	5	6	0.83
4	sehubungan → itu	5	6	0.83
5	beliau	52	88	0.59
6	berkata	120	206	0.58
7	beliau → berkata	21	28	0.75
8	dalam	152	325	0.46
9	dalam → pada	15	20	0.75
10	**dalam pada → itu**	13	20	**0.65**

Conquer Significant Segment using Textual Patterns

After all the sentences are divided into segments, the main problem is to decide which seg_i should be removed or retained. Here, the normalized weight w for textual patterns *FASP* and *FASPe* in each seg_i is used to assist the removal decision. The *FASPe Conf* value is used to determine the compression level since the list of available *FASPe* relies on it. For each source sentence s in a document, let $seg_i = \{x_1, x_2, \ldots x_n\}$, where x_i denotes the term sequence in each segment seg_i. Each term x_i in seg_i is weighted if it is a subset of *FASP* or *FASPe* where each occurrence (support) in the pattern is counted. The normalized weight w was calculated by dividing the sum of each pattern's support for each x_i that belongs to *FASP* or *FASPe* against the inspected segment length.

In performing the sentence compression, the removal decision γ for each segment was based on these five *removal constraints C*:

C1. One segment sentence was not removed. It is assumed that a one segment sentence is a complete sentence. Since each sentence is divided into segments using a comma delimiter, the removal was based on segments and not per-word deletion.

C2. After the removal, each compressed sentence *cs* length must be ≥ 2, which is based on the statistical result in Malay Summary corpus, where the shortest sentence produced by the panel with the length of 2.

C3. Each *cs* should conform to one of Malay Grammar pattern. If this constraint is not satisfied, the compression process is reverted back.

C4. The compression level for each *cs* is based on finding optimum *Conf* value within the range of 0.1–1.0. The optimum *Conf* value should minimize the error found in ***C3*** and also preserve the most significant information.

C5. The binary *removal indicator* γ for each seg_i was based on the normalized weight *w* of *FASPe* and *FASP* where each *cs* also satisfies the **C1–4**, where the value 1 from the removal indicator γ indicated that the segment should be kept since it is mostly conquered by the *FASP*. Meanwhile, the value of 0 means that the segment should be dropped or eliminated since it contains higher *FASPe* with respects that it has fulfilled all the described *removal constraints C*. However, if the score of the normalized weight is the same, the segment is not removed to avoid over compressing the sentence.

From Table 3, it can be seen that seg_1 consists of *FASPe* of "*beliau berkata*" with the weight of 1.0. The calculated *FASP* weight is 0.5 since only one matched significant pattern was discovered, thus giving the $w(FASPe) > w(FASP)$ indicating the decision to remove the segment represented by binary removal indicator of 0. Meanwhile, seg_2 consists of higher *FASP* weight, meaning that the phrase is significant and the $w(FASPe) < w(FASP)$, thus the decision is to retain the segment. Next, the phrase in seg_3 does not contained any matched *FASP* and with higher *FASPe* weight; therefore, it will be also removed. It should be stated here that the discovered *FASP* in the current document might be the same with *FASPe* from the corpus since the *FASP* is uniquely generated per-document. In this event, the *support* for each pattern will have the effect on the final weight. Finally, the new compressed sentence from the example that has fulfilled all the *removal constraints* is written as: ***"Operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancer"***, where only seg_2 is retained.

Table 3. Example of *FASP* and *FASPe* weighting for each segment using *min_conf*> 50%.

seg_i	Sentences	*FASPe* Weight	*FASP* Weight	Removal Indicator
seg_1	***Beliau berkata***	1.0	0.5	**0**
seg_2	*operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancar*	0.0	0.2	1
seg_3	***sekali gus mengelakkan musibah lain***	0.2	0.0	**0**

3.4 Summary Generation

To generate a summary, each sentence (compressed and uncompressed) is scored. In this Sentence Scoring module, the combination of linear sentence features to score

important sentence is used. The Sentence Scoring is the combination of *Surface*, *Content* and *Relevant* features. This study extends the *Content* and *Relevant* scoring mechanism by using the proposed *FASP* model as feature to represent significant information from the news article instead of using only frequent keyword.

The sum of linear combination is firstly sorted based on the highest ranking value to select content-bearing relevance sentences. However, from observation, for a single summary generation, the flow of the story is usually kept according to the position of the sentence. The selected sentence based on the scores is ranked again by its' sequential position accordingly to preserve the flow of the story. In the last step in that is the Summary Generation, the candidate sentences from the previous step that was ranked by score and position are added to the summary until the desired summary length.

4 Experiments and Findings

Experiments using the Malay news dataset have been conducted in order to evaluate the effects on using a compressed sentence to compose a summary against extracting the sentence as a whole. In the Malay Summary corpus developed by this study, the total of 100 Malay news articles was downloaded from the Bernama Library & Infolink Service (BLIS) website using a keyword search on the Events and Natural Disaster (ND) domain in Malaysia. These articles are aligned automatically with 300 human summaries of 30% length for each news article, where 2,058 sentence pairs are analyzed. In this experiment, 30 articles are selected in random from the ND and Events dataset. There are five summarization methods used for comparison in this experiment. The MYTextSumCOMP method is using *Greedy* sentence selection approach, while MYTextSumCLUST method employed *K-Means* clustering algorithm using *FASP* as the features in the Sentence Selection process. Since there is no baseline and benchmark Malay Summarizer model available for comparison, this research denotes the MYTextSumBASIC method as Baseline1. Meanwhile, Baseline2 or (Lead baseline) follows the generation of baseline summaries used in the DUC conference for evaluating English summarizer [21]. Both baselines are without the application of sentence compression task. The automatic summary evaluation is performed using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit by [22]. ROUGE evaluates the quality of a summarizer by counting overlapping units such as word sequences, word pairs and n-grams between the content of the automated summary with one or more reference summaries (human manual summaries). The results in Table 4 shows that all developed methods in the MYTextSum model using the *FASP* representation consistently perform better than the baselines method and the open source OTS summarizer. The highest F-measure result of 0.5752 from the MYTextSumCLUST method is found significant against the OTS and Baseline2 method. The results indicates that with the application of PGSC technique has improves the produced summary content when evaluated against human manual summaries.

Manual evaluation on the generated summaries from the MYTextSum model is performed following the DUC 2005 guidelines where the summary is manually evaluated for its *readability*, and *content responsiveness* based on a five-point Likert scale of

1–5. The MYTextSumCLUST method produced a readability score of 4.31 and 4.1 for content responsiveness, suggesting a better quality and readability of the compressed summaries against uncompressed ones.

Table 4. ROUGE-1 results between MYTextSum methods and other summarizers using Random dataset of 30 articles

Methods	#	Recall	Precision	F-Measure
Baseline1	3	0.5813	0.5496	0.5621
MYTextSumCOMP	2	0.5851	0.5546	0.5671
MYTextSumCLUST	*1	**0.5926**	**0.5650**	**0.5752**$^{\dagger+}$
Baseline2 (Lead)	4	0.5637	0.5386	0.5472
OTS	5	0.5551	0.4961	0.5216

5 Conclusion and Future Works

This paper contributes in developing a new SC technique based on a constrained Pattern-Growth technique tailored to the Malay language. From the experiments conducted has shown promising results using automatic and manual evaluation. The application of PGSC technique in MYTextSum model significantly improves the summary's quality against baselines (uncompressed) model summarizers. Next, this study would like to extend the SC work to cater for multi-document summarization.

Acknowledgement. This work is supported by Universiti Sains Malaysia (USM), Research University Grant (RU) by project number 1001/PKOMP/811295.

References

1. Hahn, U., Mani, I.: The challenges of automatic summarization. Computer **33**(11), 29–36 (2000)
2. Das, D., Martins, A.F.T.: A survey on automatic text summarization. Lit. Surv. Lang. Stat. II Course CMU **4**, 192–195 (2007)
3. Fang, C., et al.: Word-sentence co-ranking for automatic extractive text summarization. Expert Syst. Appl. **72**, 189–195 (2017)
4. Jing, H.: Sentence reduction for automatic text summarization. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, Association for Computational Linguistics (2000)
5. Knight, K., Marcu, D.: Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artif. Intell. **139**(1), 91–107 (2002)
6. Jing, H., McKeown, K.R.: The decomposition of human-written summary sentences. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (1999)
7. Zajic, D., et al.: Multi-candidate reduction: sentence compression as a tool for document summarization tasks. Inform. Process. Manag. **43**(6), 1549–1570 (2007)
8. Knight, K., Marcu, D.: Statistics-based summarization-step one: sentence compression. AAAI/IAAI **2000**, 703–710 (2000)

9. Galley, M., McKeown, K.: Lexicalized Markov grammars for sentence compression. In: HLT-NAACL (2007)
10. Nguyen, L.M., et al.: A new sentence reduction technique based on a decision tree model. Int. J. Artif. Intell. Tools **16**(01), 129–137 (2007)
11. Turner, J., Charniak, E.: Supervised and unsupervised learning for sentence compression. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2005)
12. Conroy, J.M., et al.: Back to basics: CLASSY 2006. In: Proceedings of DUC (2006)
13. Clarke, J., Lapata, M.: Global inference for sentence compression: an integer linear programming approach. J. Artif. Intell. Res. **31**, 399–429 (2008)
14. Filippova, K.: Multi-sentence compression: finding shortest paths in word graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics (2010)
15. Boudin, F., Morin, E.: Keyphrase extraction for N-best reranking in multi-sentence compression. In: North American Chapter of the Association for Computational Linguistics (NAACL) (2013)
16. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data 2012, pp. 43–76. Springer (2012)
17. Lin, C.-Y.: Improving summarization performance by sentence compression: a pilot study. In: Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, Association for Computational Linguistics (2003)
18. Tzouridis, E., Nasir, J.A., Brefeld, U.: Learning to summarise related sentences. In: COLING (2014)
19. Abdullah, M.T., et al.: Improvement of Malay information retrieval using local stop words. In: International Advanced Technology Congress (ATC), IOI Marriott Hotel, Malaysia (2005)
20. Alias, S., et al.: A Malay text corpus analysis for sentence compression using pattern-growth method. Jurnal Teknologi **78**(8), 197–206 (2016)
21. Jones, K.S.: Automatic summarising: the state of the art. Inform. Process. Manag. **43**(6), 1449–1481 (2007)
22. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-2004 Workshop (2004)

A FIPA-ACL Ontology in Enhancing Interoperability Multi-agent Communication

Kim Soon Gan[1], Kim On Chin[1(✉)], Patricia Anthony[2], and Abdul Razak Hamdan[3]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
g_k_s967@yahoo.com, kimonchin@gmail.com

[2] Faculty of Environment, Society and Design, Lincoln University, Christchurch, New Zealand
patricia.anthony@lincoln.ac.nz

[3] Faculty of Information Science & Technology, Center for Artificial Intelligence Technology (CAIT), Universiti Kebangsaan Malaysia (UKM), 43600 Bangi Selangor, Malaysia
arh@ukm.edu.my

Abstract. The nature of computing paradigm has shifted from centralized, static and closed to distributed, dynamic and open due to the advent and popularity of Internet. Multi-agent system (MAS) gained popularity as its characteristic match with this paradigm shift. In order for MAS to interact efficiently and communicate meaningfully, agent communication language (ACL) plays an important role. FIPA-ACL is an ACL developed by FIPA that has become the de facto standard of ACL's implementation in MAS. Another emerging trend due to Internet is the Semantic Web (SW). Semantic web is an extension of the current World Wide Web, which encodes the content of the web with well-defined meaning to allow it to be processed by machines such as computer (agents). Hence, combining the existing FIPA-ACL with semantic web can bring ACLs to another level to enhance the interoperability in MAS. In this paper, a FIPA-ACL ontology in OWL is proposed to enhance the communication between agents in MAS.

Keywords: Agent communication language · FIPA-ACL · Semantic web Ontology · OWL

1 Introduction

Internet has changed the computational interaction in today's computing environment. The computing paradigm has shifted from centralized, closed, static to distributed, open and dynamic. This paradigm shift is also referred to as open system [1, 2]. Agent technology which emerged as a new software computing paradigm matches the open system [3, 4]. Multi-agent system is made of multiple heterogeneous agents that may be distributed in different environments and can enter and leave freely in the agent community [5]. As a result, agents need to interact or communicate with one another in order to achieve some goals or tasks [6]. For agents to carry out interaction such as coordination, cooperation, collaboration and negotiation, they need to communicate in a manner that can be understood by each agent [7–10]. Agent communication language (ACL) [11] is

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 151–160, 2018.
https://doi.org/10.1007/978-981-10-8276-4_15

the communication language used to achieve interoperability between agents. Two of the most common ACLs adopted by the MAS are KQML and FIPA-ACL. KQML is the first ACL developed by DARPA knowledge sharing effort [12]. The high level communication was initially developed to exchange information and knowledge for knowledge based system but it was then extended for communication between agents. FIPA-ACL was developed by FIPA which is a non-profit organization established to promote agent technology and interoperability between agent applications [13, 14]. FIPA-ACL is adopted more widely than KQML in recent years due to the formal semantic of FIPA-ACL and affordance of FIPA in promoting agent technology through as set of specification that acts as guideline for developers and industry vendors.

Another emerging trend in Internet systems is the semantic web. Semantic web is the extension of the current web by encoding the web content resources with well-defined meaning that are machine process able [15]. In order to realize the semantic web, a set of technologies such as RDF [16], RDFS [17], and OWL [18] have been created. In addition, other technologies such as SWRL [19] and SPIN [20] were created to incorporate rules into semantic web to express constraints, rules and business logics. There is an increasing trend of applying semantic web into MAS to increase computational efficiency and to communicate meaningfully. Moreover, SW is mainly used as ontology, knowledge base, validity constraints and consistency checking. SW has also been used in ACL to encode communication. In this paper, a FIPA-ACL ontology is developed to act as standardized ontology for MAS' adoption and extend it as an ACL component in MAS to enhance the interoperability between MAS. The proposed ontology reused and structured some of the concepts in the existing ontology in order to achieve interoperability across MAS applications. Besides that, the semantic web technology, SPIN is used in developing the ontology alongside the foundation technologies such as RDF, RDFS and OWL 2.

This paper introduces and describes a FIPA-ACL ontology that can be reused and imported by FIPA-ACL agents' applications. The ontology is created using a step-by-step ontology development process. The remainder of this paper is organized as follow. Section 2 describes some of the related work in FIPA-ACL ontology. The ontology engineering methodology is described in Sect. 3. Section 4 describes the steps in FIPA-ACL ontology development and finally, Sect. 5 concludes and summarizes the work presented in this paper.

2 Related Work

In this section, some of the related works in ontology-based FIPA-ACL are reviewed to serve as the foundation and motivation of this research. This review is organized in chronological order to observe how the researches in this area have progressed. One of the earlier works is by FIPA on its specification to encode the content of the message in RDF encoding known as the FIPA RDF Content Language Specification [21]. This specification specifies how to represent objects, propositions and actions in RDF and in different RDF versions with different extensions. The motivation for encoding in RDF is to increase the level of interoperability. The advantages of RDF encoding include

extensibility, reusability, simplicity and standards for data and schemas exchange. However, RDF is just a data format for encoding and exchange, and so its expressiveness of RDF is limited.

Another semantic web technology DAML has also been proposed to encode the ACL message [22]. The richer expressivity of DAML compared to RDF and RDFS enriched the content expressed in the ACL message. In this work, a DAML ontology is defined for the communication. A demo application of ITTalks project was used to illustrate the communication encoding in DAML. Unfortunately, DAML as a cornerstone of OWL did not gain momentum as semantic web technology.

Zou continued his research effort in ACL using semantic web language as content language but this time he focused on OWL. Travel Agent Game in Agentcities (TAGA) is a FIPA-compliant framework that extended and enhanced Trading Agent Competition (TAC) scenario in Agentcities as an open MAS environment. OWL was used as the content language in FIPA-ACL messages for agents to communicate with each other and reasoning about the action and services. By utilizing the benefits of OWL, the content of messages can be more expressive, unambiguous, computer-interpretable, interoperable, has automated reasoning techniques, higher-level of interoperability between agents and meaningful content can be shared [23, 24].

AgentOWL is another research that incorporated OWL as representation for agent internal model and communication [25]. It is a MAS distributed framework that is built on top of JADE with a generic knowledge model for agent. Formal description of the model is represented using description logic. There are five main elements in the model including resources, actions, actors, context and events. AgentOWL used Jena semantic reasoner to reason the context, resource, action and knowledge domain of the agent and content of agent communication. CommonKADs is used to model and developed the MAS, the UML and AUML modeling language is used for modeling formalism. As a result, the agent knowledge model is more generic and expressive and can be easily adopted and used in other similar applications.

Pu used semantic web representation for negotiation protocol in electronic commerce [26]. Ontology is used to describe the negotiation protocol to increase the efficiency of the negotiation process. The negotiation protocol used in this research is the contract net protocol. The prototype demo was built on top of JADE and the reasoning used Jena API. This work was further extended using OWL for the agent communication process in a layered architecture.

Semantic agent model (SAM) used semantic web technologies in MAS modeling and knowledge base support [27]. In SAM, semantic web rule language (SWRL) is used in modeling the behaviour and constraints of different agents. A three layered architecture is used to model the agents which is made up of a knowledge base layer, an engine layer and a low level action layer. The different states of agent actions are modelled using extended FSM concept.

Fornara has been the one of the important key players in representing ACL social semantic model using obligation and norms [28, 29]. An obligation ontology is developed to model the obligation between debtor and creditor in temporal proposition. An external program is used to keep track of the temporal constraints in the temporal proposition to

determine the different states of the obligations. Fornara and her team has further the research into the representation of policy and artificial institution [30, 31].

It can be observed that there are some ongoing works in ACLs. However, the complete ontology for FIPA-ACL has not been developed yet. Hence, the objective of this research is to develop an FIPA-ACL ontology for communication interoperability between agents.

3 Methodologies

To develop the FIPA-ACL ontology, we used the ontology engineering approach. The ontology development process used in this work is from the Noy and McGuiness ontology engineering approach [32] which was also adopted by many of the researchers working on ontology development. This ontology development process includes several steps that require iterations for refinement and evaluation. The steps in this ontology development approach are as follow:

1. Determine the domain and scope of the ontology.
2. Consider reusing the existing ontologies.
3. Enumerate important terms in ontology.
4. Define the classes and class hierarchy.
5. Define the properties of classes.
6. Define the facets of slots.
7. Create instances.

The first step in the ontology development is defining the domain and scope of the ontology. This is done by going through a list of questions, which consist of basic questions and competency questions. By going through this list of questions, the domain and scope of the ontology can be roughly identified. The next step is to consider reusing existing ontologies by checking whether there are any existing ontologies that are suitable to be reused or extended. Reusing any existing well known ontologies can increase the interoperability of the developed ontology. The next step is to identify a list of important terms that are related to the ontology domain. This list of terms can be identified from numerous related documents or from domain experts. Normally, nouns identified will be the class of the ontology and verbs will be the relationship of the connected classes in the ontology. The list of terms will lay the foundation for defining the classes and class hierarchy in the next step. There are several approaches in developing the class hierarchy such as the top-down development approach, bottom-up development approach and the hybrid approach. The top-down development approach is based on the most general concept and is narrowed down to specialized concept. On the other hand, the bottom-up development approach is based on the most specific concept and grouping related specific concept to form general class. The hybrid approach is a combination of those two processes. After defining the classes and its hierarchy, the next step is to define the properties for these classes. These properties describe the characteristics of individuals of these particular classes. The relationship to relate the individuals between classes is also defined in this step. Next, the facets of the identified

properties values are defined. These facets values include the cardinality, value-type, domain and range. The final step is to populate instances and individuals into the ontology. These steps need to be performed in several iterations before the final ontology is developed and completed.

4 FIPA-ACL Ontology Development

4.1 Determine the Domain and Scope of the Ontology

The domain ontology will cover the FIPA-ACL specification which includes the communication acts, the interaction protocols, the content language and the message structure. The ontology will be used by agents (autonomous entity) that exchange information and knowledge that conforms to FIPA (agents are communicating using FIPA-ACL). The ontology will identify the terms and vocabularies used for FIPA-ACL conversation message exchanges and to infer the performatives that are used in the message. The ontology can then be imported by other agents that use FIPA-ACL as communication language and it can also be further extended by those who want to introduce extra performatives besides the basic set of performatives to enhance the interoperability between agent communication. Furthermore, the ontology might be able to answer the following competency questions. Which performatives should be used for the next message based on the content and the performatives of the previous message? Before performing the selected performative, a validation is carried out by verifying whether the agent knows about the content and whether the agent is able to carry out the actions by referring to the knowledge base.

4.2 Consider Reusing the Existing Ontologies

There are two related ontologies of FIPA-ACL that have been developed by Zou in TAGA [22, 23] and Dickinson in NUIN [33]. These two ontologies will serve as the base reference ontologies for developing the ontology in this paper. In Zou's ontology, no semantic model was considered in defining the FIPA-ACL ontology. In Dickson's ontology [33], only the mentalistic model in the FIPA-ACL ontologies was considered. Another ontology that will be reused in this work is the obligation ontology by Fornara [28, 29]. This obligation ontology is based on the semantic model of social commitment inferred from social interaction.

4.3 Enumerate Important Terms in the Ontology

The list of terms for this ontology are extracted from various documentation including the FIPA specifications, publications in agent communication language, agent communication semantic model, interactions between agents and others. However, the most important document is the set of FIPA-ACL specifications since this is the main guideline for developers to develop FIPA-ACL MAS. Some of these terms include performatives, message, conversation, agent, sender, receiver, content, language, encoding,

ontology, conversation-id, reply-with, in-reply-to, and reply-by. The nouns that identify the list of terms would normally be considered as classes and the verbs that identify the list of terms would be the relationships that connect the classes.

4.4 Define the Classes and the Class Hierarchy

The hybrid approach is used to define the FIPA-ACL ontology. The more salient concepts are defined first and then generalization and specialization are applied to these concepts appropriately. For example, the performative class is a super class for the different performatives subclasses and protocol classes. The list of protocols are shown in Fig. 1. Other salient classes identified as part of the ontology are shown in Table 1.

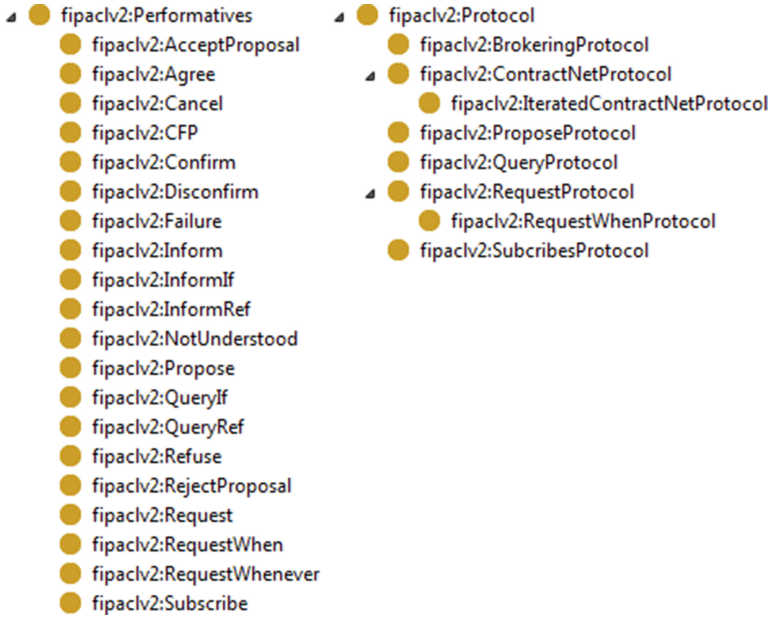


Fig. 1. FIPA-ACL ontology class hierarchies.

Table 1. FIPA-ACL ontology classes

Classes	rdf:type	SubClassOf
Agent	owl:Class	owl:Thing
AgentPlatform	owl:Class	owl:Thing
CommunicativeAct	owl:Class	owl:Thing
Conversation	owl:Class	owl:Thing
Message	owl:Class	owl:Thing
Object	owl:Class	owl:Thing
Ontology	owl:Class	owl:Thing
Protocol	owl:Class	owl:Thing
Role	owl:Class	owl:Thing
Rule	owl:Class	owl:Thing
Service	owl:Class	owl:Thing
State	owl:Class	owl:Thing

4.5 Define the Classes and the Class Hierarchy

After identifying the class hierarchy, the attributes of the classes and relationships between classes are identified. The relationships between classes are identified through the list of verbs from the terms identified. The attributes of the classes are the remaining terms that are both not classes and relationships. The relationships between classes will be identified as the object properties, whereas the attributes are identified as data properties. Table 2 shows the some of the object properties identified for the ontology.

Table 2. Object properties with domain and range

PropertyName	Domain	Range
atPlatform	Agent	AgentPlatform
hasAgentAddress	Agent	URL
hasContent	Message	Fact, Object, Action
Ontology	Message	Ontology
hasProtocol	Message	Protocol
hasReceiver	Message	Agent
hasSender	Message	Agent
hasService	Agent	Service
inReplyTo	Message	Expression
ownership	Service	Agent
replyBy	Message	Time
replyTo	Message	Agent
replyWith	Message	Expression

4.6 Define the Facets of the Attributes

The facets identified include the allowable value type, values, cardinality and other features. The cardinality for the attributes can be specified as the maximum value, minimum value and the exact value. The basic allowable value types would be based on xml schema data types permitted in OWL 2 DL. The value types can also be an instance value of classes defined. If the value type is an instance value then universal and existential quantification, set operators of union and intersection can be used to define the slot facets. The domain and range of the properties are defined to restrict the allowable value for each relationship. Some of the domains and ranges of the relationship property are shown in Table 2.

4.7 Create Instances

In this ontology no instances will be populated because the instance of the ontology will be created during the exchange of messages between agents. Some of the testing instances are created to validate the consistency and soundness of the ontologies.

5 Conclusions

This objective of this research is to create a more complete FIPA-ACL ontology based on FIPA-ACL specification that is required to facilitate the communication between agents. By combining FIPA-ACL and OWL 2 DL for the FIPA-ACL ontology, it is believed that it can provide additional advantages and overcome limitations such meaningful communication content, standardize ontology term in achieving interoperability in agent communication and allowing performative reasoning in agent communication. OWL 2 DL is a decidable subset of first order logic. Hence, using OWL 2 DL as a content language makes the logic of representation decidable compared to modal logic or first order logic. Besides, the rich expressiveness of OWL 2 DL and its incremental inferment of new knowledge can be used to represent the different types of content for FIPA-ACL including object, action and expression. In the next step, FIPA-ACL ontology will be equipped with the semantic model to allow the verification of the communication process.

Acknowledgments. This project is funded by the Ministry of Higher Education, Malaysia under RACE0014-TK-2014 and it is also partially supported by Artificial Intelligence Research Unit (AiRU).

References

1. Luck, M., McBurney, P., Preist, C.: Agent technology: enabling next generation computing (a roadmap for agent based computing). AgentLink (2003)
2. Artikis, A.: Dynamic specification of open agent systems. J. Logic Comput. **22**(6), 1301–1334 (2012)

3. Luck, M., McBurney, P., Shehory, O., Willmott, S.: Agent technology roadmap: Overview and consultation report. AgentLink III (2004)
4. Luck, M., McBurney, P., Shehory, O., Willmott, S.: Agent technology: computing as interaction (a roadmap for agent based computing). University of Southampton (2005)
5. Sycara, K.P.: Multiagent systems. AI Mag. **19**(2), 79–93 (1998)
6. Payne, T.R., Paolucci, M., Singh, R., Sycara, K.: Communicating agents in open multi agent systems. In: Proceedings of the 1st GSFC/JPL Workshop on Radical Agent Concepts (WRAC) (2002)
7. Beer, M., D'inverno, M., Luck, M., Jennings, N., Preist, C., Schroeder, M.: Negotiation in multi-agent systems. Knowl. Eng. Rev. **14**(3), 285–289 (1999)
8. Jennings, N.R.: Commitments and conventions: the foundation of coordination in multi-agent systems. Knowl. Eng. Rev. **8**(3), 223–250 (1993)
9. Olfati-Saber, R., Fax, J.A., Murray, R.M.: Consensus and cooperation in networked multi-agent systems. Proc. IEEE **95**(1), 215–233 (2007)
10. Poslad, S.: Specifying protocols for multi-agent systems interaction. ACM Trans. Auton. Adapt. Syst. (TAAS) **2**(4), 1–24 (2007)
11. Labrou, Y., Finin, T., Peng, Y.: Agent communication languages: the current landscape. IEEE Intell. Syst. Appl. **14**(2), 45–52 (1999)
12. Finin, T., Weber, J., Wiederhold, G., Genesereth, M., Fritzson, R., Mckay, D., Beck, C.: Draft specification of the KQML agent-communication language (1993)
13. FIPA.: FIPA specification part 2: Agent communication language. Technical report, FIPA-Foundation for Intelligent Physical Agents (1997)
14. FIPA.: FIPA ACL Message Structure Specification. Foundation for Intelligent Physical Agents (2002)
15. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
16. Cyganiak, R., Wood, D., Lanthaler, M.: RDF 1.1 concepts and abstract syntax. W3C Recommendation (2014)
17. Brickley, D., Guha, R.V., McBride, B.: RDF Schema 1.1. W3C Recommendation (2014)
18. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 web ontology language primer. W3C Recommendation (2009)
19. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A semantic web rule language combining OWL and RuleML. W3C Member submission (2004)
20. Knublauch, H., Hendler, J.A., Idehen, K.: SPIN-Overview and Motivation. W3C Member Submission (2011)
21. FIPA.: RDF Content Language Specification. FIPA-Foundation for Intelligent Physical Agents (2000)
22. Zou, Y., Finin, T., Peng, Y., Joshi, A., Cost, S.: Agent communication in DAML world. In: Innovative Concepts for Agent-Based Systems: First International Workshop on Radical Agent Concepts, WRAC 2003. Springer, Heidelberg (2003)
23. Zou, Y., Finin, T., Ding, L., Chen, H.: TAGA: using semantic web technologies in multi-agent systems. In: International Joint Conference on Artificial Intelligence 2003 (2003)
24. Zou, Y., Finin, T., Ding, L., Chen, H., Pan, R.: Using semantic web technology in multi-agent systems: a case study in the TAGA trading agent environment. In: Proceedings of the 5th International Conference on Electronic Commerce, pp. 95–101. ACM (2003)
25. Laclavik, M., Balogh, Z., Babik, M., Hluchý, L.: AgentOWL: semantic knowledge model and agent architecture. Comput. Inf. **25**(5), 421–439 (2012)
26. Pu, Q., Fu, S., Cao, Y., Hou, Z.: Adopting ontology and agent in electronic negotiation service. In: 8th IEEE International Conference on Cognitive Informatics, ICCI 2009, pp. 547–551. IEEE (2009)

27. Subercaze, J., Maret, P.: SAM: semantic agent model for SWRL rule-based agents. In: International Conference on Agents and Artificial Intelligence, vol. 2, pp. 244–248 (2010)
28. Fornara, N., Colombetti, M.: Ontology and time evolution of obligations and prohibitions using semantic web technology. In: International Workshop on Declarative Agent Languages and Technologies, pp. 101–118. Springer, Heidelberg (2009)
29. Fornara, N., Colombetti, M.: Representation and monitoring of commitments and norms using OWL. AI Commun. **23**(4), 341–356 (2010)
30. Fornara, N., Okouya, D., Colombetti, M.: Using OWL 2 DL for expressing ACL content and semantics. In: European Workshop on Multi-Agent Systems, pp. 97–113. Springer, Heidelberg (2011)
31. Fornara, N., Cardoso, H.L., Noriega, P., Oliveira, E., Tampitsikas, C., Schumacher, M.I.: Modelling agent institutions. In: Agreement Technologies, pp. 277–307. Springer, Netherlands (2013)
32. Noy, N.F., McGuinness, D.L.: Ontology development 101: a guide to creating your first ontology. Knowledge Systems Laboratory, Stanford University. Stanford Knowledge Systems Laboratory Technical report KSL-01-05 and Stanford Medical Informatics Technical report SMI-2001-0880 (2001)
33. Dickinson, I.J.: BDI Agents and the Semantic Web: Developing User-Facing Autonomous Applications. Doctoral Dissertation. University of Liverpool (2006)

Gamification Effect of Loyalty Program and Its Assessment Using Game Refinement Measure: Case Study on Starbucks

Ooi Wei Xin[1], Long Zuo[2(✉)], Hiroyuki Iida[2], and Norshakirah Aziz[1]

[1] Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia
{o.weixin_20731,norshakirah.aziz}@utp.edu.my
[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
{zuolong,iida}@jaist.ac.jp

Abstract. This paper explores the advantage of loyalty program in the domain of business, while Starbucks is chosen as a case study. It focuses mainly on the point system that provides a certain degree of gamification effect. It considers a game progress model of My Starbucks Rewards to derive a game refinement measure for the assessment of gamification impact. The assessment results indicate that the game element of point system in My Starbucks Rewards shows motivations towards the normal purchasing activities. On the other hand, the point system shows the decreasing of motivation effect towards customers' purchases over the time. In short, customers are experiencing unsophisticated game experience in a point system which is proved to be a short term incentive that is useful to motivate customers in the early age for a short period of time. Starbucks incorporates both point system and tier system in its loyalty program, targeting to attract new customers as well as retain them for a long time to come. However, the current study only examines the point system of Starbucks. Further research might explore more on structure of loyalty program in restaurant or food industry.

Keywords: Starbucks · Point system · Loyalty program
Game refinement theory

1 Introduction

Coffee statistics report shows that coffee shops make up the fastest growing part of the restaurant business, checking in with a 7% annual growth rate [3]. In today highly competitive marketplace, customer retention increasingly becomes the attention of businesses. Customer retention refers to the activities and actions of companies and organizations which take to reduce the number of customer

Starbucks Corporation™. Trademarks are fair used here under Fair Use for educational purpose.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 161–171, 2018.
https://doi.org/10.1007/978-981-10-8276-4_16

defections [5]. Retention of customers stands out to be so crucial because the cost of finding a new customer is about 5 to 10 times more than to keep the existing one. Loyal customers often know exactly what they want when purchasing from a certain brand and tend to spend more as well. Current customers tend to spend 67% more than new customers. Unfortunately, companies tend to lose around 13% of their customers every 5 years [11].

Therefore, loyalty of a customer is critical to the success of a business. Generally, loyalty program is defined as a reward program offered by a business to customers who frequently make purchases [13]. Loyalty program typically requires customers to register with their information and customers will be given a unique membership ID or physical membership card to be used when making purchases. From the perspective of business owner, loyalty program works to track customers' purchase behavior and history in order to recognize the loyal customers and hence reward them. Meanwhile, customer engagement is further improved when the feeling of appreciation by the company is established in the customer.

According to a statistic [4], there is a 26% rise in profit and an 11% total revenue lift when Starbucks introduced the My Starbucks Rewards program. Spokesman in Los Angeles Daily News announced that coffee chain's most loyal customers visit 18 times a month. Maritz Study in 2016 reported that more than 45% of consumers buy a product to gain rewards in a loyalty program [6]. Starbucks loyalty program had always been known to be a successful gamification mobile marketing strategy. Gamification is defined as the employment of game elements in non-game context to improve user's engagement. Early works had been done by Zuo et al. [22] on the analysis of hotel loyalty program, identifying the game experience of loyalty program in the world's largest hotel chain. In this paper, we aim to investigate the gamification effect of Starbucks loyalty program, specifically across its evolution as well as comparing among Starbucks US and Starbucks China.

This paper begins with the overview of Starbucks loyalty program and follows by reviewing the basic idea of game refinement theory. Next, we implement the game refinement assessment in Starbucks loyalty program. Finally, the conclusion gives a brief summary and critique of the findings.

2 Starbucks Loyalty Program "My Starbucks Rewards"

The first Starbucks was opened in 1971 in Seattle's Pike Place Market with just a narrow storefront. With the growing numbers of stores worldwide, Starbucks launched "My Starbucks Rewards" loyalty program in 2009. Currently having more than 24,000 stores in 70 countries, and 13 million of "My Starbucks Rewards" active members [18]. "My Starbucks Rewards" is a free loyalty program introduced by Starbucks that gives exclusive member offers and allow customers to earn rewards such as free drinks, foods and refills. To earn rewards, customers simply need to pay for any Starbucks product with a registered Starbucks card. Each time a purchase is made, customer will earn a specific amount

of stars that can be redeemed for free Starbucks treat. The terms and conditions of "My Starbucks Rewards" varies according to different countries. As of 2016, the number of Starbucks licensed stores globally is ranked by Unites States (US) as the first with 5,292 stores and follows by China with a total of 1,110 stores [17]. In this paper, we observe mainly "My Starbucks Rewards" in two countries: United States (US) and China.

2.1 US

Starbucks US had always been a great example for its loyalty program over the years. The evolution of its loyalty program is so successful that it is able to recognize and retain its loyal customers. Starbucks makes changes and improvements to its loyalty program by listening to their customers' feedback and ideas. Figure 1 illustrates the evolution of Starbucks loyalty program. In 2009, December 26, Starbucks US officially launched its loyalty program, "My Starbucks Rewards", a visit-based rewards system which consists of three levels. By opening an account and registering a Starbucks card, customer will be automatically enrolled in "My Starbucks Rewards" at the Welcome Level. Customer can earn a star each time visit Starbucks. The rewards in Welcome Level includes a free birthday beverage, and up to two continuous hours of free Wi-Fi access daily. After successfully collected five stars, customer will reach the Green Level where customer is qualified for the; benefits in Welcome Level, free beverage customization, free brewed coffee refills, free tall beverage of choice with purchase of one pound of whole bean coffee, and special offers. Customer who earns at least 30 stars will be promoted to Gold Level to enjoy all the benefits in Welcome Level and Green Level, in addition receiving a personalized gold card and a free drink with every 12 additional stars.

Fig. 1. Evolution timeline of Starbucks loyalty program

In 2012, October 16, Starbucks makes some modifications on "My Starbucks Rewards" that launched in 2009. This improvised version of "My Starbucks Rewards" basically eases the free redemption process where postcards evolve into email notification, rewards can be used for food and drink redemption and 12 stars for a free item redemption instead of 15 stars [12].

In 2016, April 12, Starbucks introduced another "My Starbucks Rewards" which is a spending-based rewards system. This rewards system has the same

benefits as the old system. However, the main difference with the previous system is that this rewards system consists of only two levels, namely Green Level and Gold Level. Customer will be automatically qualified as Green Level member once successfully registered. With each dollar spent, customer will earn two stars. Customers are required to earn 300 stars to be eligible for Gold Level promotion. In Gold Level, customers need to earn an additional of 125 stars in order to trade for a free item. Customers can enjoy the redemption of unlimited number of free items with every 125 stars earned during the year of Gold Level membership.

2.2 China

In January 1999, Starbucks entered the mainland China market by opening the first store in the China World Trade Building, Beijing [8]. "My Starbucks Rewards" in China consists of three levels, somehow similar with the old "My Starbucks Rewards" in US, that is Welcome Level, Green Level, and Gold Level. Released in the middle February of 2011 [2], "My Starbucks Rewards" had gained high popularity in China.

The star or point in "My Starbucks Rewards China" is rewarded with every spending of RMB50. In Welcome Level, members will receive several e-coupon during the membership year which includes three pieces of "buy one get one free" handcrafted beverage, one piece of free morning complimentary tall-sized beverage before 11am and one piece of free upgrade from tall to grande or grande to venti. After earning five stars, Welcome Level members will be promoted to Green Level where members can enjoy a free birthday beverage and one piece of "buy three beverages get one free" e-coupon. Within the 12 months of membership period, Green Level member will be upgraded to Gold Level after earning 25 stars, else will be downgraded to Welcome Level. In Gold Level, members are eligible for one free birthday beverage, Gold Level My Starbucks Rewards card, one free tall size beverage during account anniversary and one piece of "10 purchases get one complimentary beverage".

By comparing "My Starbucks Rewards" in US and China, there are some differences in term of structure and the rules of the loyalty programs. The main interest is to identify the main successful core structure which made up of point system that has been used in both loyalty programs which will be further discussed in Sect. 4.

3 Assessment Methodology

This section presents the game refinement theory to derive a measure of game sophistication which will be used for the assessment of game elements of Loyalty Program "My Starbucks Rewards".

3.1 Mathematical Model of Game Refinement

Classical game theory [15] originated with the idea of the existence of mixed-strategy equilibria in two-person zerosum games. This has been widely recognized as a powerful tool in many fields such as economics, political science, psychology, logic and biology. Game refinement theory is another game theory focusing on the attractiveness and the sophistication of games. The foundation of this direction was made by Iida et al. [9], in which a measure of game refinement was proposed based on the concept of information of game outcome uncertainty. A logistic model was constructed in the framework of game refinement theory and applied to many board games including chess variants and Mah Jong [10]. Later a general model of game refinement was proposed based on the game information progress model and applied to time limit games such as soccer and basketball [20]. While game theory concerns a player's winning strategy, game refinement theory is concerned with the whole game, including quality of play and entertainment.

The "game progress" is two fold. One is game speed or scoring rate, while another one is game information progress which focuses on the game outcome. Game information progress presents the degree of certainty of the game's results in time or steps. If one knows the game information progress, for example after the game, the game progress $x(t)$ will be given as a linear function of time t with $0 \leq t \leq t_k$ and $0 \leq x(t) \leq x(t_k)$, as shown in Eq. (1).

$$x(t) = \frac{x(t_k)}{t_k}\, t \tag{1}$$

However, the game information progress given by Eq. (1) is usually unknown during the in-game period. Hence, the game information progress is reasonably assumed to be exponential. This is because the game outcome is uncertain until the very end of game in many games. Hence, a realistic model of game information progress is given by Eq. (2).

$$x(t) = x(t_k) \left(\frac{t}{t_k}\right)^n \tag{2}$$

Here n stands for a constant parameter which is given based on the perspective of an observer in the game considered. If one knows the game outcome, for example after the game, or if one can exactly predict in advance the game outcome and its progress, then we have $n = 1$, where $x(t)$ is a linear function of time t. During the in-game period, various values of the parameter n for different observers including players and supporters will be determined. For example, some observers might be optimistic with $0 \leq n < 1$. However, when one feels any difficulty to win or achieve the goals, the parameter would be $n > 1$. Meanwhile, we reasonably assume that the parameter would be $n \geq 2$ in many cases like balanced or seesaw games. Thus, the acceleration of game information progress is obtained by deriving Eq. (2) twice. Solving it at $t = t_k$, the equation is shown in Eq. (3).

$$x''(t_k) = \frac{x(t_k)}{(t_k)^n} {t_k}^{n-2}\, n(n-1) = \frac{x(t_k)}{(t_k)^2}\, n(n-1) \tag{3}$$

It is assumed in the current model that game information progress in any type of game is encoded and transported in our brains. We do not yet know about the physics of information in the brain, but it is likely that the acceleration of information progress is subject to the forces and laws of physics. Therefore we expect that the larger the value $\frac{x(t_k)}{(t_k)^2}$ is, the more the game becomes exciting, due in part to the uncertainty of game outcome. Thus, we use its root square, $\frac{\sqrt{x(t_k)}}{t_k}$, as a game refinement measure for the game under consideration. We can call it GR value for short.

Early researches [16,20] had found that the game refinement values of various games which include boardgames, time-limits sports games and score-limits sports games lie between a zone value of 0.07 to 0.08 as tabulated in Table 1.

Table 1. Measures of game refinement for different type of games

Game	$x(t_k)$	t_k	GR
Chess	35	80	0.074
Go	250	208	0.076
Basketball	36.38	82.01	0.073
Soccer	2.64	22	0.073
Badminton	46.34	79.34	0.086
Table tennis	54.86	96.47	0.077

3.2 Game Progress Model of My Starbucks Rewards

In this paper, the game progress is studied from the perspective of point system in "My Starbucks Rewards". The game progress in point system can be measured by two factors: the number of free items redeemed, and the total number of items consumed. In order to get the game refinement value, Eq. (4) is proposed, where F and C represents the number of free items redeemed and total number of items consumed, respectively.

$$GR = \frac{\sqrt{F}}{C} \tag{4}$$

4 Data Collection and Analysis

This section starts by analyzing the different versions of point system in Starbucks loyalty program specifically in US, then follows by analyzing point system implemented that in "My Starbucks Rewards" in both US and China.

4.1 Evolution of Point System in Starbucks US Loyalty Program

In the evolution of "My Starbucks Rewards" since 2009, Starbucks US had continuously made minor changes on the rules in redeeming free items. Hence, the game refinement (GR) value of different versions of "My Starbucks Rewards" from the perspective of point system in Gold Level is analyzed.

Table 2. GR values of different versions of Starbucks US loyalty program (point system) in earning a free item in Gold Level

Year	Version	F	C	GR
2009	My Starbucks Rewards (visit-based)	1	15	0.067
2012	My Starbucks Rewards (visit-based)	1	12	0.083
2016	My Starbucks Rewards (spending-based)	1	16	0.063

As shown in Table 2, the point system of "My Starbucks Rewards" in US from 2009 until 2016 shows the GR value in the range of 0.063 to 0.083. The GR value is calculated and tabulated by assuming the very first free item redeemed by the customer, F and the required number of item consumption, C is computed based on the average price of a cup of Starbucks coffee which is $4 and RMB27 in US and China respectively. The higher the GR value indicates the lower the required number of consumption in order to redeem for the free item.

The changes of rules in "My Starbucks Rewards" of 2012 had shown an increase in value of GR. This means that customers in Gold Level are required to spend lesser than previously for a free item redemption. On the other hand, comparing the visit-based "My Starbucks Rewards" in 2012 with the spending-based "My Starbucks Rewards" in 2016, the GR values show a decrease of 0.02, from 0.083 to 0.063. This result may be explained by the fact that Starbucks requires its customer to spend more in order to get a free item. Before introducing the spending-based "My Starbucks Rewards", customer can spend less than $4, which is the average price of a cup of Starbucks coffee, in a transaction for 12 times to earn a free item. However, Starbucks values its loyal and high-spending customer, hence changes are made that star is earned based on amount of spending. This changes are less entertaining and more challenging for the customer who spends minimum amount to earn stars or free items.

4.2 Point System

This subsection is mainly focusing on the point system in Gold Level of "My Starbucks Rewards" in both US and China. When a customer reaches the Gold Level to enjoy the benefit of free item redemption, the effect of repeating redemption for free items on the GR value is observed.

The differences between "My Starbucks Rewards" in US and China is tabulated in Table 3. Firstly, the difference can be seen from the number of level,

Table 3. Differences between my starbucks rewards in US and China

Country	US	China
Number of level	2	3
Requirement to reach or stay in Gold Level	38 cups	47 cups
Requirement to get one free item in Gold Level	16 cups	10 cups

where "My Starbucks Rewards US" consists of only 2 levels, but total of 3 levels in "My Starbucks Rewards China". Secondly, in order for a customer to reach or stay in Gold Level, the customer is required to at least consume 38 cups of coffee in US and 47 cups in China during the membership year, taking the average price per cup is $4 and RMB27 respectively. Thirdly, the requirement for a customer to redeem for a free item in Gold Level is 16 cups in US and 10 cups in China. From the differences, we found that the requirement for a customer to reach or stay in Gold Level in US is lower than in China. However, the requirement for a customer to get one free item in Gold Level is higher in US as compared to China. With these differences, the redemption of free item for My Starbucks Rewards US and China in long term is observed in Fig. 2.

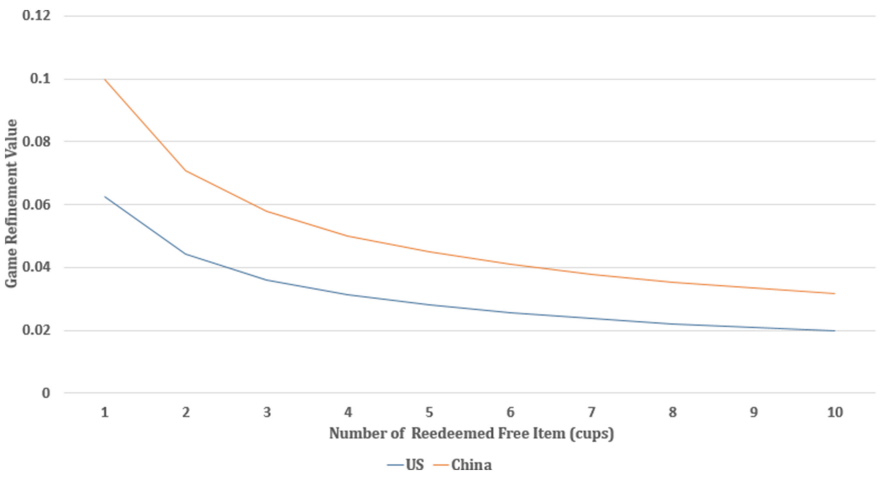


Fig. 2. Game refinement value for point system in Gold Level with increasing number of free cups redemption

In economics, Gossen [7], a Prussian economist explains that a consumer's utility which is the satisfaction derived from consuming a service or product decreases with the increasing consumption of that particular service or product. In other words, the first unit of consumption of a service or product yields more utility than the second and subsequent unit. This decrease of marginal utility

with the increase of consumption is known as Law of Diminishing Marginal Utility. Mathematically, it is represented by Eq. (5).

$$MU_1 > MU_2 > MU_3 ... > MU_n \tag{5}$$

where MU_i stands for the marginal utility with the frequency $i \in \mathbb{N}$. Marginal utility may decrease into negative utility, as it may become entirely unfavorable to consume another unit of the product. Therefore, the first unit of consumption for any product is typically highest, with every unit of consumption to follow holding less and less utility [7].

From the graph in Fig. 2, the GR value for China is relatively high which yields the value of 0.1, whereas US starts with the GR value of 0.063. Nevertheless, both curves in US and China shows the downwards sloping trend which indicates that the GR value is decreasing with the increasing number of free item redemption. This inverse relationship is identical to the Law of Diminishing Marginal Utility in economics if we assumed that the game refinement relates to satisfaction of a person. For instance, a person feels less excitement with lower GR when the number of free item redemption increases.

Hence, points system is concluded to be a normal rewards or incentives system that is able to attract new customers to purchase and join the point system loyalty program. Eventually, the GR value or customer's satisfaction will decrease. Thus, point system seems to be suitable for business that have frequent, and short term purchases.

The Law of Diminishing Marginal Utility directly relates to the concept of diminishing prices. As the utility of a product decreases as its consumption increases, consumers are willing to pay smaller dollar amounts for more of the product. Same goes to the case of free items redemption, customers become less willing to purchase the same amount of item in order to get another free item which will give less satisfaction.

5 Conclusion

Generally, the game element of point system in "My Starbucks Rewards" shows motivation towards the normal purchasing activities. From the case study on Starbucks, the GR zone value of point system is observed to be ranged from 0.063 to 0.10. "My Starbucks Rewards" in China demonstrates the highest GR value of 0.1, whereas "My Starbucks Rewards" in US shows the lowest GR value of 0.063. Meanwhile, the GR value of point system in both US and China will eventually decrease as the number of free item redemption increases. Hence, we concluded that point system shows decreasing motivation effect towards customers' purchases over the time. In short, customer is experiencing unsophisticated game experience in point system which is proved to be a short term incentive that is useful to motivate customers in the early age for a short period of time. Starbucks incorporates both point system and tier system in its loyalty program, targeting to attract new customers as well as retain them for a long time to come. However, the current study only examines the point system in the

case of Starbucks. Further research might explore more on structure of loyalty program in restaurant or food industry to determine the appropriate or universal comfortable zone for loyalty program in business domain.

Acknowledgements. This research is funded by a grant from the Japan Society for the Promotion of Science (JSPS), within the framework of the Grant-in-Aid for Challenging Exploratory Research.

References

1. BucksCard, tracking the Starbucks Gift Card. https://buckscards.com/history.htm
2. China My Starbucks Rewards card. http://www.buckscards.com/china_rewards.htm
3. Coffee Statistic (2016). http://www.e-importz.com/coffee-statistics.php
4. Fisher, E.: 2016's Most Important Customer Loyalty Statistic (2016). http://www.annexcloud.com/blog/2016/02/05/ultimate-customer-loyalty-statistics-2016/
5. Roland, T.R., Anthony, J.Z.: Customer satisfaction, customer retention, and market share. J. Retail. **69**(2), 193–215 (1993)
6. Garai, T.: Why Tiered Program Are Great For Loyalty. https://antavo.com/blog/tiered-programs-great-customer-loyalty/
7. Gossen, H.H.: Die Entwicklung der Gesetze des menschlichen Verkehrs und der daraus fließenden Regeln für menschliches Handeln. Translated into English as The Laws of Human Relations and the Rules of Human Action Derived Therefrom. MIT Press, Cambridge (1983)
8. History of Starbucks China. https://www.starbucks.com.cn/en/about/history
9. Iida, H., Takeshita, N., Yoshimura, J.: A metric for entertainment of boardgames: its implication for evolution of chess variants. In: Entertainment Computing Technologies and Applications, pp. 65–72. Springer, Boston (2003)
10. Iida, H., Takahara, K., Nagashima, J., Kajihara, Y., Hashimoto, T.: An application of game-refinement theory to Mah Jong. In: International Conference on Entertainment Computing (ICEC), pp. 333–338. Springer, Heidelberg (2004)
11. Kumar, K.: An Analysis On Brand Loyalty: A Case Study On Starbucks (2016)
12. Bahsin, K.: Starbucks Rewards Program Changes. Business Insider (2012). http://www.businessinsider.com/starbucks-rewards-program-changes-2012-9
13. Loyalty Program. http://www.investopedia.com/terms/l/loyalty-program.asp
14. Magatef, S.G., Tomalieh, E.F.: The impact of customer loyalty programs on customer retention. Int. J. Bus. Soc. Sci. **6**(8), 78–93 (2015)
15. Neumann, J.: Zur theorie der gesellschaftsspiele. Math. Ann. **100**(1), 295–320 (1928)
16. Nossal, N., Iida, H.: Game refinement theory and its application to score limit games. In: Games Media Entertainment, pp. 1–3. IEEE (2014)
17. Number of Starbucks Stores Globally, 1992–2016. Knoema. https://knoema.com/kchdsge/number-of-starbucks-stores-globally-1992-2016
18. Smith, C.: 22 Interesting Starbucks Facts and Statistics, January 2017. http://expandedramblings.com/index.php/starbucks-statistics/
19. Starbucks (2017). https://www.starbucks.com/
20. Sutiono, A.P., Purwarianti, A., Iida, H.: A mathematical model of game refinement. In: Reidsma, D. (ed.) INTETAIN 2014. LNICST, vol. 136, pp. 148–151. Springer, Cham (2014)

21. Zuo, L., Iida, H.: An analysis of sales promotion 'discount' using game refinement measurement. In: Munekata, N., et al. (eds.) LNCS, vol. 10507, pp. 487–491. Springer, Cham (2017)
22. Zuo, L., Xiong, S., Iida, H.: An analysis of hotel loyalty program with a focus on the tiers and points system. In: 4th International Conference on Systems and Informatics (ICSAI), pp. 507–512. IEEE (2017)

Rule-Based Model for Malay Text Sentiment Analysis

Khalifa Chekima(✉), Rayner Alfred, and Kim On Chin

Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
k.chekima@gmail.com, {ralfred,kimonchin}@ums.edu.my

Abstract. With the increase number of opinionated content on the web, organizations and people have shown tremendous interest in knowing other's opinions. This phenomena has attracted both the academic and the business world to pay a close attention towards the development of automated tools which helps in sentiment analysis (SA). While different well-defined approaches have been defined for English SA, the problem remains far from being solved for other languages such as Malay language, despite having more than 215 million Malay native speakers worldwide. To the author's knowledge, most of researches on Malay language SA rely heavily on the use of bag-of-words model (BOW), which resulted Malay SA to have low accuracy, as BOW model disrupts word order, breaks the syntactic structures and discards some semantic information of the text. In this paper, we propose new feature sets that refine the traditional sentiment feature extraction method and take contextual valence shifters into consideration from a different perspective compared to the earlier research concerning Malay language. The most common valence shifter (VS) are considered in this paper, this includes negation, intensifier, diminisher and contrast. Negation is considered to be the most obvious and common valence shifter of all. A new technique is proposed in this paper to handle complex negation compared to the existing techniques where only simple negations (Bigram) are handled. The proposed system is then compared with existing techniques. The final result showed improvements in Malay SA after considering valence shifter. The discussion and implication of these findings are further elaborated.

Keywords: Malay sentiment analysis · Valence shifter
Unsupervised technique

1 Introduction

In standard practice, sentiment analysis is considered as a special case of text classification, where a review text is normally represented by a bag-of-words (BOW) model. Then, statistical machine learning algorithms, such as Naïve Bayes, maximum entropy classifier, and support vector machine (SVM) are used for classification. However, the BOW model disrupts word order, breaks the syntactic structures and discards some semantic information of the text. Hence, it brings about some fundamental deficiencies including the polarity shift problem. Polarity shift refers to a linguistic phenomenon in

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 172–185, 2018.
https://doi.org/10.1007/978-981-10-8276-4_17

which the polarity of sentiment can be reversed (i.e., positive to negative or vice versa) by some special linguistic structures called polarity shifters, e.g., negation ("I don't like this movie") and contrast ("Fairly good, but not my style").

In another definition, Polarity shifters, also called "sentiment shifter" in [2], and "valence shifter" in [1] are words and phrases that can change sentiment orientations of texts. Polarity shift is a complex linguistic structure that may include explicit negations, contrasts, intensifiers, diminishers, irrealis, etc. [3, 4] reported that valence shifters are very frequent, and they cause a significant number of errors if not handled correctly, further more handling them correctly could significantly improves the performance of sentiment classification systems, as he reported that approximately 15% of valence shifter frequencies in all sentences [5]. Have conducted a statistic on the distribution of different types of polarity shift, and reported that explicit negations and contrasts covers more than 60% polarity shift structures. Negation is the most common type of polarity shifter.

Although being of importance, valence shifter issue has been investigated only recently and is now the object of an increasing attention. Additional to that, to the author's knowledge, none of the previous work concerning Malay text sentiment analysis have investigated the impact of valence shifter on Malay opinionated text, despite its importance in sentiment analysis as reported by researchers while dealing with languages such as English language. This has motivated us to further investigate the importance of valence shifter on Malay text, and to what extent proper handling of valence shifter improves Malay sentiment analysis.

In this work, a rule-based model is proposed that handles explicit Valence shifter in Malay text sentiment analysis including negation, intensifier, diminisher and contrast. The rest of the paper is organized as follow. Section 2 discusses related work. Section 3 discusses our approach, this include an explanation on of methods and algorithms adopted. Section 4 discusses the evaluation and result. Finally, we conclude our paper in Sect. 5.

2 Related Work

Sentiment analysis is a growing research field, widely explored by both the academia and the industry for diverse application such as e-health data visualization human-computer interaction, etc. [6]. However, automatically extracting polarity from text is a difficult and challenging process as it encounters varies NLP problems such as sarcasm detection, polarity shifter, and anaphora resolution.

[8] first postulated the presence of contextual valence shifters, which are contextual phenomena altering the prior polarity of a term. Afterwards, some of these phenomena (such as negative or conditional syntactic structures) were dealt with on a case by case basis [9–15]. Studies addressing the phenomenon as a whole flourished later. They aimed at best modelling the expression of opinions [3, 8, 16–18], before embedding those in a classification system. The main purposes of these studies were to determine a list of contextual valence shifters that impact the polarity of a term as well as to define the nature of this impact. However, these lists are often manually built from linguistic

intuitions and not learned from language data. Works relying on a corpus of texts to develop resources that best reflect the actual role played by the linguistic context for opinion mining are few. [19] suggested a technique to automatically select polarity shifting features in order to improve a sentiment classification system based on a machine-learning approach.

The way to handle polarity shift also differs in two types of methods namely term-counting method and machine learning method. It is relatively easy to encode polarity shift in the term-counting methods, as we can reverse the sentiment of the polarity-shifted words and phrases directly, and then sum up the sentiment score word by word [20–23] discussed some complex negation effects by using conjunctive and dependency relations among polarity words. [24] developed four rules (intra-clause rule, intra-sentence rule, extra-sentence rule, and extra- paragraph rule) for detecting different types of polarity shift, and employed a term-counting model to encode the information of polarity shifts. Empirical studies showed that their method yield much better performances than the basic term-counting approach.

[21] handled polarity shift in both term-counting and machine learning applications. However, although the system is effective in term-counting systems, it is relatively difficult to handle polarity shift in machine learning methods, perhaps because the polarity shift information is hard to be integrated into the BOW model. In handling negation, [9] proposed a simple method by adding "NOT" to sentiment words, but [11] reported that the effect is very limited in improving the sentiment classification accuracy. Some researchers attempted to model polarity shift by conducting more complex linguistic analysis. For example, [10] tried to model negation by investigating specific part-of-speech tag patterns. [21] employed syntactic parsing to capture three types of valence shifters (negative, intensifiers, and diminishers).

[25] proposes a machine learning approach to tackle negation shifting by adding the tag 'not' to every word between a negation trigger word/phrase (e.g., not, isn't, didn't, etc.) and the first punctuation mark following the negation trigger word/phrase. To their disappointment, considering negation shifting has a negligible effect and even slightly harms the overall performance. [21] explored negation shifting by incorporating negation bigrams as additional features into machine learning approaches. The experimental results show that considering sentiment shifting greatly improves the performance of term-counting approaches but only slightly improves the performance of machine learning approaches. Other studies such as [10, 26, 27] also explore negation shifting and achieve some improvements.

[28] Examined the effect of CVS on the classification of all reviews. Moreover, [28] tested the sentiments by three factors: negation, intensifier and diminisher. The first method is to count the terms and to classify all reviews based on the number of positive and negative terms in the article. [28] Applied the result of [29] to identify the positive and negative terms, as well as the negation, diminishing and intensifier terms with the data set of the film reviews [28, 30] showed that the expansion of term-counting method with CVS is able to improve the accuracy of a classification.

As a conclusion to this section, all of these studies agreed that contextual valence shifters can have diverse impacts on polarized words: inverses invert the polarity of a polarized item, intensifiers intensify it and attenuators diminish it.

3 Proposed Method

Most opinion mining systems rely on the extraction of sentiment words to detect opinions. These words, which we will rather refer to as polarized words, convey useful information about the semantic orientation (positive or negative) of a text. However, the context in which these words appear may modify their valence in many ways. Although being of importance, this issue has been investigated only recently and is now the object of an increasing attention.

Figure 1 provides an overall picture of the method adopted in this study. For text preprocessing, we used stopword list developed by [7], as for lexicon, we used Malay lexicon developed by [31] due to its high accuracy and wider coverage of Malay polar words. In essence, this work presents rule-based model for Malay sentiment analysis by taking into account valence shifter. The four most common valence shifter considered in this study are negation, contrast, intensifier and diminisher. Each of the valence shifter type adopted in this study are elaborated further in the coming sections.

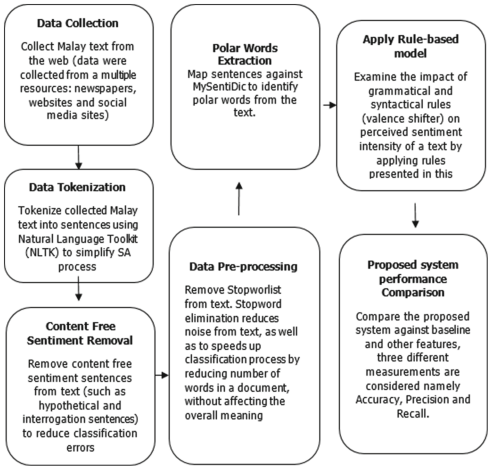


Fig. 1. Overview of the adopted method

We came to learn from the literature review conducted on sentiment analysis, that there are few errors related to sentiment classification when BOW method adopted. The most common ones are, one: classifying sentence sentiment into the opposite class, two: classifying a sentence-free sentiment into one of sentiment class (positive or negative), and three: classifying positive or negative sentence as being neutral (belonging to none of sentiment class). Consider the following example, “di mana saya bole cari restaurant yang terbaik?” using BOW, this sentence will be classified as positive due to the presence of positive word, in fact, this sentence is an interrogative sentence. Another example is hypothetical and conditional sentences, consider the following example, “jika saya jumpa baju yang cantik saya akan memebelinya” and “saya mahu makan makanan yang

sedap", again this sentences will be missed classified as a positive sentences using BOW due the presence of positive words "cantik", "terbaik" and "sedap".

As an effort to reduce the number of errors that might occur while classifying Malay opinionated text, we have identified two features that need to be considered as an early step to filter out the most common types of sentence-free sentiment. The first feature can be related to hypothetical conditional sentences. These kinds of sentences sometimes contain sentiment words that may confuse the classifier, a misclassification may occur if these sentences aren't properly handled. These types of error can be reduced by removing sentences that contains conditional words such as "`jika`", "`kalau`", "`andai kata`", "`sepatutnya`", etc., The second feature is the interrogation sentences, these types of sentences can be grouped into three groups, sentences that contain question mark "`?`", sentences that contain question word such as "`dimanakah`", "`apakah`", "`bagaimanakah`" how, and sentences that contain question lemma such as "bila", "apa", "bagaimana", "dimana", etc. Example of interrogation sentences, consider the following "`bagaimana harus kita memperbaiki keadaan?`", using the interrogation feature, this sentence can simply be identified as interrogation sentence that carries no sentiment orientation (Neutral) due to presences of lemma "bagaimana" and the question mark "?" at the end of the sentence.

3.1 Contextual Valence Shifter

There are a total of 10 rules used to tackle contrast and valence shifter including negation, intensifier, and diminisher. These rules are shown below:

POSITIVE rules:

1. Positive word or phrase: The frequency of positive words or phrase in sentence without the presence of valence shifter or modifier
2. Intensify positive: The frequency of positive words or phrase in sentence with presence of intensifier
3. Diminish negative: The frequency of negative words or phrase in sentence with presence of diminisher
4. Negate negative: The frequency of negative words or phrase preceded by negation
5. Shift negative: The frequency of negative words or phrase in sentence with presence of shifter

NEGATIVE rules:

6. Negative word or Phrase: The frequency of negative words or phrase in sentence without the presence of valence shifter or modifier
7. Intensify Negative: The frequency of negative words or phrase in sentence with presence of intensifier
8. Diminish positive: The frequency of positive words or phrase in sentence with presence of diminisher

9. Negate positive: The frequency of positive words or phrase preceded by negation
10. shift positive: The frequency of positive words or phrase in sentence with presence of shifter

Pseudo-code below explains the overall algorithm adopted to handle valence shifter including intensifier, diminisher, and negation in general. First, a sentence will be examined to reduce classification errors, if a sentence is a hypothetical or interrogation sentences, it will be ignored (pass), otherwise the sentence will be checked to identify any negation, if negation exists, word polar will be negated. Next, will check against intensifier existences, if exists, word polarity will either be increased or decreased based on presences of intensifier or diminisher (Fig. 2).

```
if sentiment word w is positive:
      if w preceded by negation:
            if w intensified:
     word count = -1 + intensifier_value
          else:  #not intensified
                word count = -1
           else: # not negated
              if w intensified:
     word count = +1 + intensifier_value
               if w diminished:
             word count = -1 #
                      else
                word count = +1
if sentiment word w is negative:
     if w preceded by negation:
           if w intensified:
   word count = +1 + intensifier_value
         else: #not intensified
              word count = +1
          else: #not negated
             if w intensified:
   word count = -1 + intensifier_value
                    else
               word count = -1
else : # if not a sentiment word
                  pass
```

Fig. 2. Pseudo-code to handle valence shifter

Negation. The most obvious shifter is negation. Negation flips valence of the term from positive to negative and vice versa. In other words, the combination of a positive evaluator with a negation turns the evaluation as a whole into a negative one. Inversely the combination of a negative evaluator with negation turn the sentence into a positive evaluation (e.g. "`dia bukan jahat` ."). In the Malay language, there are two words used as negation word, the word tidak and the word bukan as shown in Table 1.

According to Institute of Language and Literature (*Dewan Bahasa dan Pustaka*) the government body responsible for coordinating the use of the Malay language and literature in Malaysia, the word tidak is used to negate adjective phrase or verb phrase, on the other hand, the word bukan used to negate noun phrase (frasa nama) or preposition (sendi nama). To demonstrate the use of bukan and tidak, consider the following examples:

Table 1. List of negation words in malay along with proper usage

Negation	Use
Tidak	It is used to negate adjective phrase (*frasa adjektif*) or verb phrase(*frasa kerja*)
Bukan	It is used to negate noun phrase (f*rasa nama*) or preposition (*sendi nama*)

Example 1:

`Dia` **`bukan`** `pengkhianat Negara.`
`Pengkhianat` -1 combined with `bukan` ⟹ `bukan pengkhianat` +1

Example 2:

`Makanan di sini tidak sedap` (The food here is not delicious)
`sedap` +1 combined with `tidak` ⟹ `tidak sedap` -1

For simplicity purposes, positive words are underlined once, and assigned value +1, as for the negative words, they are underlined twice and assigned the value −1. Example 1 above, pengkhianat (traitor) is a negative word, when combined with negation, it invert its value to positive, this example applies rule number 4, where negative words or phrase preceded by negation is changed to positive.

Example 2 tidak sedap applies rule 9, where tidak invert the sentiment of the sentence from being positive due to presence of positive word (+1 for positive) sedap (delicious) to negative sentiment when combined with negation word tidak. As rule number 9 stated that the positive words or phrase preceded by negation is negative.

Previous works on the Malay language has proposed to only negate the first word that occurs after the negation word. However, this technique failed to handle more complex negation sentences, where the distance of polar word with negation word is more than one. Consider the following example, "`kempen ini tidak memberi apa-apa faedah kepada rakyat`", the distance from polar word and negation is 3, the existing technique which relies on the use of bi-gram fails to handle such negation. In the proposed system, we conducted an experiment on sentences that contain negation, we learned that the distance between polar word and the negation word can be up to 3 spaces away. To that, we have extended negation handling algorithm to handle negation up to three words. Figure 3 shows Pseudo-code to handle negation.

```
1. if sentiment word w is positive:
2.       flip w to negative
3.        word count = -1
4. else if sentiment word w is negative:
5.        flip w to positive
6.        word count = +1
```

Fig. 3. Pseudo-code for negation

Intensifier and Diminisher. Intensifier also known as a booster, modifies the sentiment intensity by increasing the intensity, for example very good is more intense than good alone. Diminisher, on the other hand, decreases the sentiment intensity, moderately interesting is less intense then interesting. In Malay language, intensifiers are known as *Kata Penguat*, they can be found before or after an adjective word. Table 2 lists down some of the commonly used intensifiers and diminished along with the position in which they may occur in a sentence.

Table 2. Example of malay intensifiers along with occurrence position

Intensifier	Example
Kata Penguat hadapan (front)	terlalu, paling, cukup, makin, agak, kurang
Kata penguat bebas (back)	sekali, benar, nian, betul
Kata penguat belakang (front or back)	amat, sangat, sungguh

An example of using intensifier and diminisher, consider the following examples:

Example 1: `Makanan di sini` **`sedap`** (Apply Rule 1)
Example 2: `Makanan di sini` **`terlalu`** `sedap` (Apply Rule 2)
Example 3: `Makanan di sini` **`kurang`** `sedap` (Apply Rule 8)

Example one will have a positive sentiment with score +1 due to the presence of positive word *sedap* which carries a value of +1. This example applies rule 1, where there are only positive words without the presence of valence shifter or modifier.

Example two will have a more intense positive sentiment which will have score +2, the intensifier `terlalu` intensified *sedap* from having the weight of +1 to having weight +2 without changing it's polar. This example applies rule number 2, where the value of positive is intensified due to the presence of intensifier "`terlalu`".

Example three, on the other hand, is shifted from being positive since the presence of sedap to negative since diminisher preceded the word sedap. This example applies rule 8. Based on rule 8, the sentence is negative if the positive word in a sentence is preceded by diminisher. Figure 4 shows Pseudo-code for intensifier handling.

```
1. if sentiment word w is positive:
2.       flip w to negative
3.        word count = 1+ intensifier_value
4. else if sentiment word w is negative:
5.        flip w to positive
6.        word count = -1 + intensifier_value
```

Fig. 4. Pseudo-code for intensifier

Connectors/Contrast. Contrast indicates a shift in sentiment polarity, where the sentiment of the text after the conjunction dominate the overall rating. The most widely used contrast conjunction in the Malay language is `tetapi/tapi`. There are other contrast conjunction words that have been used as well, this includes walaupun, bagaimanapun, sebaliknya, etc. To demonstrate the use of contrast, consider the following example:

Example: Makanan di sini sedap, tetapi perkhidmatan teruk.

Prediction based on BOW:

Sedap	+1
Teruk	– 1
Total score:	0

Adjusted prediction based on contrast rule:

tetapi) sedap	0
Teruk	-1
Total score	-1

As it can be noticed that contrast has a mixed sentiment, with the latter half of text manipulates the overall sentiment rating. Whatever sentiment preceded the contrast will be neutralized, as can be seen in the above example, the value of sedap is neutralized to 0 due the presence of contrast tetapi, which caused the overall of sentence score to be negative −1, compared to BOW prediction which failed to predict the real sentiment of sentence by returning neutral value 0.

The following subsection examines the impact of handling complex opinionated Malay text that contains negation, contrast, intensifier, and diminisher. We used a controlled set of experiments since the intention is to examine how well the proposed rule-based system performs with the presence of syntactic and grammatical features (valence shifter). Further details on experimental setup and evaluation of the proposed system is given in the following section.

4 Experimental Setup

Since the intention of this work is to investigate the impact of syntactic and grammatical features on Malay opinionated text, sarcasm sentences, anaphora and subjective sentences were not considered to avoid affecting the classification accuracy.

The first step involved in this experimental setup is the data collection. Initially data was to be fully collected from newspapers since the language investigated in this paper is Malay formal text, but due to the length nature of newspaper's articles of being long which may affect the annotation process in the future, few other resources have been explored as well in order to extract opinionated data, this includes www.cari.com.my a Malaysian lifestyle media forum, www.mesra.com.my, www.lazada.com.my and the well-known social media sites Facebook and Twitter. The data was collected from multiple domains to ensure the proposed rule-based system is robust, by being able to maintain stable prediction accuracy among multiple domains. These domains were randomly selected, this includes politics domain, economic domain, movie domain and fashion domain. Since the collected data is to be manually annotated, the length of a single review is set to maximum 160 characters similar to SMS size.

Next, the collected data is manually annotated and spelling checked by 7 native speakers to avoid word misspelling that may affect the accuracy performance. To ensure the accuracy of data annotation, the 7 annotators were carefully selected, all of the 7 annotators pose a credit in their Form 5 Bahasa Melayu SPM (Sijil Pelajaran Malaysia) examination. As a result, a total of 14,780 articles were successfully annotated. On a top of the automatically collected data, a total of 1500 contrived sentences were induced for the purpose to test the syntactic and grammatical impact on a text. The data is brought to its final with a total of 16,280 document, with 7798 positive reviews and 8482 negative reviews.

A total of 16 different features (F1 to F16) have been identified to examine the effect of valence shifter on Malay opinionated text as shown in Table 3. Each of the features used the Malay lexicon MySentiDic. To construct the baseline, we considered taking sentiment features that consists of frequency of sentiment word found in a document without taking into consideration its contextual polarity, in other words, baseline relies on the use of MySentiDic along with the widely used simple negation (negate the following word that occurs after negation word), baseline is denoted as F1.

To evaluate the 16 features, four types of measurement were used namely accuracy, precision, recall and F-measure. The measurements are based on the confusion matrix shown in Table 3. The following section discusses the result of the experiment conducted.

5 Result and Discussion

Accuracy is used to measure the total number of correct prediction produced by the system. Based on the results displayed in Fig. 5, feature F16 (based on all of the valence shifter types) recorders the highest accuracy improvement. This concludes

the accuracy experiment that taking into account all of the valence shifter types yields the best accuracy.

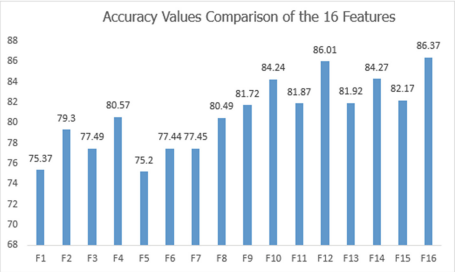


Fig. 5. Accuracy measurement

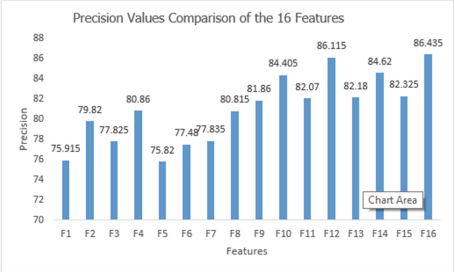


Fig. 6. Precision measurement

The precision measurement is carried out in order to measure the percentage of the correct prediction return by the system out of all prediction returned. In other words, out of the prediction provided by the system, how many percent are correct? Based on Fig. 6, feature F16 recorded the highest precision where increment is recorded to be 10.52%. This shows a number of false positive and false negative prediction were produced by the baseline, in which F16 managed to tackle (improve) by reducing the false classification.

Table 3. List of features considered for experiment

	A (Negation)	B (Intensifer)	C (Diminisher)	D (Contrast)
F1	0	0	0	0
F2	0	0	0	1
F3	0	0	1	0
F4	0	0	1	1
F5	0	1	0	0
F6	0	1	0	1
F7	0	1	1	0
F8	0	1	1	1
F9	1	0	0	0
F10	1	0	0	1
F11	1	0	1	0
F12	1	0	1	1
F13	1	1	0	0
F14	1	1	0	1
F15	1	1	1	0
F16	1	1	1	1

Recall measurement is used to measure the number of correct prediction produced by the system to the number of actual prediction (that are correct). As can be observed from Fig. 7, using F16 has improved recall by 10.84%.

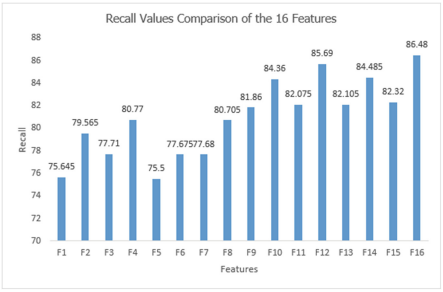


Fig. 7. Recall measurement

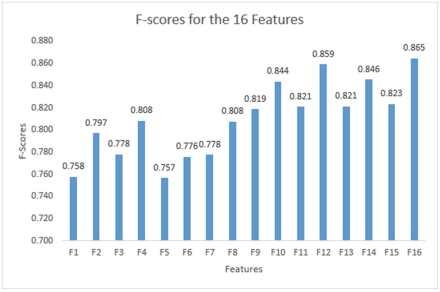


Fig. 8. F-scores measurement

In order to compare the precision and recall values, F-score also known as F-measure is adopted. The F-Score of the system is defined as the weighted harmonic mean of its precision and recall. Based on Fig. 8, F16 recorded the highest F-score of with a total of 0.865, this concludes the F-score experiment that F16 has a stable performance besides its high accuracy as recorded earlier.

6 Conclusion

In this chapter, a list of effective rules related to valence shifter handling has been introduced in order to improve Malay language opinionated text. Four types of valence shifter have been considered namely negation, intensifier, diminisher, and contrast. Among the valence shifter features discussed, and based on the experiments conducted, negation is shown to be the most common valence shifter type. It is also discovered that, by using all of the valence shifter types, it has produced the best accuracy increment. Another important discovery is using multiple valence shifters without considering negation performed lower compared to handling negation alone. As an overall conclusion, this chapter reconfirms the claim that valence shifter (negation, intensifier, diminisher, and contrast) are important features that need to be properly handled in order to improve performance of SA system in the Malay language. The proposed framework has shown a tremendous improvement in accuracy, recall and precision and F-score.

References

1. Polanyi, L., Zaenen, A. (n.d.): Contextual Valence Shifters
2. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **22**(2, 1), 1–167 (2012)

3. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011). https://doi.org/10.1162/COLI_a_00049
4. Dillard, L.: "I Can't Recommend This Paper Highly Enough": Valence-Shifted Sentences in Sentiment Classification. Master thesis (2007). http://www.tacoma.uw.edu/sites/default/files/global/documents/institute_tech/ldillard.pdf
5. Li, S., Yat, S., Lee, M., Chen, Y., Guodong, C.H.: Sentiment classification and polarity shifting, pp. 635–643, August 2010
6. Cambria, E., Hussain, A., Durrani, T., Havasi, C., Eckl, C., Munro, J.: Sentic computing for patient centered applications. In: IEEEICSP (2010)
7. Chekima, K., Alfred, R.: An automatic construction of malay stop words based on aggregation method. In: Berry, M., Hj. Mohamed, A., Yap, B. (eds) Soft Computing in Data Science. SCDS 2016. Communications in Computer and Information Science, vol. 652. Springer, Singapore (2016)
8. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, pp. 106–111 (2004)
9. Das, S.R., Chen, M.: Yahoo! for amazon: extracting market sentiment from stock message boards. In: Proceedings of the 8th Asia Pacific Finance Association Annual Conference, pp. 37–56 (2001)
10. Na, J.C., Sui, H., Khoo, C., Chan, S., Zhou, Y.: Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. Adv. Knowl. Organ. **9**, 49–54 (2004)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
12. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347–354 (2005)
13. Wilson, T., Wiebe, J., Hwa, R.: Recognizing strong and weak opinion clauses. Comput. Intell. **22**(2), 73–99 (2006)
14. Councill, I.G., McDonald, R., Velikovich, L.: What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pp. 51–59. Association for Computational Linguistics (2010)
15. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 339–346. Association for Computational Linguistics (2005)
16. Hatzivassiloglou, V., Wiebe, J.M.: Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of the 18th Conference on Computational Linguistics, vol. 1, pp. 299–305 (2000)
17. Morsy, S., Rafea, A.: Improving document level sentiment classification using contextual valence shifters. In: Natural Language Processing and Information Systems, pp. 253–258 (2012)
18. Musat, C., Trausan-Matu, S.: The impact of valence shifters on mining implicit economic opinions. In: Artificial Intelligence: Methodology, Systems, and Applications, pp. 131–140 (2010)
19. Li, S.Y.M., Lee, Y.C., Huang, C.R., Zhou, G.: Sentiment classification and polarity shifting. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 635–643. Association for Computational Linguistics (2010)

20. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD (2004)
21. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. **22**, 110–125 (2006)
22. Polanyi, L., Zaenen, A.: Contextual lexical valence shifters. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affecting Text (2004)
23. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2005)
24. Li, S., Wang, Z., Li, S., Huang, C.: Sentiment classification with polarity shifting detection. In: Proceedings of the International Conference on Asian Language Processing (2013)
25. Pang, B., Lee, L., Vaithyanathan, V.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods on Natural Language Processing, pp. 79–86 (2002)
26. Ding, X., Liu, B., Yu, P.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008 (2008)
27. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Comput. Linguist. **35**(3), 399–433 (2009)
28. Kennedyand, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. **22**(2), 110–125 (2006)
29. General Inquirer. http://www.wjh.harvard.edu/˜inquirer/homecat.htm
30. Large Movie Review Dataset. http://ai.stanford.edu/~amaas/data/sentiment/
31. Chekima, K., Alfred, R.: Non-english sentiment dictionary construction. Adv. Sci. Lett. **4**, 400–407 (2016). American Science Publisher

Proposed Scheme for Finger Vein Identification Based on Maximum Curvature and Directional Feature Extraction Using Discretization

Yuhanim Hani Yahaya[1(✉)], Siti Mariyam Shamsuddin[2,3(✉)], and Wong Yee Leng[2,3(✉)]

[1] Faculty of Science and Defence Technology,
Universiti Pertahanan Nasional Malaysia, 57000 Kuala Lumpur, Malaysia
yuhanim@upnm.edu.my
[2] UTM Big Data Centre, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
{mariyam,yeeleng}@utm.my
[3] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract. Finger vein identification has becoming increasingly noticeable biometric trait. The finger vein pattern provides high distinguishing features that are difficult to counterfeit because it resides underneath the finger skin. The performance of finger vein identification is highly depending on the meaningful extracted features from feature extraction process. Previous works have developed new methods for better feature extraction. However, most of the works focus on how to extract the individual features and not presenting the individual characteristic of finger vein patterns with systematic representation. Therefore, in this paper we propose an improved scheme of finger vein feature extraction method by adopting Discretization method. The finger vein feature extraction is based on combination of Maximum Curvature and Directional Feature (MCDF) feature extraction. After the extraction, the MCDF features value are then fed into Discretization module. The extracted features will be represented systematically by discriminatory feature values. The features values are informative enough to reflect the identity of an individual. The experimental result shows that the proposed scheme using Discretization produce identification accuracy performance above 95.0%. This shows that the proposed scheme produce good performance accuracy compared to non-discretized features.

Keywords: Finger vein · Discretization · Maximum curvature feature extraction
Directional feature extraction

1 Introduction

Biometric systems operate based on the measurement of user's physical or behavioral characteristics. It is basically a system which automatically distinguishes and recognizes a person as individual and unique through a combination of hardware and pattern recognition algorithms based on certain physiological or behavioral characteristics that are inherent to that person [1]. Biometric identification has received attentions of lots of researcher recent years. However, most of them have their own limitations. For example,

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 186–194, 2018.
https://doi.org/10.1007/978-981-10-8276-4_18

fingerprint is susceptible to counterfeit because fingerprint are easily exposed to make fake fingerprint mold. In addition, fingerprint with a scar or cur can lead to noisy biometric data [2]. Face recognition is one of the popular biometric techniques but faces major drawback in robustness against pose [3] and illumination invariance. Signature recognition is unstable over time when the individual is having stress or injury.

Considering these limitations, new biometric which depends on more robust features to enhance identification performance is required. Finger vein identification has gain increasing attention in biometrics research area [5, 6]. Finger vein identification verifies person's identity by recognizing the pattern of blood veins. The blood vessels transport the blood throughout the body to sustain metabolism using network arteries and veins. The use of such vascular structures in fingers has been investigated in the biometrics literature [4, 7] with high success. There are two unique characteristics of finger vein. First, it is hidden structure and it is difficult to steal the finger vein patterns, therefore the degree of security is high. Second, finger vein biometrics has strong antispoofing capabilities as it only ensure liveness detection.

The performance of finger vein identification largely depends on the feature extraction method. The selection of meaningful features is highly dependent on the extraction algorithm and it is crucial in finger vein identification. It leads to the main issue in identification which is how to acquire features from finger vein in order to reflect the identity of the individual. Many finger vein feature extraction methods have been proposed. Maximum curvature method proposed by Miura [11] is one of the popular methods as it showed the most promising EERs according to [19]. The method extract finger vein based on the curvature value on a cross section of a finger vein pattern. Nevertheless, the finger vein feature extracted using maximum curvature method consist of noises that can potentially decrease the identification performance. Furthermore, the values obtained provide low inter-variability and high intra-variability between the feature values of finger vein samples of an individual.

In this work, we proposed fusion feature extraction method combining maximum curvature (MC) and directional feature method with discretization method. The proposed method can enhance the identification accuracy. Our contributions consist of introducing feature discretization scheme to enhance the representation of individuality features of finger vein. Question on *How to produce unique features that can specifically identify an individual?* needs to be determined. This study will identify the additional process that can represents the finger vein features into a better representation and enhance the performance accuracy of finger vein identification. The discretization process also reduces the computation memory usage and manages the values of a feature more easily.

This paper is organized as follows: Sect. 2, presents the related work on finger vein feature extraction. Section 3, the proposed scheme for finger vein identification is described. Section 4 presents the experimental result and Sect. 5 concludes the paper.

2 Related Work

Finger vein biometric is where the vein structures randomly form a network and spread along the finger [9]. The finger vein is unique and difficult to forge as they are underneath the finger's skin surface. Finger vein pattern can only be captured by using near-infrared light-emitting diode (LED). The hemoglobin inside the blood will absorb near infrared LED light and makes the vein pattern appears as a dark line.

Based on the research works, finger vein feature extraction methods can be classified into three groups. There are vein pattern-based method, dimensionality reduction-based method and local binary-based method. Research work that implements vein pattern-based methods are repeated line tracking [10], maximum curvature [11], Gabor filter [2] and mean curvature [12]. In this method, the vein network will be segmented first and then the geometric shape structure of the finger vein network is extracted for matching. However, the segmentation results of low quality images are often unsatisfying. In dimensionality reduction-based methods, the image will be transformed into low-dimensional space to classify. Principal component analysis (PCA) [13] manifold learning [8] and two-dimensional principal component analysis (2DPCA) [14] were used in this method. Local binary-based method has been focus on local area and the extracted features are in binary formation. Local binary pattern (LBP) [15] can reflect the texture of the finger vein very well. Nevertheless, Rosdi [15] also demonstrates local line binary pattern (LLBP) can perform better accuracy than LBP as LLBP have a good feature representation method. Nevertheless, LLBP only extract horizontal and vertical line patterns.

It is important to obtain good and systematic features in order to achieve accurate identification. Large variation in intra-class and low in inter-class can leads to a poor identification performance due to the various features representing an individual. Extracted finger vein features should be represented with a standard representation in order to improve the variation between features for intra-class and interclass. The following section explains the proposed scheme for finger vein identification.

3 The Proposed Method

The proposed scheme for finger vein identification is shown in Fig. 1. It is desirable to come up with an approach that can extract features of finger vein, which can sustain the individuality features in data representation. To develop efficient finger vein identification, a primary research question in this study is how to produce the distinctive features that can identify an individual. This relate to how the extracted features systematically differentiate an individual based on the finger vein samples. Therefore, in this study, discretization method is implemented into the finger vein identification as presented in Fig. 1. The feature extraction in this framework is the combination of maximum curvature method with directional-based feature extraction (MCDF) method. Extracted feature vectors from MCDF method are the input for discretization module.

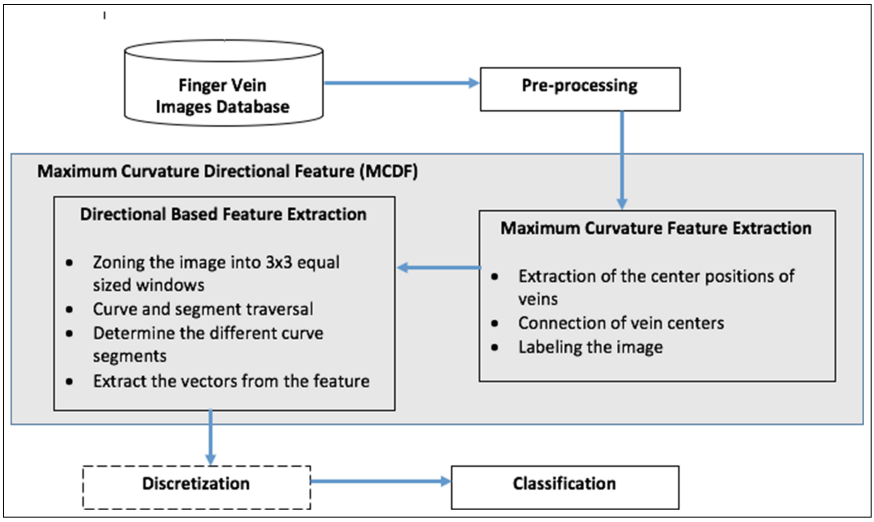


Fig. 1. The proposed finger vein identification framework

Maximum curvature was first proposed by Miura and Nagasaka [11] and this method was chosen as it showed promising EERs which is an indication for good finger vein extraction. Directional-based feature extraction proposed by [18] is also exploited in this study. According to the authors, this method extracts features that mainly focus on the different basic lines. This approach uses traversal process in each lines and return features vectors as its output.

3.1 Maximum Curvature Feature Extraction

This method has been developed to robustly extract the features of finger veins by extracting the center lines of the veins image. Centre lines are obtained by observing the curvatures of the cross sectional profiles are locally maximal positioned. The centrelines of the veins can be extracted consistently without being affected by the variations in the width and brightness of the vein. This method of finding the maximum curvature positions is against the variation in width and brightness of the vein. The positions are interconnected with each other and finally the vein pattern is detected [11].

3.2 Directional-Based Feature Extraction

This feature extraction is originally developed by Blumenstein et al. [18] for character recognition. This algorithm extracts geometric features that mainly focus on the different basic lines. This approach uses traversal process in each lines and return feature vectors as its output. As acknowledged, finger vein patterns related to the curve vein patterns and junctions. As such, this method was chosen as feature extraction method because it works well with multi-direction structures pattern.

3.3 Discretization

According to Leng and Shamsuddin [16], discretization plays an important role in the proposed scheme of finger vein identification, where its main role is to partition the continuous features into categories of discrete features. Discrete features can enhanced the performance of classification because discrete features are reduced and simplified with only relevant features. The main goal of discretization is to handle useful information from actual dataset without eliminate the original characteristics. The discretization process will define the interval of finger vein image based on the feature values extracted from MCDF for each individual. Then the range of data will be divided by the size of interval. The representation value will be determined by one interval. Feature values that have the same representation value will be grouped into the same interval. The sequence steps of discretization algorithm is presented as follows:

Defining Interval. To acquire an interval, the highest (maximum) and lowest (minimum) feature values of each individual are selected to estimate the interval for each bin. Number of bins for each individual is based on the number of features extracted from MCDF.

```
For an individual
    Min = min_features; Max = max_features;
    No_bin = no_MCDF_features;
    Interval = (Max-Min)/No_bin
```

In this study, nine bins are created for each individual as there are nine features values extracted from MCDF.

Define Representation Value for Each Bin. To define the representation value, lower and upper value of each interval is determined. Then the representation value for each bin will be determined through (upper-lower)/2 calculations.

```
For each bin
    Find lower and upper value of interval;
    RepValue = (upper-lower)/2;
```

Classify the Features into Interval. The process of assigning features into the same interval begain with the first feature and repeated until the end of the features. If the feature is in the range of the interval, then it will be initialized as discretized data with representation value of the individual.

```
For (1 to no_MCDF_features)
  For each bin
  If (feature in range of interval)
  Dis_Feature = RepValue;
```

The main motivation behind this proposed scheme is to maximize inter-class variability for all finger vein samples that do not belong to the same individual class. The issue of dimensionality caused by overlapping features can be avoided by representing the features into a set of interval.

4 Results and Discussion

The experimental results of the study used the SDUMLA-HMT finger vein dataset taken from Machine Learning and Data Mining Lab [17]. It consists of 106 subjects and composed of 3,816 finger images from index finger, middle finger and ring finger for both hands. The image is in .bmp format with 320 × 240 pixels in size.

Figure 2 presents the samples of the extracted features from MCDF method. The pre-discretized dataset (before the execution using discretization algorithm) consists of nine extracted feature vectors. The last column of the datasets represents the individual id. In Fig. 2 shows two individual that has low-inter features variability and high-inter features variability. The results after Discretization process are shown in Fig. 3. The bold values are the variances of features of the individual and it is represented with standard values as domain values (DV) of individual. DV represents the unique identity value of individuality where each individual has its own finger vein structures.

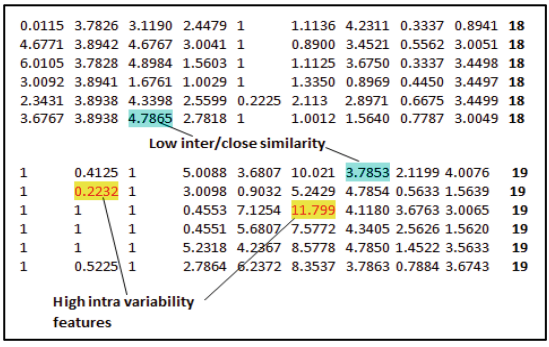


Fig. 2. Pre-discretized finger vein

Discretized Finger Vein Features of Individual No. 9

Bin	Lower	Upper	Representation Value
0	0.222506	1.15056	0.464029
1	1.15056	2.07862	1.61459
2	2.07862	3.00668	2.54265
3	3.00668	3.93474	3.47071
4	3.93474	4.86279	4.39876
5	4.86279	5.79085	5.32682
6	5.79085	6.71891	6.25488
7	6.71891	7.64696	7.18294
8	7.64696	8.57502	8.11099

Discretize Data

0.464029	**0.464029**	**0.464029**	**0.464029**	**0.464029**	5.32682	0.441234	5.326829	**0.464029**	Domain values (DV) for individual 9 is **0.464029**
0.464029	**0.464029**	**0.464029**	3.470713	8.110999	**0.464029**	0.54673	**0.464029**	**0.464029**	
0.464029	**0.464029**	**0.464029**	8.110981	6.25488	**0.464029**	4.89122	**0.464029**	**0.464029**	
0.464029	**0.464029**	**0.464029**	6.253112	1.61459	**0.464029**	2.35672	4.39876	3.43231	
0.464029	**0.464029**	**0.464029**	6.865431	4.39876	**0.464029**	1.74378	2.54265	2.41897	
0.464029	**0.464029**	**0.464029**	1.33332	**0.464029**	3.47071	1.65432	1.61459	1.23178	

Fig. 3. Domain value for individual no. 9

Roughset Toolkit (ROSETTA) is used to classify the pre-discretized and post-discretized dataset and the reducer algorithm is used to find reduct approximation in a decision table. Three built-in ROSETTA reducer algorithms used in this study are Holte1R Algorithm, Exhaustive Algorithm and Genetic Algorithm. The results for individual identification performance using pre-discretized and post-discretized datasets are reported in Table 1. From the result, it shows that the post-discretization dataset successfully achieved identification accuracy performance above 95.0%. On the other hand, the pre-discretized datasets report worse performance, which provides identification accuracy rates below 75%.

Table 1. Finger vein identification accuracy

ROSETTA built-in methods on reductions	Data types	70% train data 30% test data	60% train data 40% test data	50% train data 50% test data	Average accuracy
Holte 1R Algorithm	Pre-Dis	66.77	75.00	70.00	70.57
	Post-Dis	**100**	**100**	**97.5**	**99.17**
Genetic Algorithm	Pre-Dis	66.77	75.00	70.00	70.57
	Post-Dis	**100**	**100**	**97.5**	**99.17**
Exhaustive Algorithm	Pre-Dis	66.77	75.00	70.00	70.57
	Post-Dis	**100**	**100**	**97.5**	**99.17**

Table 2 illustrates the comparison between the proposed scheme and other feature extraction methods in terms of equal error rate (EER). The method proposed by Munalih et al. [9] extracts the finger vein features by using the combination of enhanced maximum curvature (EMC) with histogram of oriented gradient (HOG) descriptor. Their best EER result was 0.14%. Meanwhile Cao et al. [7] method uses minutiae points and curve

tracing as feature extractor. The obtained EER result was 5.82%. Nevertheless, our proposed scheme shows promising results with EER as low as 0.0121%. The result shows that the proposed scheme is better than the other methods in finger vein identification.

Table 2. Result of comparison proposed scheme and other finger vein feature extraction method

Methods	EER (%)
EMC+HOG descriptor	0.14
Minutiae points+curve tracing	5.82
Proposed Scheme	**0.0121**

Based on the two experiments conducted, it presents that the proposed scheme could enhance the accuracy of the identification performance of finger vein. It is also prove that the discretization method is able to overcome the intra-features and inter-features variability.

5 Conclusion

In this study, the propose scheme for finger vein identification using MCDF and discretization have proven that the implementation of dicretization can enhanced the identification accuracy performance results. The post-discretized data successfully achieved identification accuracy of above 95.0%. The proposed scheme also produces best results in terms of EER in comparison to the other methods. It shows that the discretized feature still carries the uniqueness features of an individual without losing important information and characteristics of individuals.

Acknowledgement. Authors would especially like to thank Ministry of Higher Education (MOHE) and UTM Big Data Centre for their support and contributions to this research study.

References

1. Newman, R.: Security and Access Control Using Biometric Technologies. Cengage Learning (2010)
2. Kumar, A., Zhou, Y.B.: Human identification using finger images. IEEE Trans. Image Process. **21**(14), 2228–2244 (2012)
3. Zhang, X., Gao, Y.: Face recognition across pose: a review. Pattern Recognit. Lett. **42**(11), 2876–2896 (2009)
4. Nasir, S.E., Shamsuddin, S.M.: Proposed scheme for palm vein recognition based on linear discrimination analysis and nearest neighbour classifier. In: Proceedings of the International Symposium on Biometrics and Security Technologies, pp. 67–72 (2014)
5. Khalid, S.I., Radzi, S.A., Saad, N.M., Hamid, N.A., Saad, W.H.: Finger vein biometrics identification approaches. Indian J. Sci. Technol. **9**(32), 1–8 (2016)
6. Lee, E.C., Jung, H., Kim, D.: New finger biometric method using near infrared imaging. Sensors **11**, 2319–2333 (2011)

7. Cao, D., Yang, J., Shi, Y., Xu, C.: Structure feature extraction for finger vein recognition. In: Second Asian Conference on Pattern Recognition, pp. 567–571 (2013)
8. Liu, Z., Yin, Y.L., Wang, H., Song, S., Li, Q.: Finger vein recognition with manifold learning. J. Netw. Comput. Appl. **33**(3), 275–282 (2013)
9. Munalih, A.S., Thian, S.O., Andrew, B.J., Connie, T.: Enhanced maximum curvature descriptors for finger vein verification. Multimed. Tools Appl. **76**(5), 6859–6887 (2016). Springer
10. Muira, N., Nagasaka, A.: Feature extraction of finger-vein pattern based on repeated line tracking and its application to personal identification. Mach. Vis. Appl. **15**(4), 194–203 (2004)
11. Muira, N., Nagasaka, A.: Extraction of finger vein pattern using maximum curvature points in image profiles. IEICE Trans. Inf. Syst. **E90-D**(8), 1185–1194 (2005)
12. Song, W., Kim, T., Kim, H.C., Choi, J.H., Kong, H.J., Lee, S.R.: A finger vein verification system using mean curvature. Pattern Recognit. Lett. **32**(11), 1541–1547 (2011)
13. Wu, J.D., Liu, C.T.: Finger vein pattern identification using principal component analysis and neural network technique. Expert Syst. Appl. **38**(5), 5423–5427 (2011)
14. Yang, G.P., Xi, X.M., Yin, Y.L.: Finger vein recognition based on (2D) PCA and metric learning. J. BioMed. Biotechnol. **2012**, 1–9 (2012)
15. Rosdim, B.A., Shing, C.W., Suandi, S.A.: Finger vein recognition using local line binary pattern. Sensors **11**, 11357–11371 (2011)
16. Leng, W.Y., Shamsuddin, S.M.: Fingerprint identification using discretization technique. Int. J. Comput. Electr. Autom. Control Inf. Eng. **6**(2), 240–248 (2012)
17. Homologous Multi-Modal Traits. SDUMLA-HMT. http://mla.sdu.edu.cn/sdumla-hmt
18. Blumenstein, M., Verma, B.K., Basli, H.: A novel feature extraction technique for the recognition of segmented handwritten characters. In: 7th International Conference on Document Analysis and Recognition, pp. 137–141 (2003)
19. Ton, B.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: International Conference on Biometrics Compendium, pp. 1–5 (2013)

Word-Based Classification of Imagined Speech Using EEG

Noramiza Hashim(✉), Aziah Ali, and Wan-Noorshahida Mohd-Isa

Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
noramiza.hashim@mmu.edu.my

Abstract. Imagined speech is a process where a person imagines the sound of words without moving any of his or her muscles to actually say the word. If the brain signals of a person imagining the speech can be used to recognize the actual words intended to be spoken, this could be a huge step towards helping people with physical disabilities such as locked-in syndrome to have effective communication with others. This can also prove to be useful in situation where visual or audible communication is undesirable, for instance in military situation. Recent advancement in technologies and devices for capturing brain signals, particularly electroencephalogram (EEG), has made the research in recognizing imagined speech possible. While these are still in early years, published studies have shown promising results in this particular area of research. Current approaches in recognizing imagined speech can generally be divided into two, syllable-based and word-based. In this paper, we proposed a simple word-based approach using Mel Frequency Cepstral Coefficients (MFCC) and k-Nearest Neighbor (k-NN) towards recognizing two simple words using EEG signals. Despite its simplicity, the results obtained show some improvements to other studies based on dry EEG electrode device.

Keywords: Imagined speech · EEG · k-Nearest Neighbor
Mel Frequency Cepstral Coefficients

1 Introduction

We regularly use verbal and non-verbal communication and it is very vital in our daily life. However, in situation where visual or audible communication is undesirable and for persons unable to speak because of physical disabilities such as locked-in syndrome, or advanced amyotrophic lateral sclerosis (ALS), unspoken or imagined speech might be used to help them to communicate.

Imagined speech refers to imagining the sound of a given word without the intentional movements from the lips or tongue. The signals are produced through anticipation of intended speech information due to neural activities in the course of speech production. Researchers are exploring the idea of interpreting information from such signals, captured through devices connected to the brain such as the electroencephalography (EEG) electrodes. Although research in this area are still in its infancy, the potential of imagined speech applications can be seen in many areas. In medical field, patients

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 195–204, 2018.
https://doi.org/10.1007/978-981-10-8276-4_19

suffering from locked-in syndrome or ALS can use imagined speech to communication their basic wants and needs. It can also be useful for military applications where non-audible and non-visual communications can offer a better and/or safer alternative to verbal communication. For game applications, imagined speech can be used to provide control and enhance users' experience.

2 Related Works

A comprehensive overview of the different types of technology used for silent or imagined speech has been presented by [1], which includes not only EEG, but also electromagnetic articulography (EMA), surface electromyography (sEMG) and electrocorticography (ECoG). Among these, EEG presents a particular interest because it is non-invasive, relatively simple, economical, and insensitive to environments with large amounts of audible noise.

Although directly inferring a person's words or sentences is not feasible currently due to limited understanding of how human brain generates and interprets speech related signals, researchers have shown that EEG signals can be used to classify vowels or a handful of words with promising results. Works using imagined speech can be divided into two approaches: by syllables and by words.

In regards to syllable-based approach, [2] studied imagined vocalizing and mouthing of /a/ and /u/. Their results show that motor cortex activation associated with imaginary vowel can be classified using common spatial pattern and a non-linear classifier. However, their results also suggested that subsequent classification may have picked up mostly muscle movements instead of the imagined speech itself. Due to this, [3] investigated imagined speech without muscle movement with /ba/ and /ku/, covertly spoken at a specific rhythms, provided by audio cues. This vocabulary was chosen due to their lack of semantic meaning. The features were calculated based on Hilbert envelope followed by the creation of an averaging class template to be matched with filters for classification. The reported average precisions range from 52% to 74%. [4] used the same database but employed a different classification approach where autoregressive coefficients were used as features followed by k-Nearest Neighbor (k-NN) classifier, with a maximum of 81% accuracy obtained.

The work proposed by [5] examined different classifiers such as k-NN, Naive Bayes and neural network. Features were extracted using mean feature extractor and discrete wavelet transform. An overall accuracy of 81.9% was obtained using neural network as for a four-class classifier. In [6], features derived from average power, mean, variance and standard deviation were fed into a back propagation neural network. An average result of 44% were obtained. Classification of pairwise syllables were explored in [7], where signal features such as mean, variance, standard deviation and skewness were calculated followed by sparse regression model-based feature selection before it was used for classification using a feed forward neural network. The accuracy up to 99.41% was obtained for classification on the best pair of syllables.

Under the word-based approach, a study performed by [8] managed to obtain good recognition performance (42%). However, [9] suggested that it might have resulted from

temporal correlated artefacts in the brainwaves since the words were presented in blocks. This was justified by the fact that block data would contain less noise and block mode made it easier to think about words in a consistent way. In their study, five words without high semantic content were chosen. This study also concluded that block mode obtained results were above chance (45.5%) and other modes (reordered block, random, sequential) only performed at chance level.

The previously mentioned studies used wet electrodes where 16 EEG channels using 128 cap montage were used. More recent studies by [10, 11] used dry electrodes devices, which offered more portability and ease of use. The authors in [10] used Emotiv, a 14-channels dry electrodes EEG device and proposed the use of discrete wavelet transform with four different classifiers; Naive Bayes (NB), Random Forests (RF), support vector machine (SVM), and Bagging-RF, resulting in above chance results in classifying five words with semantic meanings. In [11], Neurosky, a single dry electrode device was used in an attempt to classify two words; /yes/ and /no/. An average accuracy of 56% was obtained for offline testing.

These studies have suggested that there is some distinctive information contained in the EEG data for different imagined speech. However, it can be seen that classifications of imagined speech based on syllables have higher accuracy rates compared to classifications of imagined speech based on words.

Our work also attempts to classify imagined speech based on words using the dry electrode device Emotiv. The next section discusses the methodology employed for this work. Then, we describe the different experiments conducted in order to optimize the performance, including some discussions of the results obtained followed lastly by the conclusion.

3 Methodology

3.1 Data Collection

In this study, we attempt to classify two words, Yes and No, using Emotive Epoc ®, a 14-channels sensor device. Four volunteers have agreed to participate as subjects, all healthy males with age between 23 and 25 years old. Only one type of gender is chosen and the age of the subjects is kept more or less consistent in order to reduce any possible influence or variation in the brain signal patterns due to these two factors. As a comparison, the number of subjects used in similar EEG experiments are two [12], five [7] and seven [11].

Subjects are briefed about the overall experiment, where visual cue showing the start and the end of the experiment session will be shown to them. The experiment is conducted in a quiet environment. A subject is seated on a chair and facing a screen while visual cue of the word is displayed on a white background (Fig. 1).

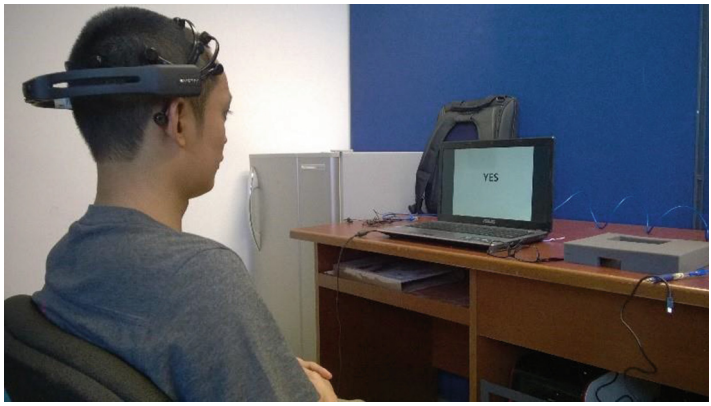


Fig. 1. Experiment environment

Without any mouthing or physical movements, subjects need to imagine the shown word. There are five sessions for each subject. Each subject is required to imagine the words ten times for each session, with a break in between. Then, a longer break time will be given to subject between each word and at the end of each session. The duration for each session are illustrated in Fig. 2. For one subject, there are 50 trials for each word. In total, there are 400 trials for four different subjects.

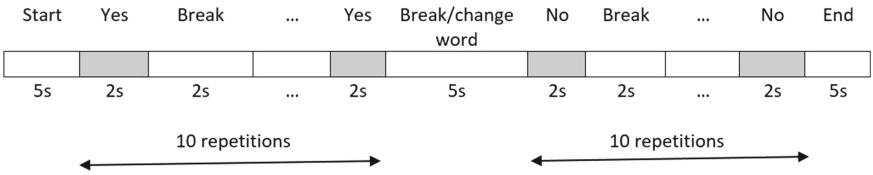


Fig. 2. Duration of a trial

3.2 Feature Extraction

Data from the Emotiv neuroheadset is sampled at 128 Hz for each sensor with Analog to Digital Conversion of the data. Digital notch filters at 50 Hz and 60 Hz and a built-in digital 5th order sinc filter are also integrated in the headset, which serve to remove the power line frequency. The signals are high pass and low pass filtered with an effective bandwidth of 0.16 Hz to 43 Hz [13]. The raw data is transmitted in packets from the neuroheadset to a USB dongle.

Based on the specification of Emotiv neuroheadset, each sensor reads the relative value of voltage compared to two reference electrodes. Therefore, each value of the data vector has approximatively 4200 μV added to it. To remove this reference component, a high pass Butterworth filter is applied.

For each trial, the Mel Frequency Cepstral Coefficients (MFCC) is extracted. [14] states that EEG signal can be considered as quasi-stationery; it is a slowly time varying

signal over a short time period and non-stationary over a long time period. Due to this reason, MFCC, a feature extraction methods used to extract speech features can be used for extracting features from EEG signals.

For this, the signal is analyzed in short overlapping window frame. Each frame is defined to be a product of a shifted window with the speech sequence. A mel is a unit of pitch defined so that pairs of sounds which are equidistance in pitch are separated by the same number of mels. The mel frequency is computed using [14]:

$$Mel(f) = 2595\, log_{10}\left(1 + \frac{f}{100}\right) \tag{1}$$

A filterbank that models the human ear ability to resolve frequencies in logarithmic scale is computed. It is an array of band-pass filters, can be rectangular or triangular and equally spaced along the mel-scale. MFCCs are calculated from the log filterbank using [14]:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{n} m_j \cos\left(\frac{\pi_i}{N}(j - 0.5)\right) \tag{2}$$

where m_j is the filterbank amplitude, N is the number of filterbank channel and c_i are the cepstral coefficients.

3.3 Classification

The feature vectors obtained from MFCC are then used for k-NN classification. K-NN is widely used as the baseline classifier for many domains due to its simplicity and ease of implementation.

For a given testing sample, k-NN algorithm searches for the nearest neighbour between the testing sample and all of the training data in order to produce the output. The nearest neighbour is determined by calculating the minimum distance between the points in the testing sample and the points in the training data. Most commonly, the Euclidean distance metric is used. Other distance functions can also be used such as the cosine similarity measure and correlation [15].

For two vectors X and Y having M dimensional feature space, the distance functions are defined as follows:

$$Euclidean\ distance: \quad dist(X, Y) = \sqrt{\sum_{i=1}^{M} \frac{\left(x_i - y_i\right)^2}{M}} \tag{3}$$

$$Cosine\ similarity\ measure: \quad sim(X, Y) = \frac{\vec{X}.\vec{Y}}{\left|\vec{X}\right|\left|\vec{Y}\right|} \tag{4}$$

$$Correlation: \quad corr(X,Y) = \frac{\sum_{i=1}^{M}(x_i - y_i)(y_i - u_i)}{\sqrt{\sum_{i=1}^{M}(x_i - y_i)^2 (y_i - u_i)^2}} \tag{5}$$

In [15], the authors have concluded that different distance functions will influence the results of the k-NN classifier used on different types of medical data i.e. categorical, numerical and mixed types. For most of their numerical datasets, Euclidean distance has been reported to provide the best performance. As EEG data may also be categorized as numerical data, we would like investigate the effect of different distance functions on k-NN classification result. Therefore, in our work, we will compare the classification of the k-NN classifier using the three distance functions discussed previously.

4 Results and Discussion

4.1 Electrodes Selection

Studies have shown that imagined speech activates two areas in the brain; Wernicke's and Broca's, two essential language areas involved in speech comprehension and production [16]. For Emotiv neuroheadset, these areas have been identified and selected, where only the left side [10] and both left and right sides [17] are considered, respectively. A study by [12] suggested that the most prominent electrode positions were located not only on the left side of the brain nearest to Broca's and Wernicke's areas, but also nearest to the motor cortex.

In our work, we experiment with different electrodes selection based on the Emotiv electrode positions in Fig. 3. Results in Table 1 show that the best accuracy is obtained from 6 electrodes selected from the left side of the brain. This result supports the study by [12], which shows that prominent electrode positions are located in the left side of the brain and cover areas close to motor cortex, Broca's and Wernicke's areas. These 6 electrodes will be chosen for the rest of our experiment.

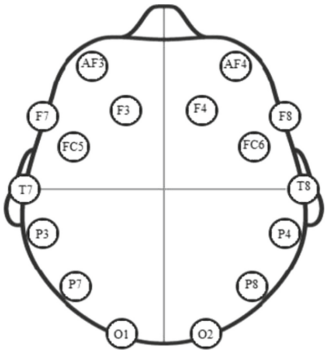


Fig. 3. Emotiv electrode positions [13]

Table 1. Average accuracies for different electrodes selection cases.

Area	Brain side	Electrodes	Accuracy
Wernicke's and Broca's	Left	F7, FC5, T7, P7	52.99%
Wernicke's and Broca's	Left and right	F7, F8, FC5, FC6, T7, T8	49.54%
Motor cortex, Wernicke's and Broca's	Left	AF3, F7, F3, FC5, T7, P7	53.53%

4.2 Distance Functions

We investigate the effect of using different distance function to evaluate the nearest neighbour distance of the k-NN classifier. For this, three types of distance functions are considered; Euclidean, cosine and correlation distances for evaluation.

Figure 4 shows the classification accuracy using different values of k, where k represents the number of nearest point considered for classification decision. From the results in Fig. 4, it can be seen that k = 3 provides the best result for the distance functions, with an accuracy of 63%. Euclidean and cosine distances produce the worst and best performances, respectively.

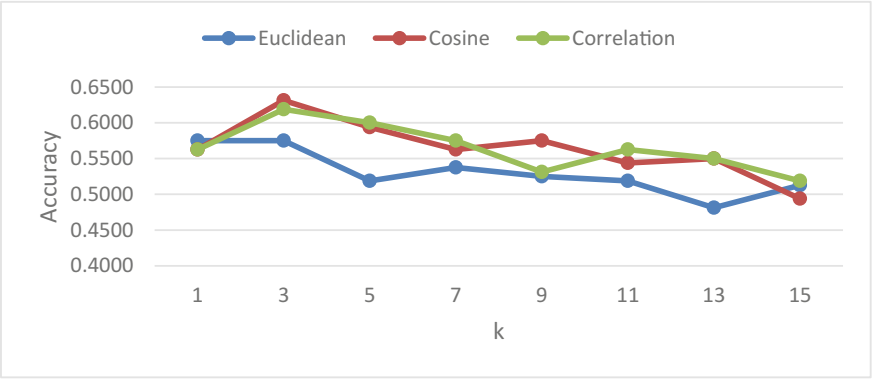


Fig. 4. Classification accuracies of using different distance functions.

4.3 Subjects

Table 2 shows the classification accuracies of the word 'Yes' and 'No' separately and for each subject individually. For some subjects, the accuracy for one word differs largely from the other word, i.e. for subject 4, the classification accuracy for the word 'No' is at 85% compared to 25% for the word 'Yes'. The word 'No' is also classified with slightly higher accuracy (57.5%) than the word 'Yes' (51.25%). On the other hand, the average classification accuracies when using both words in the testing set are quite similar for all subjects.

Table 2. Classification accuracies for different subjects and words.

	Yes	No	Subject average
S1	0.55	0.60	0.58
S2	0.70	0.35	0.53
S3	0.55	0.50	0.53
S4	0.25	0.85	0.55
Word average	0.51	0.58	

4.4 Average Accuracies

For testing, the dataset were divided in a 60%–40% train-test ratios. Five different sample sets were tested and the results are shown in Table 3. The average accuracy of the five sample sets is 58%. When compared to similar works based on dry electrode EEG devices and for word-based imagined speech classification, our result shows a slight improvement (refer to Table 4).

Table 3. Classification accuracy of different sample sets.

Sample set	Set1	Set2	Set3	Set4	Set5
Accuracy	0.63	0.55	0.58	0.57	0.58

Table 4. Comparison of our work to related works on word-based EEG classification of imagined speech.

Research work	Scope of work	EEG device	Average accuracy
Toward a silent speech interface based on unspoken speech [10]	5 words /arriba/, /abajo/, /izquierda/, /derecha/, /seleccionar/	Emotiv	Above chance (20%)
Recognition of unspoken words using electrode electroencephalographic signals [11]	2 words /yes/ & /no/	Neurosky	56%
Our work	2 words /yes/ & /no/	Emotiv	58%

5 Conclusion

Our work attempts to classify two word-based imagined speech from EEG signals. The feature extracted is MFCC, a common function used in speech processing, followed by k-NN for classification. Despite its simplicity, the results obtained show an improvement when compared to similar works, with an average accuracy of 58%. Our experiment has concluded that the brain areas that contain the most prominent electrodes for Emotiv neuroheadset, based on the 10–20 electrode placement system, are located around the

motor cortex, Broca's and Wernicke's areas. We have also compared three distance functions for the k-NN classifier and found that the cosine distance produces the best result, with a maximum of 63% accuracy.

In our work, only 4 healthy male subjects have participated as volunteers. As seen in Sect. 4.3, there is a large variation of classification accuracy among the subjects, which can be influenced by the subject's brain activity during the experiment as well as artifacts in the signal. Further investigation on these factors might improve the final results.

In future, it might be interesting to include more variation to the subjects groups. Subjects that are unable to communicate verbally due to some medical condition might not have similar brain signal for imagined speech. Also, is the choice of optimum electrode positions for left handed and right handed subjects the same? Different feature extraction and classification methods need to be explored in order to identify the ones that are most efficient and accurate.

Although research on word-based EEG signal of imagined speech have not managed to have the same success as its syllable-based counterpart, it has been shown that EEG signals do contain some discriminative components that can be used for classification purpose. A low cost system that uses dry EEG electrodes such as Emotiv and Neurosky and capable of deciphering simple intentions from its user will be most beneficial especially to patients with medical condition such as ALS.

Acknowledgement. The author's wishes to thank MOHE Fundamentals Research Grant (FRGS) with grant no FRGS/2/2014/ICT07/MMU/03/1 obtained through Multimedia University for supporting this research.

References

1. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Commun. **52**(4), 270–287 (2010)
2. DaSalla, C.S.: Single-trial classification of vowel speech imagery using common spatial patterns. Neural Netw. **22**(9), 1334–1339 (2009)
3. Zmura, M.D., Deng, S., Lappas, T., Thorpe, S., Srinivasan, R.: Toward EEG sensing of imagined speech, pp. 40–48 (2009)
4. Brigham, K., Kumar, B.V.: Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In: 2010 4th International Conference on Bioinformatics Biomedical Engineering (iCBBE), pp. 1–4 (2010)
5. Barak, O., Nishith, K., Chopra, M.: Classifying Syllables in Imagined Speech using EEG Data, pp. 1–5 (2014)
6. Kamalakkannan, R., Rajkumar, R.: Imagined speech classification using EEG. Adv. Biomed. Sci. Eng. **1**(2), 20–32 (2014)
7. Min, B., Kim, J., Park, H.J., Lee, B.: Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. Biomed. Res. Int. **2016**, 1–11 (2016)
8. Wester, M., Schultz, T.: Unspoken Speech: Speech Recognition Based On Electroencephalography (Master Thesis). Institut für Theoretische Informatik, Universität Karlsruhe (TH), Karlsruhe (2006)

9. Porbadnigk, A., Wester, M., Calliess, J., Schultz, T.: EEG-based speech recognition: impact of temporal effects. In: International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS) (2009)
10. García, A.A.T., García, C.A.R., Pineda, L.V.: Toward a silent speech interface based on unspoken speech. In: Van Huffel, S., Correia, C.M.B.A., Fred, A.L.N., Gamboa, H. (eds.) BIOSIGNALS, pp. 370–373. SciTePress (2012)
11. Salama, M., Elsherif, L., Lashin, H., Gamal, T.: Recognition of unspoken words using electrode electroencephalograhic signals. In: The Sixth International Conference on Advanced Cognitive Technologies and Applications, pp. 51–55 (2014)
12. Riaz, A., Akhtar, S., Iftikhar, S., Khan, A.A., Salman, A.: Inter comparison of classification techniques for vowel speech imagery using EEG sensors. In: 2014 2nd International Conference on Systems and Informatics, ICSAI 2014, pp. 712–717 (2014)
13. Moazzami, M.-M.: EEG Signal Processing in Brain-Computer Interface (Master Thesis). Michigan State University (2012)
14. Nguyen, P., Tran, D., Huang, X., Sharma, D.: A proposed feature extraction method for EEG-based person identification. In: International Conference on Artificial Intelligence (2012)
15. Hu, L.-Y., Huang, M.-W., Ke, S.-W., Tsai, C.-F.: The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus **5**(1), 1304 (2016)
16. Martin, S., Brunner, P., Iturrate, I., Millán, J.D.R., Schalk, G., Knight, R.T., Pasley, B.N.: Word Pair classification during imagined speech using direct brain recordings. Sci. Rep. **6**, 25803 (2016)
17. Rojas, D.A., Ramos, O.L., Saby, J.E.: Recognition of Spanish vowels through imagined speech by using spectral analysis and SVM. J. Inf. Hiding Multimed. Signal Process. **7**(4), 889–897 (2016)

Sentiment Analysis of Malay Social Media Text

Khalifa Chekima(✉) and Rayner Alfred(✉)

Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
k.chekima@gmail.com, ralfred@ums.edu.my

Abstract. Since early 2000, sentiment analysis has grown to be one of the most active research areas in Natural Language Processing (NLP). Since then, researchers have shown a tremendous interest in building automated Sentiment analysis applications for English language and non-English languages such as Arabic Language, French language, Deutsch language, Chinese language, Italian language, etc. Yet, very limited researches have been attributed to Malay opinionated social media text despite the big number of Malay native speakers which recorded to be approximately 215 million native speaker worldwide. In this paper, a framework is proposed to tackle some of the most common challenges posed by Malay social media text (informal text). Among the features discussed in this paper are the handling of *Bahasa Rojak* also known as Mix language (Malay-English language), the handling of *Bahasa SMS*, the proper handling of Emoticon and finally the handling of Valance shifter. As a result, RojakLex lexicon was constructed consists of 4 different lexicons combined together, namely (1) MySentiDic: a Malay lexicon, (2) English Lexicon: Translated version of MySentiDic, (3) Emoticon lexicon: a combination of 9 different well known lists of commonly used online emoticons, (4) Neologism lexicon: consists of common neologism words used in Malay social media text. The proposed system shows tremendous improvement in accuracy by recording 79.28% compared to baseline which recorded 51.38% only. Discussion and implication of these findings are further elaborated.

Keywords: Sentiment analysis · Malay sentiment analysis · Malay lexicon
Natural language processing · Unsupervised technique · Data analytics

1 Introduction

Sentiment analysis (SA), also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1]. It represents a large problem space.

There are two common approaches adopted by most researchers while dealing with SA, which may be classified into machine learning approach also known as supervised approach and lexicon based approach also known as unsupervised approach [2].

Contact author at k.chekima@gmail.com for a copy of RojakLex.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 205–219, 2018.
https://doi.org/10.1007/978-981-10-8276-4_20

Researchers dealing with supervised approach mostly utilized one of machine learning techniques namely Support Vector Machine (SVM), Neural Network (NN), Naïve Bayes (NB), Maximum Entropy (ME) to build their classifiers [2], etc.

According to [3], Classifiers built using supervised methods reach a high accuracy in detecting the polarity of a text as shown in [4–6]. However, even though such classifiers perform splendidly in the domain they are trained on, their performance drops significantly when the same classifier is used in different domains.

As for unsupervised approach, researchers used either Dictionary Based Approach or Corpus Based Approach. These techniques involve calculating orientation for a document from the semantic orientation of words or phrases in the document [7]. Dictionaries for lexicon based approaches can be created manually, as described in [8], or automatically, using seed words to expand the list of words [7, 9, 10]. According to [11], even though lexical approach does not invariable outperform machine learning method, yet its overall track record is better.

One advantage of using unsupervised technique is the flexibility and simplicity to improve the accuracy by embedding certain rules to handle wrong polarity classification, such as negation handling, polarity shifter handling, intensity handling, question sentences handling, etc. Another advantage of unsupervised approach over supervised approach is the analysis speed. A corpus that takes a fraction of a second to analyze with lexicon approach can take hours when using more complex models like SVM (if training is required) or tens of minutes if the model has been previously trained. Second, the lexicon and rules used by unsupervised technique are directly accessible, not hidden within a machine-access black-box. Lexicon based approach is therefore easily inspected, understood, extended or modified.

To the author's knowledge, most of the work done concerning SA in Malay language were built for a specific domain using a supervised technique like in [12–17]. Despite the high accuracy recorded as a result of using these techniques, their performances drop precipitously when the same classifier is used in different/multiple domains. Among other drawbacks are, they require (often extensive) training data which are, as with validated sentiment lexicons, sometimes troublesome to acquire especially for Non-English languages such as Malay language due to limitation of resources. Second, they depend on the training set to represent as many features as possible (which often, they do not especially in the case of the short, sparse text of social media). Third, they are often more computationally expensive in terms of CPU processing, memory requirements, and training/classification time (which restricts the ability to assess sentiment on streaming data). Fourth, they often derive features "behind the scenes" inside of a black box that is not (easily) human interpretable and are therefore more difficult to either generalize, modify, or extend (e.g., to other domains).

On the other hand, researchers who adopted the second technique (unsupervised technique) while dealing with Malay SA relied on the use of a less efficient lexicon, along with the use of less effective technique called Bag Of Words (BOW) representation as an information extraction/representation. BOW model disrupts word order, breaks the syntactic structures and discards some semantic information of the text which results a drop in sentiment analysis accuracy.

In this work, we investigated some of the most common features that contributes in improving Malay social media opinionated text SA, including the language used in Bahasa Rojak, Bahasa SMS, the use of emoticons and the use of valence shifter. From there we proposed to construct an effective lexicon to handle informal text, along with proper handling of valence shifters which BOW failed to handle.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 discusses the methodology adopted, this includes construction of mix lexicon, handling of Bahasa SMS, handling of valence shifter and finally construction of emoticons lexicon. Section 4 is the experimental setup. Section 5 discusses the result. We concluded our paper in Sect. 6 by carrying out discussion and future work.

2 Related Work

This section is divided into two subsections, the first section focus on work conducted concerning Malay SA using the supervised approach. The second section focus on work conducted using the unsupervised technique.

2.1 Supervised Approach

In this study [17], they investigated how feature selection methods contribute towards the enhancement of Malay sentiment analysis performance. They conducted experiments using seven different features selection methods (FSM) namely IG, PCA, Relief, Gini Index, uncertainty, Chi-squared and SVM-based, and three different supervised machine-learning classifiers namely, SVM classifier, Naïve Bayes classifier and k-nearest neighbour classifier. They concluded that there is no superior classifier for all feature selection algorithms nor there is superior FSM method for all data sizes. Another paper [13] investigated the suitability of using Artificial Immune System (AIS) technique for opinion mining on Malaysian movie review. The proposed technique is then compared with other three well-known traditional machine learning techniques namely Naïve Bayes, Support Vector Machine and K-Nearest Network. They concluded that the AIS is not suitable to mine opinion in Malaysian movie review dataset.

This paper [14] introduces a new Reverse Porter Algorithm (RPA) stemming technique for Malay text. First, the new proposed RPA is compared to the existing stemming technique named Backward-Forward Algorithm (BFA). Next, two experiments will be carried out to examine the accuracy performance for the sentiment analysis on Malaysian Newspaper using Artificial Immune Network (AIN). Each of the two experiments utilized one of the two stemming algorithm mentioned above. As a result, SCIN using Reverse Porter Algorithm outperform Backward-Forward Algorithm by 0.60% with each recorded the accuracy of 53.67% and 53.07% respectively.

[15] Conducted an exploratory research on opinion mining of online movie reviews written by Malaysian. They tested data using machine learning classifiers namely; Support Vector Machine, Naïve Baiyes and k-Nearest Neighbour. The result illustrates that the performance of these machine learning techniques without any pre-processing

of the micro-texts or feature selection is quite low. They concluded that, extra steps such as noise removal/data cleaning are required to execute opinion mining using machine learning approach.

A Malay Mixed Text Normalization Approach (MyTNA) and a feature selection technique based on Immune Network System (FS-INS) have been introduced in the opinion mining process using machine learning approach [18]. The purpose of MyTNA was proposed to normalize noisy texts in online messages. The results show that accuracy values of opinion mining using Naïve Bayes (NB), k-Nearest Neighbor (kNN) and Sequential Minimal Optimization (SMO) slightly increased after the introduction of MyTNA and FS-INS.

Another study has been conducted that focuses on normalizing informal Malay micro-text [19]. This paper proposes a normalization technique of noisy terms that occur in Malaysian micro-texts. They identified few commonly used word patterns by Malaysian, such as the use of symbols, use of numbers, etc. As a result, about 5000 noisy texts were extracted from 15000 documents that were created by Malaysians. The result shows up to 5% improvement in accuracy values of opinion mining.

2.2 Unsupervised Technique

An utilized lexicon based approach was introduced to perform sentiment classification of Malay Facebook posts [20]. The aim of this paper is to identify the opinion mining and sentiment analysis components for extracting both English and Malay words in Facebook. First, they transformed unstructured information from Facebook into meaningful lexicons by removing symbols and stopwords. The stopwords used in this research was borrowed form English stopwords. Next, these lexicons are stored into a database after being manually classified into happy (positive), unhappy (negative) or emotionless. For every happy words matched, the counter is increased accordingly. The same technique applied to unhappy words matched. They used simple prediction method by comparing the percentage between happy and unhappy emotions from processed comments. They concluded that users are satisfied (happy) if the percentage of happy sentiment is greater than the unhappy sentiment and vice versa.

A lexical based method has been introduced to analyse Facebook comments written in Malay language [21]. They used two types of lexical dictionaries; the first dictionary was manually constructed with scores of $+1$ for a positive word and -1 for a negative word. The second dictionary is inspired by the use of SentiWordnet, where words have their corresponding negative and positive weight. In order to classify sentence's overall sentiment, they use three different calculation techniques namely, term score summation, term counting and average on comments.

A new model for Malay sentiment analysis on Malay movie review has been proposed that was based on a combination of supervised (machine learning K-NN) and unsupervised (lexicon, Wordnet) machine learning approaches [12]. 11 features were used and these were grouped under four categories: Features based on the presence and frequency of sentiment words, features based on the level of the sentence, features based on the polarity level of sentiment words and features based on the conditional probability of subjective words. The main goal of this research was to introduce features that can

enhance Malay sentiment analysis and classification. They considered bag-of-words (unigram) as baseline for sentiment analysis. They conclude that, feature set has a positive effect on Malay sentiment classification compare to baseline, this can be observed in the improvement of precision, recall and F-measure values, where improvement were recorded at 16.96%, 10.80% and 18.84% respectively compared to base line.

A lexicon based for Malay sentiment analysis [22] has been proposed to develop sentiment lexicon induction for the Malay language. To achieve this, they first map WordNet Bahasa onto the English WordNet to construct a multilingual word network, next, they used a dictionary-based approach and a supervised classifier for classifying words with their sentiment polarities. The algorithm was evaluated against the General Inquirer lexicon, claiming that it performs with accuracy that is comparable to human accuracy.

Based on the papers reviewed related to Malay SA, there are few limitations which can be associated to either the lack of resources or to the use of poor techniques. Limitations related to resources can be linked to the lexicons adopted, as most researchers use only Malay lexicon that consists of Malay opinionated word, this has leads to ignoring non-Malay words (mainly English opinionated words).

Limitation related to techniques can be grouped into three different groups. The first group related to pre-processing techniques, where data is cleaned and made available to be computationally processed. Some of the pre-processing steps pose a negative impact when applied to sentiment analysis despite their positive impact when applied to text mining. For example, removing punctuation and symbols may lead to loss of important data especially in informal text where emoticons are built using symbols, removing these symbols leads to removing emoticons, and by removing emoticons leads to losing important data as emoticons are intensively used in both social media text as well as in Short Message Service (SMS) text.

The second part is related to information extraction. Most researchers adopted the BOW technique. Despite its simplicity, BOW performed poorly when dealing with complex sentence such as sentence containing negation, intensifiers, diminisher, etc. The third part is related to prediction techniques, where most researchers relied on the use of Majority voting method solely by summing up the frequency of positive terms and the negative terms, the final prediction class is awarded to the polar that recorded higher frequencies. This technique proven to be effective when dealing with simple sentences, yet, deploying MV alone as prediction method without considering other features such as valence shifter and contrast leads to wrong prediction.

In this research, the scope was focused on investigating features related to Malay informal text which have not been tackled previously, which are believed to have major contribution towards SA accuracy. Among these features are emoticons as well as mix language BahasaRrojak. Malaysians tend to use mix language along with emoticons especially text involving social media. For instance, consider the word “best” and the symbol “👍” in the following two sentences, “kamera ni 👍” and “kamera ni best”. If mix language and emoticons are ignored, such sentences will be either ignored or miss classified by the classifier. Other important feature is Bahasa SMS, which is famous among Malaysians especially in social media text.

3 Methodology

This section is divided into few subsections according to the flow of the proposed framework in Fig. 1. First, data collection is discussed followed by data tokenization, data pre-processing, polar word extraction, valence shifter handling and finally prediction of sentence's final sentiment. A news lexicon called RojakLex is constructed in this study which will be further discussed in the coming sections.

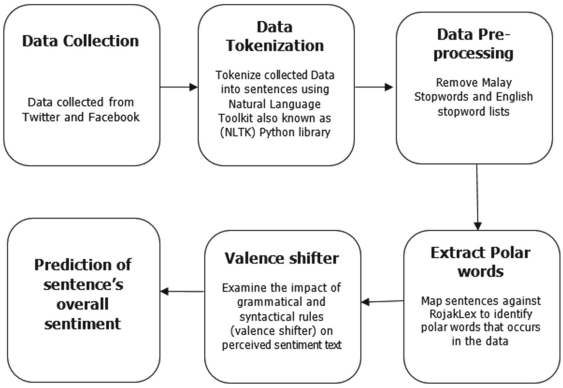


Fig. 1. Proposed framework to handle Malay informal text (Social Media Text)

3.1 Data Collection

The first step involved in the proposed system is to retrieve the right data (Malay opinionated text). Since the intention of this chapter is meant for Malay informal text, data were mainly collected from Twitter. Data collected is annotated by native speakers. The collected data were preserved to its original form, in other words, no spelling checker was performed on the data. A total of 8,026 tweets were collected.

Figure 2 shows a sample of data collected from Twitter on movie review, as can be observed from sample data, text used in Malay informal text is more complicated compared to the text used in a formal Malay text. This is due to the use of Bahasa Rojak, emoji, Bahasa SMS, etc. This will be further discussed in the coming subsections.

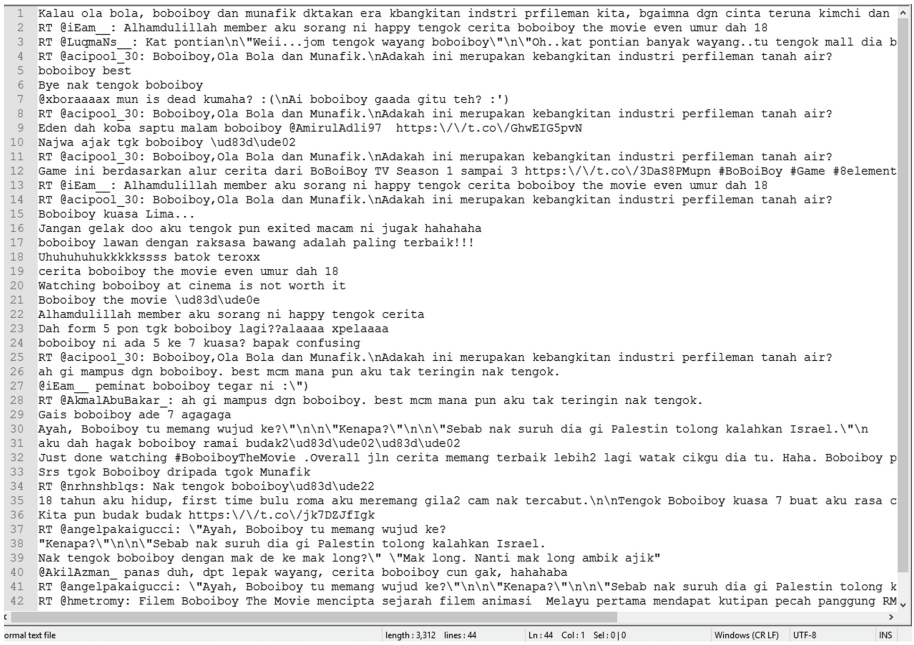


Fig. 2. Sample of data collected from Twitter on the movie BOBOI

3.2 Data Tokenization

To chunk data into sentences, NLTK toolkit was used to tokenize the data. NLTK is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) written in the Python programming language. Once the data is tokenized, next data is pre-processed.

3.3 Data Pre-processing

To transform data from raw data into a data that can be computationally processed, first step is to remove noise. As can be observed from Fig. 2, data contains unwanted symbols (noise) such as "\n", hashtag "#", to remove the unwanted symbols without removing or affecting the emoji present in a text, we used regular expression to target a removal of specific symbols. Next, Malay stopwords and English stopwords were removed from text. To remove Malay stopword list, we used the list constructed by [23], as for English stopwords, we used list retrieved from http://norm.al/2009/04/14/list-of-english-stopwords/. There are some of words have been excluded from English stopword list which are believed to lower sentiment accuracy if they are used, such as although, and, cannot, can, but, could, couldn't, cry, has, should, not, hasn't.

3.4 Polar Word Extraction

To be able to extract polar words that contributes towards sentence's sentiment concerning informal Malay text, the need to have a lexicon that contains not only Malay polar words, but English polar words, emoji, and also polar words that are written in short form, because Malaysians tend to use Bahasa Rojak, Emoji and Bahasa SMS when expressing their opinion especially in social media sites such as Facebook and Twitter. To address the lexicon issue, RojakLex was constructed constisting of English-Malay lexicon (Bahasa Rojak lexicon), Bahasa SMS lexicon, and emoji lexicon. The following subsection discusses the construction of RojakLex.

Bahasa Rojak. The most common and widely adopted language in Malay social media text is called Bahasa Rojak (Mix Language). Rojak language can be described as a Malaysian pidgin (trade language) formed by code-switching among two or more of the many languages of Malaysia. For instance, "`Kamera ini memang best`", this language consists of two languages, Malay language "`kamera ini memang`" and English language "best".

Based on the observation of the type of languages used by Malaysian in Bahasa Rojak text, Malay language and English language was the most dominant languages along with some local slang. To handle Bahasa Rojak, the first step was to construct a lexicon that contains both English and Malay opinionated words. To construct the mix lexicon, Malay lexicon called MySentiDic developed by [24] was used along with its translated version, next these two lexicons were combined to form the mix lexicon. Figure 3 shows a sample of the constructed mix lexicon. This lexicon mainly constructed to handle sentences that contain both English and Malay opinionated words. This lexicon can overcome the weakness posed by previous technique where every time an English word is encountered, it will be translated to Malay first, then will be mapped against Malay lexicon to retrieve its polar, which happen to be time consuming.

```
Sombong Tenggelam frivolousness Malas mengotorkan uncongenial ditikam men-
jerit mengganggu menderita undynamic unsupportable Cussed bodoh unscientific
stereotaip letih Sessile foul mirage bladder_disorder aggression ill_nature
parthenogenetic condole rupture scold superficially lorn fatcat sputter por-
tentous immature kwashiorkor sinking
```

Fig. 3. Sample of Malay-English lexicon

Bahasa SMS. Other than Bahasa Rojak, Malaysians tend to use Bahasa SMS (Short Message Services language) also known as Sistem Mesej Singkat while expressing their opinion online. According to Dewan Bahasa dan Pustaka, SMS language is the language used to communicate via SMS channels, and in most cases the language used in Bahasa SMS is non-standard language/ informal language. Abbreviation and combination of words are used arbitrarily. For instance, "`kamera ni x best`", the "`x`" in the sentence is an example of Bahasa SMS referring to negation "no" or "`tidak`". To handle Bahasa SMS, we utilized the rules proposed by Dewan Bahasa dan Pustak

(Panduan Singkatan Khidmat Pesanan Ringkas (SMS) Bahasa Melayu, 2008). Among these rules are the removal of vowels in a word, dropping certain characters, using first and last character in a word, the use of "tak", "x" replacing negation, etc. Table 1 outlines examples of commonly used Bahasa SMS by Malaysians.

Table 1. List of some of the words word and their corresponding short form (Panduan Singkatan Khidmat Pesanan Ringkas (SMS) Bahasa Melayu, 2008)

Word	Short form	Word	Short form	Word	Short form
Adik	adk/ dik	kilowatt	kwj	gram	gm
atas	ats	kompleks	kondo	halaman	hlm atau hal.
atas nama	nama a.n	kondominium	kump	halang	hlg
awak	awk	kumpulan	krg	hari	hr
awal	awl	kurang	lpr	hari bulan	hb
ayah	ayh	lapar	lwn	hektogram	hg
bagi	bg	lebih	lps	hidup	hdp
bahagian	bhg	lewat	ksh	hospital	hosp
bahasa	bhs	Kapal	kwn	Inci	in
baju	bj	lorong	lrg	Ini	ni
Balak	blk	makan	mkn	institute	inst
bangun	bgn	malam	mlm	Itu	tu
bapak	bpk/pak	malass	mls	jadi	jd
batang	btg	mana	mn	jalan	jln
berat	brt	masak	msk	jangan	jgn
bila	bl	masih	msh	kampung	kg
bodoh	bdh	megabait	mb	Kapal	ksh
bukan	bkn	memang	mmg	kasih	kwn
Buku	bk	mereka	mrk	kawan	kpd
Bulan	bln	miligram	mg	kepada	krn
dan sbgnya	dsb	Negeri/Negara	neg	kilobait	kg
jangan	jgn	nombor	no.	takut	tkt
Dapat	dpt	orang	org	tapi	tp
Dari	dr	pada	pd	tidak	tdk
daripada	drpd	padang	pdg	halaman	hlm atau hal.
datang	dtg	pagi	pg	halang	hlg
datuk	tuk	paling	plg	hari	hr
dekat	dkt	walaupun	w.pun	Ini	ni
dengan	dgn	payah	pyh	institute	inst
dengar	dgr	perut	prt	Itu	tu
depan	dpn	petang	ptg	jadi	jd

Despite the rules introduced by Dewan Bahasa and Pustaka as a guide line in writing Bahasa SMS, Malaysians tend not to follow these rules fully, this has led to the issue of neologism presence in Malay opinionated social media text. Although there is an increasing

number of research related to text normalization and neologism detection in English, the problem remains far from being solved for Malay language, a limited number of research have been undergone to tackle the problem of neologism.

In this study, we adopted a semi-automatic approach to address this issue. Human centric approach is adopted to translate neologism terms. First data were collected from twitter and Facebook due to the common use of neologism in these two social media sites. As mentioned earlier, a total of 8,026 comments were successfully collected. Next, the data was tokenized into words separated by white space, the correctly spelled English and Malay words along with proper nouns were removed from the corpus. To remove correctly spelled Malay words, dictionary Kamus Dewan Edisi Keempat (2013) was used. Table 2 lists the number of words used in the dictionary listed based on alphabet orders. The remaining words were arranged in descending order (from large to small) based on their frequencies in different sentences. Terms that exceeded certain threshold were kept (threshold were set to 3), and for those words that didn't pass the threshold were discarded. Next, these words were manually translated. As a result, the second lexicon was constructed consists of neologism lexicon. Human centric approach is considered to be essential in the process of translating neologism words. Table 3 lists a number of neologism terms found in Malay opinionated text. This simple method was able to reduce noisy terms used in Malay opinionated text. One disadvantage of this technique is the frequent update on neologism lexicon is required due to the continuous emerging of neologism word especially in Malay informal text.

Table 2. Number of words present in dictionary based on alphabet order

Alphabet	No. of words	Alphabet	No. of words	Alphabet	No. of words
A	1699	J	978	S	3221
B	1938	K	3571	T	2117
C	1154	L	1600	U	433
D	1163	M	1907	V	142
E	557	N	617	W	272
F	342	O	320	X	18
G	1465	P	2436	Y	61
H	707	Q	20	Z	95
I	610	R	1378		

Emoticon Lexicon Construction. Another common feature used in Malay social media text is emoticons. Over the years, people start to embrace the use of emoticons as an alternative to face-to-face visual cues while communicating through computer-mediated communication. Emoticons are typically made up of typographical symbols such as, ")", "=", "–", ":", or "(" and commonly represent facial expressions. Emoticons can be read either normal way, like "(^ ^)" (a happy face), or sideways, like ":-(" (a sad face), Consider the following example, "`saya (^ ^) handphone ni`", although no sentiment word present in this sentence, yet this sentence conveys a positive expression due to presence of the positive emoji "(^ ^)".

Table 3. Examples of Neologism in Malay opinionated social media text

Neologism	Word	Neologism	Word	Neologism	Word
nape	kenapa	cite	Cerita	tima	Terima
4	Untuk	gak	Juga	Jer	Sahaja
U	awak	fon	Telefon	lom	Belum
bes	Bagus	leh	Boleh	coz	Sebab
K	ok	Ngan	Dengan	cite	cerita
I	Saya	Ari	Hari	X, tak	tidak
Y	Kenapa	cam	Macam	giler	gila
N	Dan	Tq	Terima kasih	sorang	seorang

According to [25], emoticons normally are used in three different ways. First, emoticons can be used to express sentiment when sentiment is not conveyed by explicit positive or negative words in a text. Second, emoticons are used to intensify sentiment that already been conveyed by polar words (positive negative words). Third, emoticons can be used to disambiguate sentiment, for instance in sarcasm sentences where explicitly expressed polar words are to be negated.

Since similar emoticons are used across different languages, from the literature review conducted concerning available emoticons list used in sentiment analysis, we have identified 9 different potential resources, 8 of the lists are used in [25], and 1 used in [26]. These lexicon lists are believed to cover most of the commonly used emoticons on social media. These lists were combined to form a large lexicon while ignoring duplicate and other emoticons that do not carry any form of sentiment such as flag representation, body part representation, animal representation etc. As a result, Emoticon Lexicon was constructed consists of more than 650 emoticon features that include positive emoji, for instance, (-: (-; (8 (: (; (= (^: (^; (O: *: 0: :-)) :-D :-p :> :] :};-);] =) o;) and negative emoji that includes $:):); 0_ o :(:-(|-0 |-o ||-: :-(.

To properly handle Malay social media text, all of the lexicon constructed namely, mix English-Malay lexicon, neologism lexicon, emoticon lexicon were combined to form one big lexicon called RojakLex. RojakLex contains Malay-English polar words, neologism, Bahasa SMS words and Emoticons.

Valence Shifter Handling. Relying on RojakLex lexicon alone without considering other features in English language was not effective enough to handle Bahasa Rojak. For instance, "`kamera ni canggih, but mahal`", using RojakLex will not be able to recognize the presence of contrast word "but" which transformed the overall sentiment from positive to negative. Due to that, an extra English-Malay features need to be handled in order to ensure the accuracy of Bahasa Rojak sentiment analysis as well as to overcome the weaknesses as a result of using BOW. To achieve this, valence shifter were handled including negation, intensifier, diminisher as well as contrast. Table 4 list some of the valence shifter list in English and Malay. Negation flips the polar from negative to positive and vice versa, intensifier intensifies the polarity, and diminisher diminishes and sometime negates the original polarity. Contrast on the other hand indicates a shift in sentiment polarity, where the sentiment of the text after the conjunction dominates the overall rating.

Table 4. List of Malay and English language valence shifter

Valence	Malay	English
Negation	tidak, x, bukan, tak	aint, arent, cannot, cant, couldnt, darent, didnt, doesnt, ain't, aren't, can't, couldn't, daren't, didn't, doesn't, dont, hadnt, asnt, havent, isnt, mightnt, mustnt, neither, don't, hadn't, hasn't, haven't, isn't, mightn't, mustn't
Intensifier	amat, nian, betul, sungguh, benar, sekali, paling, agak, sungguh, terlalu, sangat	absolutely, amazingly, awfully, completely, considerably, decidedly, deeply, effing, enormously, entirely, especially, exceptionally, extremely, fabulously, flipping
Diminisher	jarang, kurang	Almost, barely, hardly, just, enough, kind of, kinda
Contrast	tetapi, walaubagaimanapun, walaupun, meskipun, sekalipun, namun	Although, even though, but, tough, despite, however, yet, regardless

The final step involved is to predict the overall sentiment of a sentence. Once the values retrieved for both negative and positive words that occurs in a sentence after considering the existence of valence shifter, a simple prediction method is used to decide the overall sentiment of a sentence. The polarity (sentiment) of a given sentence s that contains a list of positive expressions P and a list of negative expressions N is defined as:

$$polarity(s) = \begin{cases} positive \sum_{p \in P} count_{pos}(p) > \sum_{n \in N} count_{neg}(n) \\ negative \sum_{p \in P} count_{pos}(p) < \sum_{n \in N} count_{neg}(n) \end{cases} \quad (1)$$

4 Experimental Setup

Since most of researchers who adopted lexicon based approach relies solely on lexicons that contain Malay polar words only, the baseline is based on Malay lexicon (MySentiDic) along with syntactic and grammatical features (valence shifter). The proposed system on the other hand utilizes RojakLex along with syntactic and grammatical features (valence shifter). Machine learning prediction have been reported to perform well on English data, we identified three well known machine Learning approaches to be included in this experiment namely Naïve Bayes Classifier, Maximum Entropy and support vector machine. To train machine learning classifiers, 20% of annotated data was used to train the classifier. To conduct the machine learning experiments, scikit-learn was used, a machine learning library for python programming language.

5 Result and Discussion

The result displayed in Table 5 supports the claim in this study that the proper handling of Malay-English word along with emoji and neologism can have a tremendous improvement on Malay social media text compared to using a lexicon that contains

Malay polar words only, this is due to the writing style of Malaysian whereby they intend to include English words, emoticons and short form text in their writing styles. A total of 27.9% improvement has been recorded by the proposed system compared to baseline. The proposed system even outperformed well-known machine learning. One of the main reason machine learning did not perform well is due to their nature by depending heavily on the training set to represent as many features as possible which often, they do not especially in the case of the short, sparse text of social media. In other words, machine learning requires big training data, the bigger the better which in most cases are hard to obtain.

Table 5. Result comparison

Test condition	Accuracy
Proposed system	**79.28**
Baseline	51.38
Support Vector Machine (SVM)	57.96
Maximum Entropy (ME)	49.44
Naïve Bayes (NB)	53.50

6 Conclusion and Future Work

This paper investigated features that contribute towards the improvement of Malay informant text sentiment analysis. As a result, a total of four features have been identified namely Bahasa Rojak and Bahasa SMS handling, Emoji handling and valence shifter handling. Despite the high accuracy recorded by the proposed system, the need to frequently update the lexicon especially neologism list is essential, due to the continuous emerging of new neologism words. As a future work, it is important to propose an algorithm which relies lesser on human intervention, in other words an algorithm that automatically handles new emerging of neologism word concerning Malay social media text is needed. Another feature that can be considered as future work is the handling of sarcasm and anaphora concerning Malay informal text.

References

1. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, vol. 22(2), no. 1, pp. 1–167. Morgan & Claypool, San Rafael (2012)
2. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture, pp. 70–77 (2003)
3. Das, S.R., Chen, M.Y.: Yahoo! for Amazon: sentiment extraction from small talk on the web. Manage. Sci. **53**(9), 1375–1388 (2007)
4. Morinaga, S., Yamanishi, K.: Mining product reputations on the web. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 341–349 (2002)

5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
6. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424, July 2002
7. Wiebe, J.M.: Learning subjective adjectives from corpora. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, no. 1, pp. 735–741 (2000)
8. Hatzivassiloglou, V., et al.: Predicting the semantic orientation of adjectives. ACM Trans. Inf. Syst. **21**(4), 315–346 (2009)
9. Wiebe, J.: Tracking point of view in narrative. Comput. Linguist. **20**(2), 233–287 (1994)
10. Hearst, M.A.: Direction-based text interpretation as an information access refinement. In: Text-Based Intelligent Systems, pp. 257–274. L. Erlbaum Associates Inc., Hillsdale (1992)
11. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of 12th International Confrence on World Wide Web, pp. 519–528 (2003)
12. Alsaffar, A., Omar, N.: Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis. J. Comput. Sci. **11**(4), 639–644 (2015)
13. Samsudin, N., Puteh, M., Hamdan, A.R., Nazri, M.Z.A.: Is artificial immune system suitable for opinion mining? In: 2012 4th Conference on Data Mining and Optimization (DMO), pp. 131–136, September 2012
14. Isa, N., Puteh, M., Mohamad, R., Raja, H.: Sentiment Classification of Malay Newspaper Using Immune Network (SCIN), vol. III (2013)
15. Samsudin, N., Puteh, M., Hamdan, A.R.: Bess or xbest: mining the Malaysian online reviews. In: 2011 3rd Conference on Data Mining and Optimization (DMO), pp. 38–43, June 2011
16. Puteh, M., Isa, N., Puteh, S., Redzuan, N.A.: Sentiment mining of Malay newspaper (SAMNews) using artificial immune system. In: Proceedings of the World Congress on Engineering, vol. III (2013)
17. Alsaffar, A., Omar, N.: Study on feature selection and machine learning algorithms for Malay sentiment classification. In: 2014 International Conference on Information Technology and Multimedia (ICIMU), pp. 270–275 (2014)
18. Samsudin, N., Puteh, M., Hamdan, A.R., Ahmad, M.Z.: Mining opinion in online messages. Int. J. Adv. Comput. Sci. Appl. **4**(8), 19–24 (2013)
19. Samsudin, N., Puteh, M., Hamdan, A.R., Nazri, M.Z.A.: Normalization of common noisy terms in Malaysian online media. In: Proceedings of the Knowledge Management International Conference, pp. 515–520, July 2012
20. Zamani, N.A.M., Abidin, S.Z.Z., Omar, N., Abiden, M.Z.Z.: Sentiment analysis: determining people's emotions in facebook 2 related work. In: Proceedings of the 13th International Conference on Applied Computer and Applied Computational Science, pp. 111–116 (2014). ISBN 978-960-474-368-1
21. Shamsudin, N.F., Basiron, H., Saaya, Z., Abdul Rahman, A.F.N., Zakaria, M.H., Hassim, N.: Sentiment classification of unstructured data using lexical based techniques. J. Teknol. **77**(18), 113–120 (2015)
22. Darwich, M., Azman, S., Noah, M., Omar, N.: Inducing a domain-independent sentiment lexicon in Malay, no. 1 (2012)
23. Chekima, K., Alfred, R.: Automatic construction of Malay stopword list. In: Berry, M.W., Mohamed, A., Yap, B.W. (eds.) Soft Computing in Data Science. Springer, Singapore (2016)

24. Chekima, K., Alfred, R.: Non-english sentiment dictionary construction. Adv. Sci. Lett. **4**, 400–407 (2016)
25. Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., Kaymak, U.: Exploiting emoticons in sentiment analysis. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 703–710 (2013)
26. Gilbert, E.: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

Modeling Dengue Hotspot with Bipartite Network Approach

Woon Chee Kok[1(✉)], Jane Labadin[1], and David Perera[2]

[1] Department of Computational Science and Mathematics, Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
woonchee.kok@gmail.com, ljane@fit.unimas.my

[2] Institute of Health and Community Medicine, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
dperera@unimas.my

Abstract. Dengue poses a large economic burden in Malaysia includes among other endemic countries. In order to detect the likely hotspots that breeds mosquito vectors, this study aims to formulate a contact network model of dengue transmission where the research scenario is characterised by spatial data that is complex and difficult to be modelled. The bipartite network modeling approach can address the homogenous limitation seen in deterministic models by projecting the research scenario into two sets of node: human hosts and locations visited by the human. The data of human movements are collected and aggregated from Sarawak State Health Department while the environmental data are obtained from Kuching Meteorological Department. All data are pre-processed and formulated into a targeted model which consists of eight human nodes and nineteen location nodes and a test model which consists of three human nodes and eight new incoming location nodes. The link weight between two sets of node is quantified using summation rule which combines the environmental predictors for instance temperature, precipitation, humidity, human and vector characteristics. The location nodes in targeted and validated models are ranked using a web-based search algorithm according to the respective ranking values. As a result, the ranking values between the targeted and validated model shows strong ranking similarity with good Spearman rank correlation coefficient ($\rho > 0.80$; $p < 0.001$). The ranked locations can help public health authorities to prioritize the locations for vector control to remove the hotspot which results in the reduction of the spread of dengue disease.

Keywords: Bipartite network · Dengue hotspot · Rank location
Spatial data

1 Introduction

Hotspot detection of mosquito-borne diseases for instance dengue, malaria and zika is a key to ensure eradication [1]. In this study, hotspot is defined as the prime location of mosquito breeding site. Public health targets the hotspot and eliminate the vector effectively [2, 3]. Typical studies of vector-borne disease transmission aim to examine

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 220–229, 2018.
https://doi.org/10.1007/978-981-10-8276-4_21

the relationship between the disease occurrences, distribution of the susceptible and infected at different locations [4–7]. These studies mainly applied statistical approach for combining the epidemiological data, demographic data and environmental predictors such as temperature and precipitation to predict disease outbreaks [5, 7–9]. The approaches required a large data set to understand the long-term disease trends in general in order to achieve statistical significance [10]. These approaches introduced a simple assumption which assumes the sub-populations to be well-mixed and homogenous [11]. This assumption has limitation to be applied into location detection studies due to the different levels of human contact that relate to different factors such as age, social and travel patterns. Many attempted to combine population dynamics with the spatial element by forming partial differential equations (PDE). This eventually turns a simple model to a dispersal-reaction model that range from one- to two- and even three-dimensions of linear or non-linear PDE [12, 13]. This complex model is computationally expensive to solve the non-linearity in PDE [14, 15].

Network modeling employs network theory which is a subset of a graph theory that can depict the complexity of the real world problem [16]. Network theory based approaches are widely applied in infectious disease modeling [17, 18]. These studies use the network topology for instance lattice and network to determine the potential risk carried by the human host in disease transmission [18] and to understand the heterogeneity of human movement effect in disease transmission [17]. Network based approach is reported to be able to provide effective results with the least amount of data [19, 20]. In addition, network modeling can incorporate human movement which is one of the factors that caused dengue incidence to surge [21, 22]. The node and edges in network theory can be used to represent disease transmission dynamics. It is observed that the three main components in Epidemiological Triangle (ET) (see Fig. 1(a)) are always interdependent in disease transmission [23]. The interdependence of the components encourages the application of bipartite network modeling (BNM). In addition, the location component (L) specifies the spatial features of certain environment properties (N). At the same time, the environmental properties characterise a particular location that is different from another location. Hence the ET are simplified into the basic building block (see Fig. 1(b)) as a structure in a network model. The basic block consists of two vertices which are the human host and the location.

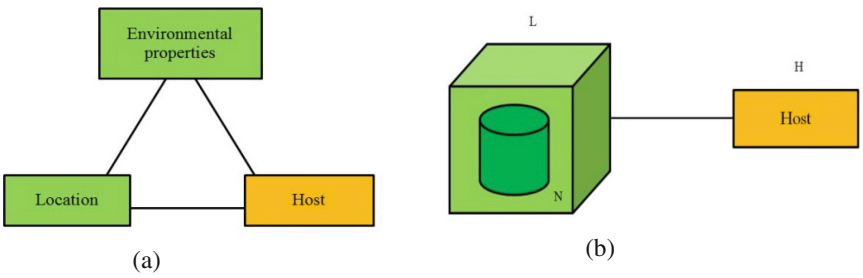


Fig. 1. (a) Epidemiological triangle; (b) Basic building block of bipartite network

This study aims to formulate a contact network to model the possible hotspots of dengue transmission in Kuching, Sarawak with the integration of patients movement. Public health keeps track of the whereabouts of a patient two weeks before they showed symptoms. This study aims to assist public health to conduct effective control interventions especially when the budget, manpower and resources are limited in district-level. Vector control through fogging and removal of mosquitoes breeding sites are the main intervention of dengue prevention in Malaysia [24, 25]. Therefore, identifying dengue hotspots and prioritising the risk level of locations according to the vector density can help public health authorities to target the high risk area first thus preventing waste of resources.

2 Materials and Methods

The methodology applied in this study is extracted from the work that used BNM approach to model the habitat suitability of Irrawaddy dolphin and malaria hotspots [20, 26]. The methodology applied in this study involves (i) pre-processing the dataset; (ii) formulating the bipartite graph structure; (iii) formulating the bipartite network; and (iv) ranking the location node according to vector density.

2.1 Pre-processing Data

The targeted dengue patients in this study is the patients from Epidemiological Week (EW) 28 until 31 in year 2015 in Kuching. The environmental data involved precipitation, humidity, maximum and minimum temperature. There are two network models formulated by using the data from EW 28 until 31, namely targeted model and validated model. The targeted model uses data from EW 28 and 29 (12^{th} until 25^{th} July 2015) while the validated model uses data from EW 30 and 31 (26^{th} July-8^{th} Aug 2015). The targeted model comprises of eight human nodes and nineteen location nodes with twenty data points where each visit of the human is treated as one data point. The validated model consists of three patients with positive dengue tests for the validation purpose and consists of twenty-seven location nodes with eleven data points. Each data point is identified with the particular location GPS coordinates (in latitude, x and longitude, y), altitude (Al, in meter, m), average of the maximum and minimum surface air temperature (T, in degree Celsius, °C), mean relative humidity (H, in percentage, %), daily precipitation amount (Pre, in millimetres, mm), the frequency of a human visiting a location (Fh), total duration of stay of one human in a location (Du, in seconds, s), frequency of one location is visited by human (Fl) (see Fig. 2). The human node refers to the dengue patient and this can be obtained from the investigation form. The location node is the location where visited by the patients two weeks prior the onset date. Each location declared is more than 400 m from one another as it is the maximum flight range of the Aedes mosquito.

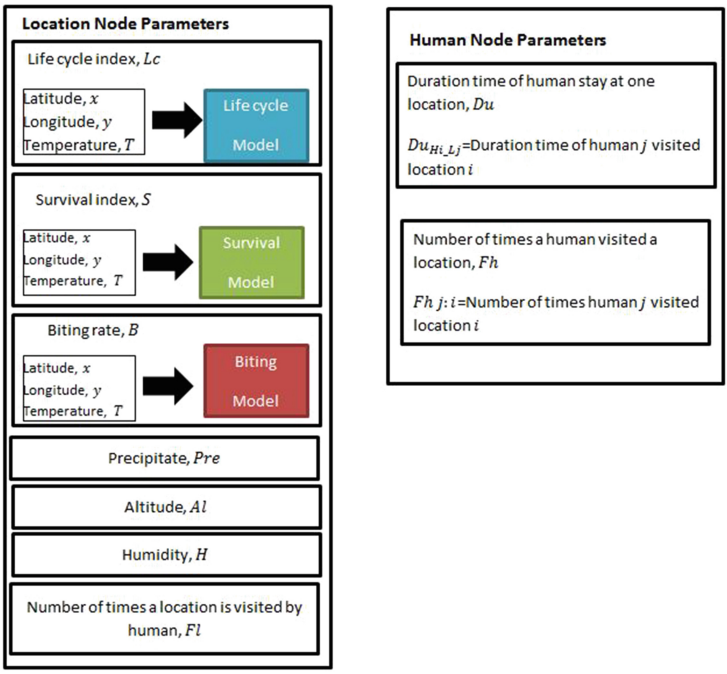


Fig. 2. Parameters used to formulate the network model

2.2 Formulation of Bipartite Network Model

A bipartite graph is an unweighted graph for visual representation of bipartite network while a bipartite network is a weighted graph that gives the functional relationship of its elements including the nodes and their respective links. Therefore, quantifying the parameters for location, human nodes and the link weights are important. The location nodes will then be presented as ranked, where a higher ranked location indicates that the location has higher vector density and is a more critical location that need to be tackled first. As a result, a web-based search algorithm will be used to rank the location according to the computed ranking value which is termed as a dengue hotspot ranking (DHR) value. The bipartite network model will be termed as Bipartite Dengue Contact (BDC) network and the processes of formulating this network are as discussed below.

Formulation of Bipartite Graph

Figure 1(b) is a basic building block of a BDC graph structure for a BDC network. The host vertex (H) in the Fig. 1(b) is applied in this study and replaced by the dengue patient's vertex and termed as human node (H). The targeted model includes patients with positive and negative dengue serological test. For validation, the validated model only includes the patient with positive results to compare the Spearman rank correlation

coefficient (SRCC). The location vertex is denoted as location node (L) where the distance between the location nodes are more than 400 m. Furthermore, the edge linked the human and location nodes and this link weight is quantified to measure the contact strength in BDC network. Greater link strength indicates a more frequent contact between the human and the location. Location with higher contact strength suggests that the particular location has higher possibility to develop into a high risk breeding site.

Eight patients with positive and negative dengue serological test are recorded during EW 28 and 29. Thus, there are eight human nodes labelled as H1, H2, H3, H4, H5, H6, H7 and H8. There are nineteen locations labelled as L1, L2, L3, L4, L5, L6, L7, L8, L9, L10, L11, L12, L13, L14, L15, L16, L17, L18 and L19. The data points mentioned in 2.1 are the twenty edges in the BDC network. BDC graph is defined in (1)–(4). From (4), the edge '*H1L1*' represents the link formed when a dengue patient node with label H1 visited the location node with label L1.

$$BDC_{DEN\_KCH} = BDC\ (H, L, E) \text{ where} \tag{1}$$

$$H = \{H1, H2, H3, H4, H5, H6, H7, H8\} \tag{2}$$

$$L = \{L1, L2, L3, L4, L5, L6, L7, L8, L9, L10, L11, L12, L13, L14, L15, L16, L17, 18, L19\} \tag{3}$$

$$\mathrm{E} = \left\{ \begin{array}{c} H1L1, H1L9, H2L2, H2L10, H2L15, \\ H3L3, H3L11, H3L16, H4L4, H4L12, \\ H4L17, H4L19, H5L5, H6L6, H6L7, \\ H6L18, H7L7, H7L13, H8L8, H8L14 \end{array} \right\} \tag{4}$$

Parameter Quantification for Location Node

The parameters that are utilised to contribute to the link weight values are life cycle duration (Lc), vector survival (S), vector biting rate (B), altitude (Al), humidity (H), precipitation (Pre) and number of visits by human to one location (Fl). The quantification of Lc and S explained next are not available in the past literatures and will be discussed below.

Life cycle duration (Lc) measures the duration of mosquito development from egg hatching to adult at every locality in BDC network. The life cycle duration is inversely proportional to the dengue transmission rate where the shorter the time taken for a complete life cycle leads to an increase in the dengue infection rate. Due to its known dependence on temperature, the life cycle of *Aedes* mosquito plays important role in efforts to understand the effects of environmental property on the dengue transmission. Life cycle duration is defined as a function that comprises of three attributes, temperature, latitude and longitude of the location node. In this section, the polynomial fitting that employs `polyfit` tool in MATLAB is used to determine the life cycle function. This was done by using the data gathered from an experimental research that

examine the effects of daily fluctuations on larval development time of *Ae. Aegypti* collected from Thailand [23]. At the end, a function of degree 6 in (5) is determined to provide life cycle duration with the temperature given.

$$Lc(t) = -0.633t^6 - 0.786t^5 + 1.488t^4 + 1.153t^3 - 0.408t^2 - 0.758t - 0.504 \quad (5)$$

Vector survival (S) measures the survival probability at a locality as an indicator of vector survival rate at one locality. A higher vector survival contributes to a higher vector capacity in one locality. If the mosquitoes lived longer, they were more likely to be infected with the dengue virus [23, 24]. Therefore, the vector survival rate in one locality directly influences the dengue transmission rate. Similarly, polynomial fitting technique is used to formulate a function to describe the vector survival rate with the data aggregated from an experimental research that shows survival rate of *Aedes* mosquito and temperature ranged from 10–34 °C [25]. Finally, a function of degree 6 in (6) is determined to provide the survival rate with the temperature given.

$$S(t) = 1.3908t^6 - 0.2951t^5 - 3.8642t^4 + 1.3217t^3 + 1.2971t^2 - 0.1412t + 0.591 \quad (6)$$

The parameter of the number of times a location is visited by human (Fl) refers to the frequency of the human visits to one locality. It is included to capture the effect of the visiting of the dengue patient to a location. A link matrix, $Link_{Matrix_{BDC\ network}}$ is created in (7) where L_i is the location node i of BDC network, H_j as the human node j of BDC network, $i \in \{1, 2, \ldots, 19\}$ and $j \in \{1, 2, \ldots, 8\}$.

$$Link\_Matrix_{BDC\ network}(L_i H_j) = \begin{cases} 1, \text{ if } L_i \text{ is linked to } H_j \\ 0, \text{ if } L_i \text{ is not linked to } H_j \end{cases} \quad (7)$$

Parameter Quantification for Human Node

The parameters for human node include the duration time for the human stay at one location (Du) and the frequency of a human visited a location (Fh). Du records the total duration of a human stay at one location which standardize in unit of seconds (s). For instances, the human node, H2 visited the location L10 from 7:30 am until 4 pm. The total Du for L10 is 8.5 h or 428400 s.

$Fhj : i$ measures the number of times a human visited a location and the equation is defined in (8) where $i \in \{1, 2, \ldots, 19\}$ and $j \in \{1, 2, \ldots, 8\}$.

$$Fhj : i = \begin{cases} n \text{ if } H_j \text{ visited } L_i \, n \text{ times where } n \in Z^+ \\ 0 \text{ if } H_j \text{ not visit } L_i \end{cases} \quad (8)$$

Link Edge Quantification

Link edge refers to the contact strength between the human and the location and the link weight is termed as dengue contact strength (DCS). The stronger the strength, the greater the degree of attachment between the human and the specific location which contributes to a higher degree of human contact with the specific location.

The summation rule is used to compute the contact strength by applying the equation given in (9) where $i \in \{1, 2, \ldots, 19\}$ and $j \in \{1, 2, \ldots, 8\}$. The values of all parameters are normalised before the DCS computation.

$$\begin{aligned} DCS_{i:j} &= \left(\sum Location\_Node\_Parameters_i\right) + \left(\sum Human\_Node\_Parameters_{j:i}\right) \\ &= (Lc_i + S_i + B_i + Al_i + H_i + Pre_i + Fl_i) + \left(Du_{j:i} + Fh\, j : i\right) \end{aligned} \tag{9}$$

Implementation of Ranking Algorithm

The Hypertext-Induced Topic Search (HITS) search algorithm is applied to rank the location nodes in the BDC network [27]. The main reason of choosing HITS over Page-Rank search algorithm is that the HITS relates both authority and hub pages and this study involves two node types. Thus, the resulted DCS matrix from previous section serves as an input to this adapted HITS algorithm which has been applied in previous study on malaria. The output of this algorithm is principals eigen pairs and the values of the eigen pairs are taken as the DHR values in this study. HITS algorithm rank the location and human nodes according to the DHR values which corresponds to the vector density.

3 Results and Discussion

The complete BDC network formulated successfully and there are seven parameters identified for location nodes and two parameters identified for human nodes. The values attached to the edges are the DCS values of each human-location link. From these nineteen location nodes, location node lab with label L7 has the highest node degree with two links formed with human nodes, H6 and H7. The degree of the nodes does not directly affect the ranking of locations and this is shown in the ranking result in Table 1.

In the table, the result of the location ranking of targeted model is tabulated. From the result, L4 was ranked first with the highest DHR value compared to other location nodes although L4 has a node degree of 1. On the other hand, L7 with the highest node degree of 2 only ranked eighth. The ranking of the nodes' degree did not go into the same direction as the ranking of location. L4 is the location node with the highest ranking. This imply that L4 is the possible location with the highest vector density.

Similarly, the DHR values are computed for the validated network by using the methodology above and the ranking result is tabulated in Table 2. There are eleven location nodes and three human nodes in the validated network. For model validation, Spearman's rank correlation coefficient (SRCC) is employed to measure the ranking correlation between the overlapped location nodes in targeted and validated network. The SRCC of these network, $\rho_{ValidLoc}$ is calculated in (10) where the $d_{TargetValidate}$ is the difference in the ranking place of the identical location node a, while N_{Loc} refers to the number of identical location nodes that appear in targeted and validated network. The SRCC of these network is 1.00 ($\rho > 0.80$; $p < 0.001$) and this indicates a strong

Table 1. Location ranking with DHR in BDC targeted network

Ranking	Ranked location	DHR
1	L4	1
2	L12	0.947297
3	L17	0.937019
4	L19	0.881315
5	L3	0.001617
6	L11	0.00144
7	L16	0.001427
8	L7	7.49×10^{-7}
9	L6	4.64×10^{-7}
10	L18	4.19×10^{-7}
11	L13	2.89×10^{-7}
12	L10	3.3×10^{-116}
13	L2	3.1×10^{-116}
14	L15	2.9×10^{-116}
15	L1	0
16	L5	0
17	L8	0
18	L9	0
19	L14	0

Table 2. Location ranking with DHR in BDC validated network

Ranking	Ranked location	DHR
1	L4	1
2	L2	0.855065
3	L22	0.847884
4	L27	0.803807
5	L26	0.792811
6	L24	0.766107
7	L20	8.62E-06
8	L5	7.24E-06
9	L23	7.07E-06
10	L25	6.76E-06
11	L21	8.33E-21

and positive ranking correlation between targeted and validated network and hence the model is validated.

$$\rho_{ValidLoc} = 1 - \frac{6 \sum_{a=1}^{N_{Loc}} \left[\{ d_{TargetValidate} \}_a \right]^2}{N_{Loc}(N_{Loc}^2 - 1)} \tag{10}$$

4 Conclusion

This study formulated two network models of dengue transmission in Kuching from EW 28 until 31. The location nodes in the targeted network consists of the locations visited by the patients with positive and negative results in the serological test. In the validated network, the hotspots are determined by high DHR values. The SRCC result demonstrated a strong and positive correlation between these network models hence validated

the network models. This study also suggests that the network modeling approach can provide a promising result with the least amount of data ($n < 30$). The output of this study can be incorporated in the *eNotifikasi* system used by the officers from the Ministry of Health to monitor dengue transmission and curb the outbreaks [28].

Acknowledgements. The authors thank Universiti Malaysia Sarawak for the support in this research with the grant number F08/SpFRGS/1601/2017. Our heartfelt thanks also go to Sarawak State Health Department and Sarawak Meteorological Department for providing the research data.

References

1. Aziz, S., Aidil, R.M., Nisfariza, M.N., Ngui, R., Lim, Y.A.L., Yusoff, W.W., Ruslan, R.: Spatial density of Aedes distribution in urban areas: a case study of breteau index in Kuala Lumpur, Malaysia. J. Vector Borne Dis. **51**(2), 91 (2014)
2. Nagao, Y., Thavara, U., Chitnumsup, P., Tawatsin, A., Chansang, C., Campbell-Lendrum, D.: Climatic and social risk factors for Aedes infestation in rural Thailand. Trop. Med. Int. Health **8**(7), 650–659 (2003)
3. Ritchie, S.A., Johnson, B.J.: Advances in vector control science: rear-and-release strategies show promise… but don't forget the basics. J. Infect. Dis. **215**(Suppl_2), S103–S108 (2017)
4. Johansson, M.A., Dominici, F., Glass, G.E.: Local and global effects of climate on dengue transmission in Puerto Rico. PLoS Negl. Trop. Dis. **3**(2), e382 (2009)
5. Rueda, L.M., Patel, K.J., Axtell, R.C., Stinner, R.E.: Temperature-dependent development and survival rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae). J. Med. Entomol. **27**(5), 892–898 (1990)
6. Sarfraz, M.S., Tripathi, N.K., Tipdecho, T., Thongbu, T., Kerdthong, P., Souris, M.: Analyzing the spatio-temporal relationship between dengue vector larval density and land-use using factor analysis and spatial ring mapping. BMC Public Health **12**(1), 853 (2012)
7. Suaya, J.A., Shepard, D.S., Siqueira, J.B., Martelli, C.T., Lum, L.C., Tan, L.H., et al.: Cost of dengue cases in eight countries in the Americas and Asia: a prospective study. Am. J. Trop. Med. Hyg. **80**(5), 846–855 (2009)
8. Jeefoo, P., Tripathi, N.K., Souris, M.: Spatio-temporal diffusion pattern and hotspot detection of dengue in Chachoengsao province, Thailand. Int. J. Environ. Res. Public Health **8**(1), 51–74 (2010)
9. Wu, P.C., Lay, J.G., Guo, H.R., Lin, C.Y., Lung, S.C., Su, H.J.: Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. Sci. Total Environ. **407**(7), 2224–2233 (2009)
10. Esteva, L., Vargas, C.: Coexistence of different serotypes of dengue virus. J. Math. Biol. **46** (1), 31–47 (2003)
11. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**(4), 599–653 (2000)
12. Garnett, G.P.: An introduction to mathematical models in sexually transmitted disease epidemiology. Sex. Transm. Infect. **78**(1), 7–12 (2002)
13. Kon, C., Labadin, J.: Reaction-diffusion generic model for mosquito-borne diseases. In: 2013 8th International Conference on Information Technology in Asia (CITA), pp. 1–4. IEEE (2013)

14. Cheng, A.D., Golberg, M.A., Kansa, E.J., Zammito, G.: Exponential convergence and H-c multiquadric collocation method for partial differential equations. Numer. Methods Partial Differ. Equat. **19**(5), 571–594 (2003)
15. Crandall, M.G., Ishii, H., Lions, P.L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. **27**(1), 1–67 (1992)
16. Craft, M.E., Caillaud, D.: Network models: an underutilized tool in wildlife epidemiology? Interdisc. Perspect. Infect. Dis. **2011**, 12 (2011)
17. Belik, V.V., Geisel, T., Brockmann, D.: The impact of human mobility on spatial disease dynamics. In: International Conference on Computational Science and Engineering, CSE 2009, vol. 4, pp. 932–935. IEEE (2009)
18. Gardner, L.M., Sarkar, S.: Risk of dengue spread from the Philippines through international air travel. Transp. Res. Rec.: J. Transp. Res. Board **2501**, 25–30 (2015)
19. Salathé, M., Kazandjieva, M., Lee, J.W., Levis, P., Feldman, M.W., Jones, J.H.: A high-resolution human contact network for infectious disease transmission. Proc. Nat. Acad. Sci. **107**(51), 22020–22025 (2010)
20. Ying, L.C., Labadin, J., Chai, W.Y., Tuen, A.A., Peter, C.: Applying bipartite network approach to scarce data: modeling habitat suitability of a marine mammal species. Procedia Comput. Sci. **60**, 266–275 (2015)
21. Sutherst, R.W.: Global change and human vulnerability to vector-borne diseases. Clin. Microbiol. Rev. **17**(1), 136–173 (2004)
22. Viennet, E., Ritchie, S.A., Williams, C.R., Faddy, H.M., Harley, D.: Public health responses to and challenges for the control of dengue transmission in high-income countries: four case studies. PLoS Negl. Trop. Dis. **10**(9), e0004943 (2016)
23. Carrington, L.B., Armijos, M.V., Lambrechts, L., Barker, C.M., Scott, T.W.: Effects of fluctuating daily temperatures at critical thermal extremes on Aedes aegypti life-history traits. PLoS One **8**(3), e58824 (2013)
24. Lambrechts, L., Paaijmans, K.P., Fansiri, T., Carrington, L.B., Kramer, L.D., Thomas, M.B., Scott, T.W.: Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti. Proc. Nat. Acad. Sci. **108**(18), 7460–7465 (2011)
25. Tun Lin, W., Burkot, T.R., Kay, B.H.: Effects of temperature and larval diet on development rates and survival of the dengue vector Aedes aegypti in north Queensland, Australia. Med. Vet. Entomol. **14**(1), 31–37 (2000)
26. Eze, M., Labadin, J., Lim, T.: Contact strength generating algorithm for application in malaria transmission network. In: 2011 7th International Conference on Information Technology in Asia (CITA 11), pp. 1–6. IEEE (2011)
27. Eze, M., Labadin, J., Lim, T.: Structural convergence of web graph, social network and malaria network: an analytical framework for emerging web-hybrid search engine. Int. J. Web Eng. Technol. **9**(1), 3–29 (2014)
28. Kok, W.C., Labadin, J., Mohammad, A., Wong, K.S., Chang, Y.L.: Android-based disease monitoring. In: International Conference on Information and Communication Technology (ICICTM), pp. 97–103. IEEE (2016)

Data Fusion Based on Self-Organizing Map Approach to Learning Medical Relational Data

Rayner Alfred[1(✉)], Chong Jia Chung[1(✉)], Chin Kim On[1(✉)], Ag Asri Ag Ibrahim[1(✉)], Mohd Shamrie Sainin[1(✉)], and Paulraj Murugesa Pandiyan[2(✉)]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
{ralfred,chinkimon,awgasri,shamrie}@ums.edu.my, jiachung93@gmail.com

[2] School of Mechatronic Engineering, Universiti Malaysia Perlis, Arau, Perlis, Malaysia
paul@unimap.edu.my

Abstract. Amount of data generated and stored in relational databases has motivated numerous researchers to study and develop learning algorithms on learning relational data mining. One of the most important relational tasks is to discover knowledge from relational data for a better decision making. Despite that, various representations can be generated using the same data by applying the Self-Organizing Map (SOM) methods in clustering relational data. This can be achieved by tuning the parameters used in Self-Organizing Map (SOM), such as the number of clustering, weights, seeds, epoch and others. Thus, this paper proposes a summarization method that applies SOM as the main algorithm to cluster relational data and applies the concept of data fusion in order to get better results in learning relational data. Input data obtained from Dynamic Aggregation of Relational Attributes will be clustered using the SOM method by tuning the SOM parameters. Results generated will be fused and embedded into the target table to form a single representation. A few representations will be formed and fed into the classifiers (J48 Decision Tree and Naïve Bayes classification model) as input data. Throughout the experiments conducted, representations that are extracted by tuning the number of cluster produced better results compared to the representations that are extracted by tuning the other parameters. Overall, the data summarization approach based on individual data fusion is found to perform better compared to the other types of data fusion. In addition to that, the clusters based data fusion with average number of clusters provided better accuracy performances compared to clusters based data fusion with small and large number of clusters.

Keywords: Relational data mining · Clustering · Data summarization
Self-Organizing Map · Data fusion · Dynamic aggregation relational attributes

1 Introduction

Traditional data mining approaches applied a single table of dataset in learning new knowledge. Nowadays, data mining approaches have been extended to multi-relational

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 230–240, 2018.
https://doi.org/10.1007/978-981-10-8276-4_22

data mining approaches that look for patterns involved in multiple tables (relations) from a relational database instead of looking for patterns involved in a single table. In a relational database, a target table has one-to-many relationship with a non-target table. A target table is defined as a table that consist rows of unique labelled object while a non-target table is a table that consists of rows of object where multiple rows can be linked to a single object stored in a target table. The most important component of relational database is the relationship that exists between tables. It defines a connection between a set of tables that are logically related to each other via the information each contains. The Dynamic Aggregation of Relational Attributes (DARA) algorithm is one of the approaches that introduced by Alfred in learning relational databases [1–4]. DARA summarizes all the records kept in the non-target tables that have many-to-one relationships with records kept in the target table by converting them into a term-record matrix based on the TF-IDF concept. Then, DARA clusters these records into groups where these multiple records in non-target table are associated to a particular record in target table. The summarized data generated will then be embedded into the target table and fed into any classifiers. The accuracy of the classification task is highly depended on the representations of data since the predictive accuracy of any classification tasks can be influenced by the performance of the data summarization process on the basis of the descriptive accuracy of the summarized data; while the descriptive accuracy is highly influenced by the representation of the records stored in the non-target tables that are associated with records stored in target table [1–3].

Based on previously published works, DARA algorithm is able to the descriptive accuracy of the summarized data. However, the performance of DARA is highly dependent on the effectiveness of the clustering techniques in specifying the appropriate number of clusters and seeds of the cluster. Nevertheless, the summarized data can also be represented in many forms and these multiple forms can be used as a multiple perspective of the same data for better knowledge discovery purposes. By applying the concept of data fusion using clustering, all these multiple forms of representations can be integrated together to form a single table. The aim of the data fusion is to produce a more enrich and useful information by combining all the summarized data together. Since the performance of the classification task is highly influenced the data representation and the quality of the extracted knowledge [4], this paper proposes a standard framework that learns relation data based on data fusion using Self-Organizing Map (SOM) to determine the best data summarization approach based on data fusion. SOM algorithm is selected since SOM is powerful and convenient tool for exploring, clustering and visualizing data as well as. In addition to that, the position of the seeds of the clusters can be addressed effective when using SOM to cluster the data.

The rest of the paper follows, where Sect. 2 will highlight some related works on learning relational data. Section 3 will describe the self-organizing map based data fusion approach to learning relational data. Section 4 discusses and analyses the results obtained. Section 5 will conclude the paper.

2 Related Works

Learning is an important aspect of and challenges in data mining. Through learning relational data, hidden knowledges and patterns as well as the relationships among data are able to be identified and to be extracted. Most existing approaches are based on propositional approach and look for patterns in a single data table. Nowadays, most real-world data are often stored in relational databases due to the limited storage of a single table especially when enormous and complex data is involved. These situations have motivated plenty of researchers to study and develop learning algorithms on relational data mining [5]. The data stored in a database may consist of a lot of informative information but this data have to be processed before any relevant information can be extracted. The information extracted can assist the process of critical decision making relevant set of constructed features [4]. Majority researchers make use of the inductive logic programming, relational association rules, relational classification rules, relational clustering rules, relational distance-based methods in learning relational data. Nevertheless, these existing learning approaches consider the learning algorithm as passive process that makes use of the information presented to them [3].

Dynamic Aggregation of Relational Attributes (DARA) algorithm [1–4] was introduced to learn data stored in relational databases. In the previous experiments, it has shown that high predictive accuracies can be achieved by using DARA algorithm to learn relational data [6]. The DARA algorithm applied the concept of feature construction that summarizes information stored in the non-target tables by clustering them into groups, where these multiple records exist in non-target tables is correspond to the data stored in the target table. The data summarization method employs the vector space model (Term Frequency-Inverse Document Frequency (TF-IDF) weighted frequency matrix) as a relational data model, where the representation of information stored in multiples non-target tables will be analyzed and transformed into data representation in a vector space model [3]. In other words, DARA is aimed to map data stored in the non-target table to the records stored in the target table based on the feature patterns or instances constructed and selected in terms of TF-IDF weighted frequency matrix. Alfred has strongly emphasized that the input representation of data for data mining is important for data summarization process [4]. In data summarization, the representation of the record stored in the target table influences the descriptive accuracy of the summarized data. If these summarized data are fed into a classification model as one of the input features, the predictive accuracy of the classifier will also be affected. Despite that, a set of data can be represented in plenty ways and all these data representations might be contained some useful and relevance knowledges or information. Multiples data representations can be combined by applying data fusion concept to learn the relational data.

The main idea of data fusion is to combine two or more data representations obtained from the same data sources into one that is more enrich and meaningful information than any previous original data sources [7]. Plenty of data fusion frameworks have been developed both within the research and business environments to act as guidelines and aid the development of fusion systems by creating the most appropriate algorithm for the defined problem [8]. Initially, data fusion is referred as merging of homogeneous data and now, moving towards heterogeneous data. Data fusion techniques had been

widely employed on multi-sensor environments, life sciences and others, fusing data from several different data sets; however, recently researchers perceived that these techniques can also be applied to other domain, such as text processing. Recently, many applications faced the requirement of data fusion using clustering due to the information contained in a single data source is limited by its specific observation. Hence, combining several observations may create a better, clearer vision and comprehensive understanding of the problem [9]. Different researches can be mentioned such as investigating the memory persistence of the bacteria by combining multiple experimental observations of a bacterium under different conditions and evolutional times [10], combining text mining data and bibliometric data to explore the structure mapping of journal sets [11], retrieving correlated or complementary information about the underlying functional partitions of genes and proteins [12].

Fusing multiple data sources can yield some significant advantages to learning task or information retrieval. One of the advantages is that user of such a system can obtain a complete yet concise summary of all existing data without accessing and interpreting all the information sources separately. The results are complete because there is no left-over object; concise as there is no object is duplicated and the data presented to the user is without contradiction [13]. For example, in bioinformatics, fusing multiple data sets representing different organisms [14] or different tissue types [15] improves the understanding of the underlying biological process. Besides, data fusion is able to search the pseudo relevant documents by measuring relative performance of information retrieval systems and use the data generated to rank the retrieval systems [10]. Recently, a new method has been presented for bearing fault diagnosis using the fusion of two primary sensors: an accelerometer and a load cell [16]. This waterfall fusion model is proposed and resulting an excellent performance in bearing diagnosis. Test results demonstrate that there is a promising improvement in accuracy and reliability of bearing fault diagnosis.

Todays, studies have been extended from supervised learning contexts to unsupervised learning that discovering the hidden structure in unlabeled data. Besides aiming to investigate how data fusion can be applied in unlabeled data, studies also interested to build a single representation for multiple sources of heterogeneous data. A research has also been done in which data fusion with clustering can be concluded into two main categories, clustering ensemble and fusing similarity matrices [10]. An ensemble is a collection of individual classifiers with different parameters which may lead to a higher generalization than when it is working individually. Figure 1 illustrates that each classifier operates separately and generates a solution that is combined by ensemble creating a single representation [17].

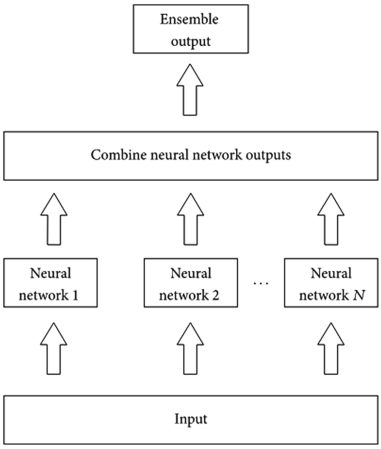


Fig. 1. Blocking diagram of an ensemble

The most commonly used clustering method is the *k*-mean clustering, a partitional clustering algorithm that is based on the square error-criterion. A few researches have been conducted by applying the concept of data fusion using *k*-mean clustering and based on the obtained finding, the accuracy performance can be improved compared to the classical clustering method on a human colon and rectal cancer data [18], isolated word recognition [19] and classification rates [20]. Evidently, data fusion mechanisms significantly improved the quality results of unsupervised learning algorithms. Recently, Complementary Ensemble Clustering (CEC) has been introduced that consists of a framework in which *k*-means algorithm is applied to a weighted, linear combination of the co-association matrices obtained from separate ensemble clustering of different data modalities. CEC is an extension of ensemble clustering that combines the co-association matrices of ensemble clusters from different modalities into one aggregate co-association matrix that is subsequently used for obtaining the consensus clustering. Ensemble clustering extracts information from different clustering of a given data modality based on one or many clustering algorithms and their corresponding parameters. In the experiment conducted, CEC has been applied on two biomedical datasets and all cases of CEC showed superior performance in both datasets [21].

Nowadays, many real-world applications make extensive use of high dimensional data and issues related to visualizing highly dimensional data through low dimensional maps to discover regularities and extract information effectively have been increased. Self-Organizing Map (SOM) model was one of the best known Artificial Neural Network (ANN) with unsupervised learning that introduced by T. Kohonen to discover information and the patterns of the data [22]. SOM performs a topology-preserving mapping from the high-dimensional input space onto a low dimensional display (usually one or two-dimensions), which plots the similarities of the data by grouping similar data items together. It uses competitive learning techniques to train the network and the nodes competing among themselves to display the strongest activation to a given data. Although several improvements and variants to SOM have been conducted and they

have been very successful in several real applications areas, but there is a need to improve the quality of the solution. In order to improve the quality of solution generated, researchers tend to use the techniques of machine learning ensembles. This ensemble combines a set of model decisions of the phenomenon under study to obtain a global decision in order improve the individual answers. A model has been created known as Fusion-SOM, an ensemble of SOM that are combined by fusing prototypes that are modeling similar partitions and their neighborhood relation are given by the edges that measures the similarity between the fused nodes [22]. They claimed that the architecture of the resulting model improved the results in term of quality and robustness, and the topological representation of the single model.

Researches have been done to prove that DARA and data fusion can be used to learn relational data and enhanced the data representation for classification purposes. However, studies related to fusing DARA data with SOM has not been carried out so far. Therefore, this paper proposes a standard framework that learns relational data based on data fusion using SOM to generate a better summarization results.

3 Data Fusion Based on Self-Organizing Map

The aim of this work is to implement and extend the work proposed by Sia et al. [2] by summarizing the data using SOM instead of using the traditional *k*-means clustering. In previous work, relational data is learned by summarizing records stored in a non-target table using several feature construction methods followed by selecting a set of multiple features of summarized data that best describes the data stored in the non-target table. Besides, the set of summarized data that is selected is then embedded to the target table as a new feature-value and thus, J48 classifier is applied on the target table to perform the classification task. In the experiment, they proved that the proposed method is capable of selecting multiple features of summarized data that enrich the representation of all the records stored in the non-target table that are treated as an input data for classification task, and yet, they emphasized again that the representation of the input data is important to get a higher predictive accuracy of the classification task.

Two set of medical relational datasets (e.g., mutagenesis and hepatitis [2]) are used to evaluate the proposed framework. Both these data are converted to term-record matric based on the TF-IDF representation which is based on the weighted frequency matrix. Both relational datasets consist of three different documents (B1, B2 and B3 for Mutagenesis; H1, H2 and H3 for Hepatitis) together with a target table and each document will be imported as an input. Based on the literature review, clustering can be used to perform the data fusion by merging multiples data representations obtained from the same data sources to study the relationships between data representations and the predictive accuracy. In this work, SOM method is selected as the main clustering method as the issue related to the seeds of the clusters can be addressed effectively. The advantage of using SOM to cluster data and visualize high dimensional data into a lower dimension is that the seeds of the clusters are determined automatically by SOM algorithm and this will ease the analysis process.

Figure 2 illustrates the proposed framework of using data fusion based on the self-organizing map. All the non-target tables are transformed using DARA algorithm [1–4] into TF-IDF representation in the form of document text matrix in which clustering or SOM algorithms can be used to group them into different groups. Then, SOM algorithm can be used to cluster and produce multiple different results by adjusting the SOM's parameters (e.g., number of clusters and epoch values). With this different cluster results, they can be embedded into the target table to be fed into classifiers.

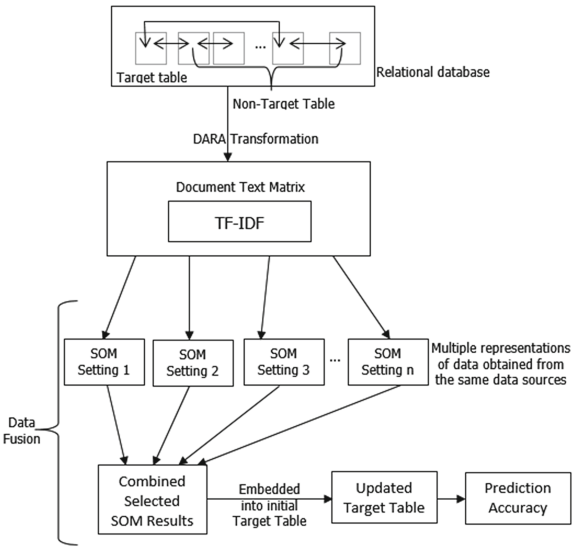


Fig. 2. Structure of the proposed data fusion based on SOM

First, this data will be tuned based on individual data fusion (e.g., P1 through P30) with 30 sets of different parameter settings based on number of clusters (e.g., 4, 8, 12, 16, 20, 24, 28, 32, 36 and 40) and epoch values (e.g., 10000, 100000 and 1000000) as shown in Table 1. All the obtained results will be embedded into the target table accordingly and labelled as P1 to P30 accordingly. The purpose of breaking down the data into smaller pieces based on different settings individually is that the assessment can be performed for each data separately. Eventually, the performance of individual data fusion can be compared against the fused data (based on cluster and epoch) that are described in Table 2. In Table 2, each row represents set of data fusion that is combined based on number of clusters and epoch's values. Each set of representations will be embedded into the target table and will be fed into the classifier as an input to be evaluated. In this work, J48 decision tree and Naïve Bayes model were chosen as both models are more efficient than others complex classification models. For both classifiers, the predictive accuracy of a 10-fold cross validation is used to evaluate the performance of the proposed method. Lastly, all the predictive accuracies generated from the classification task will be analyzed and compared, to make an assumption or conclusion about the performance of data summarization approach based on data fusion.

Table 1. 30 sets of combination parameter settings with the label

Parameter setting labels	Parameter settings	
	Number of clusters	Epoch
P1–P10	4, 8, 12, 16, 20, 24, 28, 32, 36, 40	10,000
P11–P20	4, 8, 12, 16, 20, 24, 28, 32, 36, 40	100 000
P21–P30	4, 8, 12, 16, 20, 24, 28, 32, 36, 40	1 000 000

Table 2. Representations of data with label and parameter settings

Labels	Based on parameter settings labeled in Table 1	Clusters	Epoch
C1	P1, P11 & P21	4	
C2	P2, P12 & P22	8	
C3	P3, P13 & P23	12	
C4	P4, P14 & P24	16	
C5	P5, P15 & P25	20	
C6	P6, P16 & P26	24	
C7	P7, P17 & P27	28	
C8	P8, P18 & P28	32	
C9	P9, P19 & P29	36	
C10	P10, P20 & P30	40	
E1	P1, P2, P3, P4, P5, P6, P7, P8, P9 & P10	–	10,000
E2	P11, P12, P13, P14, P15, P16, P17, P18, P19& P20	–	100,000
E3	P22, P22, P23, P24, P25, P26, P27, P28, P29 & P30	–	1000,000

4 Results and Discussion

All the results are tabulated and average based on each representation are calculated and recorded in Table 3. Besides, the best parameter settings as well as worst parameter settings will also be discussed in Table 4.

Based on Table 3, it was observed that the higher average accuracy can be obtained when the target datasets are embedded with individual data fusion representation (e.g., P1–P30). The performance accuracy is higher when the target table is embedded with data fused using individual with Naïve Bayes classification model. On the other hand, data summarization based on epoch values has the worst performance compared to others, while the data summarization based on number of cluster has the moderate performance in most cases. Data summarization based on data fusion (number of clusters and epoch value) has the worst performance compared to the results obtained using the data summarization based on individual parameter settings is due to relevant features are constructed in the representation. The features contained in the representation have relatively minor effects and this eventually causes a huge impact when a large number of features with relatively minor effects taken together.

By comparing the accuracy performance for the data representation based on number of clusters and epoch value, data represented based on number of clusters is slightly

better compared to epoch values. Once again, this showed that the data fusion based on epoch values has not produced a good data representation for the classification task.

Table 3. Average of data accuracy

Dataset		Type of representations	Individual P1–P30	Based on	
				Clusters C1–C10	Epoch value E1–E3
Mutagenesis	B1	J48	0.8072	**0.8114**	0.8002
		Naïve Bayes	**0.8642**	0.8122	0.7311
	B2	J48	0.8092	0.8121	**0.8302**
		Naïve Bayes	**0.8597**	0.7993	0.7049
	B3	J48	0.8082	0.8115	**0.8333**
		Naïve Bayes	**0.8333**	0.8173	0.8167
Hepatitis	H1	J48	0.6375	0.6475	**0.6846**
		Naïve Bayes	0.5839	0.5849	**0.5961**
	H2	J48	**0.5928**	0.5916	0.5898
		Naïve Bayes	**0.5927**	0.5713	0.5243
	H3	J48	**0.5980**	0.5980	0.5857
		Naïve Bayes	**0.5916**	0.5585	0.5094

Table 4. Summary of data accuracy

Best performance							
Data	Classifiers	Individual		Settings based on			
				Cluster		Epoch	
B1	J48	**P21**	**0.8351**	C7	0.8246	E1	0.8091
	Naïve Bayes	**P24**	**0.8886**	C7	0.8670	E2	0.7395
B2	J48	P8, P9	0.8351	C3	0.8298	**E2**	**0.8459**
	Naïve Bayes	**P15**	**0.8833**	C7	0.8345	E3	0.7193
B3	J48	P6	0.8462	C9	0.8354	**E3**	**0.8509**
	Naïve Bayes	**P19**	**0.8518**	C4	0.8412	E2	0.8202
H1	J48	P19	0.6819	**C7**	**0.6978**	E1	0.6939
	Naïve Bayes	P1–P12	0.5965	**C6**	**0.6180**	E2	0.6119
H2	J48	Majority	0.5972	C1–5	0.5972	**E1**	**0.6112**
	Naïve Bayes	Majority	0.5972	**C8**	**0.6011**	E2	0.5290
H3	J48	**P1–P28**	**0.5984**	**C8**	**0.5984**	E1	0.5924
	Naïve Bayes	**P19**	**0.6024**	C6	0.5905	E1	0.5182
	Naïve Bayes	P24	0.5563	C2	0.5161	E3	0.5041

Table 4 clearly shows that most of the best performances throughout these experiments can be obtained using the individual data fusion representation when the epoch values are greater (100,000 and 1,000,000). It was also observed that when the clusters based data fusion is embedded into the target table, the best performance can be achieved

when number of cluster is not too small and not too large (e.g., C6–C9), where the number of clusters ranges from 24 to 32 clusters.

5 Discussion

In this paper, we have proposed a standard framework that summarizes data based on data fusion to generate a better and relevant summarization results. The input data was trained by using SOM algorithm with different set of parameter settings. All the results generated are fused back to the initial target table. The updated target table was extracted into few data representations based on number of cluster and epoch value. The performance of these summarization approach generated are evaluated by passing the data summarizations into the classifiers as an input. Based on the results, data summarizations with individual data fusion performed better compared to other types of data fusion studied in this paper. It was also observed that when the clusters based data fusion is embedded into the target table, the best performance of the classification task can be achieved when number of cluster is not too small and not too large.

Acknowledgments. This work has been supported by the Research Grant Scheme project funded by the Ministry of Education (MOE), Malaysia, under Grants No. FRG0382-ICT-2/2014.

References

1. Alfred, R.: DARA: data summarisation with feature construction. In: 2008 Second Asia International Conference on Modelling and Simulation (AMS) (2008)
2. Sia, F., Alfred, R., Chin, K.O.: Learning relational data based on multiple instances of summarized data using DARA. In: Soft Computing Applications and Intelligent Systems, pp. 293–301 (2013)
3. Alfred, R.: Optimizing feature construction process for dynamic aggregation of relational attributes. J. Comput. Sci **5**, 864–877 (2009)
4. Sia, F., Alfred, R.: Evolutionary-based feature construction with substitution for data summarization using DARA. In: Conference on Data Mining and Optimization (DMO), Langkawi (2012)
5. Kavurucu, Y., Senkul, P., Toroslu, I.: Concept discovery on relational databases: new techniques for search space pruning and rule quality improvement. Knowl.-Based Syst. **23**, 743–756 (2010)
6. Nevilie, J.: Statistical Models and Analysis Techniques for Learning In Relational Data (2006)
7. Alfred, R.: Summarizing relational data using semi-supervised genetic algorithm-based clustering techniques. J. Comput. Sci. **6**, 775–784 (2010)
8. Haghighat, M., Abdel-Mottaleb, M., Alhalabi, W.: Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition. IEEE Trans. Inf. Forensics Secur. **11**, 1984–1996 (2016)
9. Esteban, J., Starr, A., Willetts, R., Hannah, P., Bryanston-Cross, P.: A review of data fusion models and architectures: towards engineering guidelines. Neural Comput. Appl. **14**, 273–281 (2005)
10. Yu, S., Moor, B.D., Moreau, Y.: Clustering by heterogeneous data fusion: framework and applications. Internal Report (2008)

11. Wolf, D.M., Fontaine-Bodin, L., Bischofs, I., Price, G., Keasling, J., Arkin, A.P.: Memory in microbes: quantifying history-dependent behavior in a bacterium. PLoS ONE **3**, e1700 (2008)
12. Liu, X., Yu, S., Moreau, Y., Moor, B.D., Glanzel, W., Janssens, F.: Hybrid clustering of text mining and bibliometrics applied to journal sets. In: Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM (2009)
13. Sitaram Asur, D.S., Parthasarathy, S.: An ensemble framework for clustering protein–protein interaction networks. Bioinformatics **23**, 29–40 (2007)
14. Bleiholder, J., Naumann, F.: Data fusion. ACM Comput. Surv. **41**, 1 (2008)
15. Ponnapalli, S.P., Saunders, M.A., Loan, C.F.V., Alter, O.: A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PLoS One **6**, 1–11 (2011)
16. Acar, E., Plopper, G.E., Yener, B.: Coupled analysis of in vitro and histology tissue samples to quantify structure-function relationship. PLoS One **7**, 1–14 (2012)
17. Safizadeh, M.S., Latifi, K.: Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell. Inf. Fusion **18**, 1–8 (2014)
18. Pasa, L.A., Costa, J.A.F., Medeiros, M.G.D.: A contribution to the study of ensemble of self-organizing maps. Math. Prob. Eng. 1–10 (2015)
19. Gönen, M., Margolin, A.A.: Localized data fusion for kernel k-means clustering with application to cancer biology. In: NIPS 2014, Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada (2014)
20. Moshiri, B., Eslambolchi, P., HoseinNezhad, R.: Fuzzy clustering approach using data fusion theory and its application to automatic isolated word recognition. Int. J. Eng. (IJE) Trans. B **16**, 329–336 (2003)
21. Ribas, A.D., Colonna, J.G., Figueiredo, C.M.S., Nakamura, E.F.: Similarity clustering for data fusion in wireless sensor networks using k-means. In: WCCI 2012 IEEE World Congress on Computational Intelligence, Brisbane, Australia (2012)
22. Yu, S., Tranchevent, L., Liu, X., Glanzel, W., Suykens, J.A., Moor, B.D., Moreau, Y.: Optimized data fusion for kernel k-means clustering. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1031–1039 (2012)

A Review on Outdoor Parking Systems Using Feasibility of Mobile Sensors

Md Ismail Hossen[1], Michael Goh[1(✉)], Tee Connie[1], Azrin Aris[2], and Wong Li Pei[3]

[1] Faculty of Information and Science Technology, Multimedia University, Cyberjaya, Melaka, Malaysia
ismailmmu@gmail.com, {michael.goh,tee.connie}@mmu.edu.my
[2] Emerging & Technology Partnership, TM Business Solution, Telekom Malaysia, Kuala Lumpur, Malaysia
azrin.aris@tm.com.my
[3] School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Pulau Pinang, Malaysia
lpwong@usm.my

Abstract. An efficient outdoor parking system is a crucial need for smart cities to monitor the occupancy of outdoor parking. Currently, there are many external sensors-based parking systems available for indoor parking. These external sensors either need to be installed at the parking slot or attached with vehicles at fixed positions. Hence, the deployment cost is very high for such implementation. Due to high cost and complex network configuration, external sensors are not preferred for Smart Outdoor Parking Systems (SOPS). Several SOPS have been deployed to solve outdoor parking problems. Understanding existing SOPS approaches is essential to develop a robust and effective outdoor parking system. In this paper, we present a review of the various SOPS. We have addressed the most important aspects including technical, economical, accuracy, open issues and challenges of the existing SOPS. Based on the review, a recommendation has been proposed to improve outdoor parking system.

Keywords: Outdoor parking · Smartphone-embedded sensors
Smart parking system · Sensors fusion · Activity recognition

1 Introduction

Parking has been a serious problem in urban cities. The number of vehicles is increasing day by day with the growth of population and economic development. It is extremely difficult and frustrating for drivers to find a parking place in congested areas. According to a study by Parknet [1], it costs $78 billion per year in the form of 4.2 billion lost hours and 2.9 billion for gas in busy roads in the U.S. Another study by White [2] showed that 45% of the traffic generated by automobiles circulation for finding parking places had caused a series of problems like air pollution, traffic congestion, and waste of energy in the New York City. The lack of outdoor parking system leads to a number of illegal parking, heavy traffic on roads, and causes distress for drivers. Hence, monitoring

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 241–251, 2018.
https://doi.org/10.1007/978-981-10-8276-4_23

outdoor parking availability is a significant mission for developing smart city which can guide drivers to find empty parking spaces easily.

Most of the existing outdoor parking systems used tokens or tickets that count the number of cars entering and leaving the parking premises. Many of these systems used external sensors such as sensor under the road surfaces [3], or sensor attached to the side of the vehicles [4]. These methods were implemented to provide outdoor parking occupancy information. Nevertheless, these solutions could neither reduce drivers' anxiety to find a parking space nor provide satisfactory parking management for outdoor parking. Therefore, there is an urgent need for a solution for outdoor parking for a smart city.

We have presented a review on outdoor parking systems using four different approaches; namely computer vision, GPS, wireless based and smartphone-based systems. We have given an insight into the flexibility and cost of implementing these SOPS to provide a precise overview to design a more flexible and cost efficient SOPS. The remainder of the review paper is organized as follows: In Sect. 2, we divide different SOPS into four categories according to their deployment techniques. The categories are vision-based techniques, GPS/Satellite-based techniques, wireless network-based techniques and smartphone-based techniques. Section 2.3 demonstrates a comparison and discussion of the different approaches, in Sect. 3, a recommendation is provided and finally the conclusion is drawn in Sect. 4.

2 Categorization of Smart Outdoor Parking Systems

2.1 Vision-Based Techniques

Vision-based parking system has become attractive with the rapid development in the image processing and computer vision fields [5]. The core idea of vision-based SOPS is to use image processing and computer vision techniques to analyze CCTV video streams [5]. It analyses the content captured by the video cameras to detect and find empty parking space. Hampapur et al. [6] presented an object recognition with adaptive background subtracting and striking movement discovery method to identify moving vehicles. The background subtraction method subtracted the current image from a reference image to find region of interest (ROI). Adaptive background subtraction continuously updated the background changes and generated a mask of where the images were stable. The method detected entering and leaving of cars by identifying the location of the vehicles. Wu et al. [7] described an approach that combined Principal Component Analysis with wavelet change for feature extraction and applied Support Vector Machine (SVM) for vehicle recognition. Bong et al. [8] presented a bi-stream empty parking slot detection approach, in which one stream examined the number of pixels within a certain intensity threshold to determine the presence of a car in the parking slot and other stream eradicated false detection produced by shadows of vehicles, utilizing edge detection and median filter. A ParkLotD system was introduced by Ichihashi et al. [9] to enable a camera-based system for practical outdoor parking detection with high performance. The cameras return real-time available parking spaces and adaptive background subtraction was used to overcome illumination and shadow effects. Similarly,

Lin et al. [10] demonstrated a vision-based parking method for outdoor parking which set up cameras around the parking zone, sends real-time information to a center database that enabled drivers to discover available parking spaces or monitor the parked space through wireless communication device. The system used a statistical method in color image sequences to find out the suitable color for each parking space and the foreground was extracted using color information.

Overall, vision-based techniques are better than sensor based techniques in terms of cost since it does not require additional sensors in each parking slot [11]. In addition, vision-based SOPS is easy to deploy because the cameras setup does not require intricate configuration. Furthermore, vacant parking spots can be identified precisely by analyzing the parking space images. However, there are also some limitations with vision-based systems. The cameras must be in the positions from where they can monitor the whole parking area. Sometimes, high-resolution cameras are high-priced [12]. Figure 1 below demonstrates the block diagram of vision based SOPS.

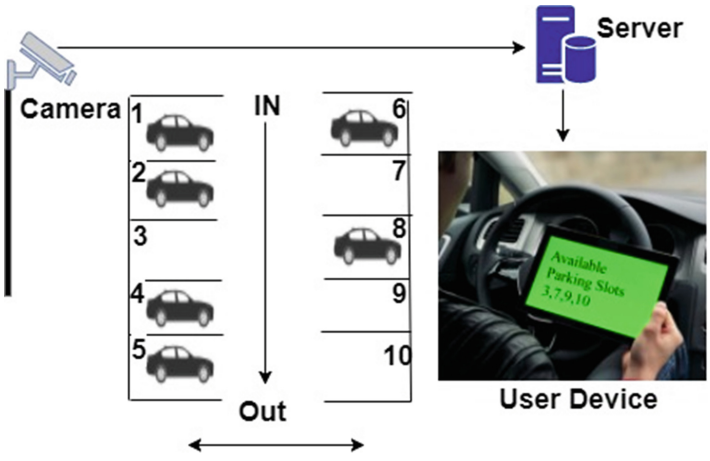


Fig. 1. Block diagram of vision based SOPS

2.2 GPS/Satellite-Based Techniques

The Global Positioning System (GPS) is a location system based on a constellation of around 24 satellites circling the earth at heights of approximately 11,000 miles. GPS gives exact positioning 24 h a day, anywhere in the planet. GPS calculates distance between a GPS satellite and a GPS beneficiary by measuring the amount of time it takes for a radio signal to travel from the satellite to the recipient. Many smart outdoor parking systems take advantage of GPS services to find parking locations. It provides information about the location and availability of parking spaces near the destination of the driver by transmitting signal from satellite to the receivers. A block diagram of GPS based SOPS is shown below in Fig. 2. A number of GPS-based smart outdoor parking systems have been proposed. Gahlan et al. [13] designed a parking management approach by using GPS whose primary purpose was to determine the location and to

send available parking spaces to the driver. The system could also provide feedback about the current location of the drivers. A scientific solution based on past and current status of the drivers was presented by Pullola et al. [31] by utilizing Poisson process to model vacancies of available parking slot. The system helped drivers to choose the place with maximum probability of being available. A location-based parking system was presented by Chon et al. [14] to find the nearest available parking lots for the drivers. The system helped to find parking place in areas like campus, airports. However, the system did not provide information about availability of parking spots.

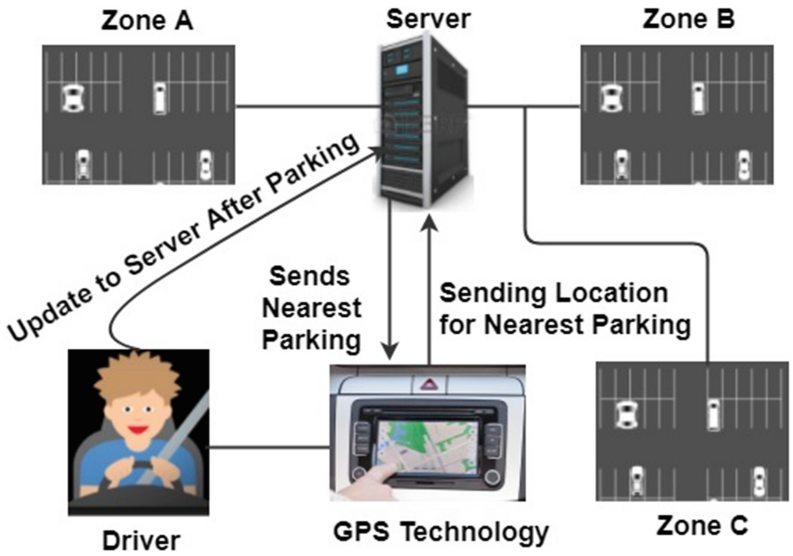


Fig. 2. Block diagram of GPS based SOPS

The good side of GPS-based technology is that it provides real-time location services and guides towards the destination. Thus, it is flexible to use and can obtain knowledge of unknown parking places. Besides, it eliminates the need to install expensive sensors for getting parking places. Conversely, GPS receives location data from the satellite which takes times due to transmission delays. In addition, it depends on the weather to receive clear signals from the satellite. Apart from that, GPS does not guarantee a parking place as it can only provide information about the current state of the parking zone [16].

Overall, vision-based techniques are better than sensor based techniques in terms of cost since it does not require additional sensors in each parking slot [11]. In addition, vision-based SOPS is easy to deploy because the cameras setup does not require intricate configuration. Furthermore, vacant parking spots can be identified precisely by analyzing the parking space images. However, there are also some limitations with vision-based systems. The cameras must be in the positions from where they can monitor the whole parking area. Sometimes, high-resolution cameras are high-priced [12]. Besides, the tracking systems might sometimes give false positive result due to occlusion effects and lighting conditions.

2.3 Wireless Network-Based Approach

Wireless network (WN) is a computer network that uses radio waves for connecting devices, business network and applications. There is an increased interest in wireless-based parking system. The use of WN in smart outdoor parking systems enables people to avoid costly process of introducing sensors and cables. Hanif et al. [15] demonstrated a smart outdoor parking system that allowed users to book their parking slot through Short Message Services (SMS). A wireless communication instrument named micro-remote terminal unit (micro-RTU) was used to confirm booking with details such as lot number and password. Password was used to enter into parking area with a validity of certain period of time. A microcontroller named Peripheral Interface Controller (PIC) was used to automate the system that was capable of storing information of vacant parking spaces, controls entree to the parking area by allowing or denying access. Besides, a system to monitor the entry and exit of vehicles for outdoor parking zones was presented in [17]. It utilized wireless network of photo electronics sensors and a network technology called 6LowPAN over IEE802.15.4 for communication among devices such as computers, routers and mobile devices. 6LowPAN stands for IPv6 over Low-Power Personal Area Network and it allows IPv6 packets to be carried efficiently using small link layer. Gu et al. [18] presented a system named Street Parking System to monitor road side parking spaces using wireless communication with magnetic sensors and ZigBee technology. The system used 3-axes magnetic sensor and an algorithm based on cyclic detection state machine for detecting parking vehicles. A similar outdoor parking system was proposed by Reve et al. [19] that used infrared sensors and radio frequency for communication. Figure 3 below shows a block diagram of wireless based SOPS.

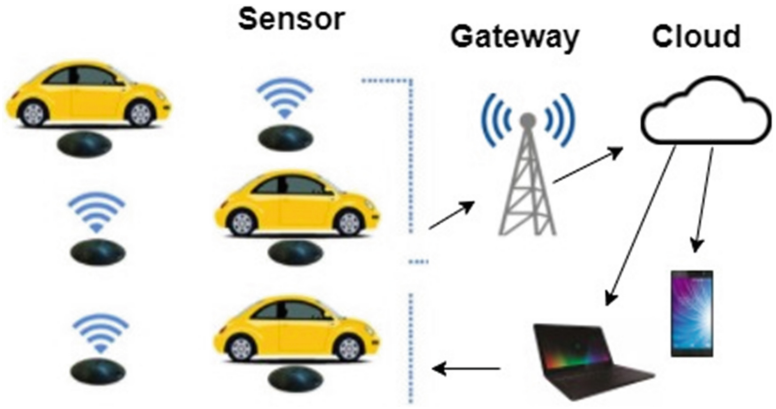


Fig. 3. Block diagram of wireless based SOPS

Wireless technology provides a number of advantages. The wireless technology can be found in homes, workplaces and cars. Therefore, the deployment of wireless based parking is rampant. Besides, the technology can be used for both outdoor and controlled

zone parking systems. WN-based SOPS can detect the passage of cars and communicate with the data center in real-time.

Wireless sensors network-based technique is considered a low cost solution for smart parking. But it typically consists of various sensors such as fluxgate magnetic detector, active infrared detector which are expensive to implement. For example, a study described a method that used wide-area sensor network which had video cameras, microphone, and motion detector [20]. When many drivers use WN-based smart parking systems, the bandwidth of the infrastructure becomes overloaded. Again, there is chance of interference and signal distortion due to external factor such as dust, storm, and fogs.

2.4 Smartphone-Based Techniques

In the past few years, smartphones experience a vast technological growth with the incorporation of new hardware and embedded sensors. Modern smartphones are equipped with various sensors such as accelerometer, barometer, gyroscope and GPS. Smartphone sensors are being used for crowdsensing which collects information from user's mobile. A number of SOPS have proposed to utilize crowdsensing capability [21–26]. Context recognition or user activity recognition technology is used in such systems. Activity recognition can effectively detect activities such as walking, standing, driving, and running. The flow of detected activities, e.g. driving-standing-walking infers the activity when a driver is parking. Analyzing activity recognition for parking can be implemented without using costly sensors. The architecture of smartphone-based SOPS is demonstrated in Fig. 4 below.

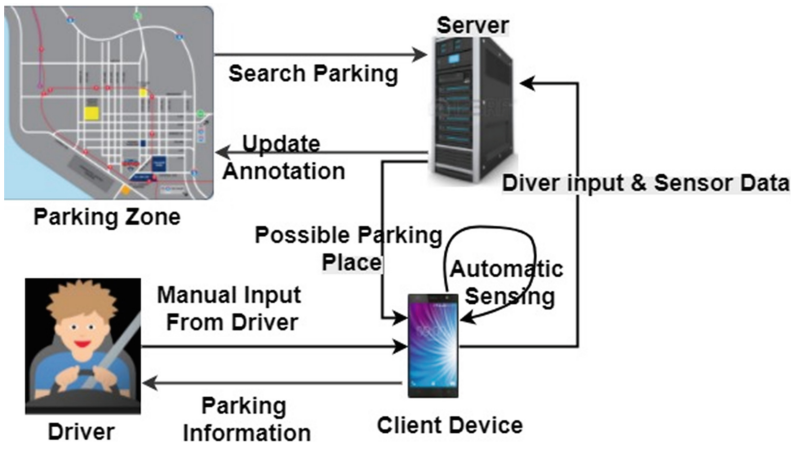


Fig. 4. Block diagram of smartphone-based SOPS

In a study by Salpietro et al. [21], the users' activities were identified automatically through the analysis of accelerometer and gyroscope data to determine the transition mode from car to walking, and vice-versa. To detect activities of users, driving-to-walking or walking-to-driving, sensors in smartphone and Bluetooth were used. Driver's

actions such as walking, driving were classified using supervised machine learning. The solution has helped to solve the problem of installing external sensors in parking areas to track car occupancy. On the contrary, GPS, accelerometer, geospatial data, and Bluetooth data were used by Stenneth et al. [22] to automatically detect when and where a driver parked a car, or released the parking spot. A novel approach named Probabilistic Internal Navigation (ProbIN) [24] that used Bayesian Probabilistic framework was used to address the problem of noisy sensors reading. It was a novel statistical approach for mapping low-cost internal sensors to get the user's position for navigation. Lan and Shih [23] proposed a phone-based driver tracking system that studied the trajectory of the driver. They proposed a way to detect when a driver was about to leave a parking place. A waist-mounted and map-matching algorithm is used to get accurate prediction results. Another solution was presented by Biondi et al. [25] which gathered the user contextual information that was displayed as an approximated distribution of free parking spots using a map. A bluetooth low energy advertisement was used to detect passengers to reduce bias that might be introduced by multiple users in the same car. Nawaz et al. [30] introduced ParkSense, a robust signature matching approach based on bacon reception ratio and an approach based on the rate of change of Wi-Fi beacons to detect driving. In order to sense parking events, ParkSense leveraged the ubiquitous Wi-Fi beacons. It used Wi-Fi signature matching approach to detect driver's presence to the parked vehicle. PocketParker, a crowdsourcing system presented by Nandugudi et al. [26], detected driver's movement utilizing smartphone's GPS and accelerometer sensors. The system stored collected data in a centralized database and broadcast the information to the drivers using Wi-Fi and cellular network. Table 1 shows a summary of the existing smartphone embedded smart outdoor parking systems.

Table 1. Summary of different smartphone based SOPS

Study	Sensors	Storage	Accuracy
Stenneth et al. [22]	Accelerometer, Gyroscope and Bluetooth	Distributed	Transportation mode detection 92.5%, parking occupancy 100%
Nguyen et al. [24]	GPS, Accelerometer and Gyroscope	Centralized	87% for parking occupancy detection
Biondi et al. [25]	GPS and Bluetooth		–
Nandugudi et al. [26]	Accelerometer and GPS		Overall accuracy 92.5%
Nawaz et al. [30]	Wifi-beacon		Identify above 90% users return

We observe that mobile-based parking system uses two types of storages: distributed or centralized. Distributed storage stores data in devices that are not located in the same physical location. On the other hand, a central Database Management System (DBMS) manages all the distributed storages in a unified location. Distributed storage system can process all the requests to access data by balancing the load among several servers. To keep the data up to date in distributed storage system, it requires additional software. Therefore, this incurs additional cost and complexity. On the other hand, centralized database keeps all the data in a single place such as a server or mainframe computer.

Centralized database is easy to maintain since all the data resides in the same location. But the server of the centralized database becomes slow if there are multiple requests to the system.

In general, mobile phone-based system is suitable for outdoor parking. It is the most economical solution among the alternatives. Mobile-oriented outdoor parking system provides the best flexibility since the mobile phones can be carried everywhere. It has also helped to solve the problem of having complex infrastructure or expensive external sensors installation and maintenance. Mobile phone provides an economical smart parking solution by taking advantage of modern mobile phone's powerful sensors. Besides, the system is easy to use since the application is available on the phone. It also provides other features such as a booking system to allow users to reserve a parking slot in advanced.

Nevertheless, there are some disadvantages of using smartphone-based outdoor parking systems. To track location for the car parking systems, the user's location information is needed and this can lead to user privacy breach. Accelerometer and GPS are two important sensors for designing smartphone-based SOPS. Accelerometer is orientation and position dependent, and it needs complex training and processing to achieve good accuracy. Sometimes, the GPS sensor might miss the GPS signal in urban areas [28]. Again, some system takes manual input from the users which might be inflicted by false information created by free riders and selfish liars. In addition, there might be problems when a user books the parking a way too early while the other nearby drivers cannot book the parking place as it has already been blocked. Therefore, application developers need to consider many scenarios to solve or reduce these kinds of problem.

3 Discussions and Recommendation

This section discusses the different SOPS techniques and provides a comparison among these techniques. Table 2 shows the comparison of deployment flexibility, accuracy and cost of the SOPS techniques. We observe that vision-based technique is easy to deploy, and it requires less maintenance cost. However, the tracking systems might sometimes give false positive result due to occlusion effects and lighting conditions [33]. However, it incurs cost on the driver to purchase a smart phone. On the contrary, GPS-based systems are easy to deploy because it does not require infrastructure design and maintenance cost since there is no external sensor installed and parking detection is accurate [33]. Nonetheless, it only provides information of the parking area but not the exact vacant spot. WN-based SOPS approaches depend on additional sensors and the systems require complex infrastructure, which is difficult to deploy [29]. It becomes worse when the entire infrastructure need to be revamped. The accuracy of parking detection is very accurate [33]. Smartphone-based system, on the hand, provides easy deployment with no or very low installation and maintenance cost. But this techniques sometime delays to detect parking and user location privacy has to be disclosed [12].

Table 2. Comparisons different of outdoor parking techniques

Categories	Easy to deploy	Maintenance cost	Startup cost	Accuracy
Vision based	Yes	Moderate	High	False detection occurs
GPS based	Yes	No	Low	Accurate
Wireless based	No	Low	High	Accurate
Smartphone based	Yes	No	No	Delay occurs

From this study, it is clear that smartphone-based SOPS is the most economical solution as compared to the other methods. Although the system has some drawbacks such as sensor sensitivity, this can be resolved by using supervised machine learning techniques [32]. It is also possible to utilize fusion approach that combines data from multiple sensors to increase the accuracy of recognizing drivers' activity. Since mobile-based parking system does not require external hardwires installation and maintenance cost, it could be considered as one of the best options for outdoor parking service providers. Besides, the growth of mobile technology and the ease of mobile phone usability would also facilitate the development of mobile-based SOPS.

4 Conclusion

An efficient real-time outdoor parking system can eradicate the frustration of users for getting parking occupancy in outdoor places such as shopping malls, universities, hospitals. In this review, we present the latest deployments of outdoor parking systems using different approaches including vision-based, GPS-based, wireless-based and mobile-based. The main focus is drawn to the advantages, disadvantages, deployment, and cost of the various systems. Overall, there is a need to implement a cost-effective system to reduce drivers' hectic in finding parking spaces. An analytical comparison is presented and several issues are deliberated. The highlighted issues can be further explored for enhancement of the existing systems.

References

1. Mathur, S., Kaul, S., Gruteser, M., Trappe, W.: ParkNet: a mobile sensor network for harvesting real time vehicular parking information. In: Proceedings of 2009 MobiHoc S3 Work, MobiHoc S3 2009, p. 25, May 2009
2. White, P.: No Vacancy: Park Slope's Parking Problem. Transportation Alternatives. Transalt.org (2017). https://www.transalt.org/news/releases/126, https://www.transalt.org/newsroom/releases/126. Accessed 31 July 2017
3. San Francisco Park website. http://sfpark.org/
4. Nick, T.: Did you know how many different kinds of sensors go inside a smartphone? Phone Arena (2017). http://www.phonearena.com/news/Did-you-know-how-many-different-kinds-of-sensors-go-inside-a-smartphone_id57885. Accessed 24 May 2017

5. Kłosowski, M., Wójcikowski, M., Czyzewski, A.: Vision-based parking lot occupancy evaluation system using 2D separable discrete wavelet transform. Bull. Polish Acad. Sci. Tech. Sci. **63**(3), 569–573 (2015)
6. Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H., Pankanti, S.: Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. IEEE Sig. Process. Mag. **22**(2), 38–51 (2005)
7. Wu, J., Zhang, X., Zhou, J.: Vehicle detection in static road images with PCA and wavelet-based classifier. In: Proceedings of IEEE Intelligent Transportation Systems Conference, vol. 1, pp. 740–744 (2001)
8. Bong, D.B.L., Ting, K.C., Lai, K.C.: Integrated approach in the design of car park. IAENG Int. J. Comput. Sci. **35**(1), 7–14 (2008)
9. Ichihashi, H., Notsu, A., Honda, K., Katada, T., Fujiyoshi, M.: Vacant parking space detector for outdoor parking lot by using surveillance camera and FCM classifier. In: IEEE International Conference on Fuzzy System, pp. 127–134 (2009)
10. Lin, S.F., Chen, Y.Y., Liu, S.C.: A vision-based parking lot management system. In: Proceedings of IEEE International Conference System Man Cybernetics, vol. 4, pp. 2897–2902 (2007)
11. Masmoudi, I., Wali, A., Jamoussi, A., Alimi, A.M.: Vision based system for vacant parking lot detection: VPLD. In: IEEE International Conference on Computer Vision Theory Application (VISAPP), vol. 2, pp. 1–8, January 2014
12. Moses, N., Chincholkar, Y.D.: Smart parking system for monitoring vacant parking. Int. J. Adv. Res. Comput. Commun. Eng. **5**(6), 717–720 (2016)
13. Gahlan, M., Malik, V., Kaushik, D.: GPS based parking system, vol. 5, no. I, pp. 2053–2056 (2016)
14. Chon, H.D., Agrawal, D., Abbadi, A.E.: NAPA: nearest available parking lot application. In: 18th International Conference on Data Engineering, pp. 496–497 (2002)
15. Hanif, N.H.H.M., Badiozaman, M.H., Daud, H.: Smart parking reservation system using short message services (SMS). In: 2010 International Conference on Intelligent and Advanced Systems (ICIAS), pp. 1–5 (2010)
16. Faheem, Mahmud, S.A., Khan, G.M., Rahman, M., Zafar, H.: A survey of intelligent car parking system. J. Appl. Res. Technol. **11**(5), 714–726 (2013)
17. Vera-Gómez, J.A., Quesada-Arencibia, A., García, C.R., Moreno, R.S., Hernández, F.G.: An intelligent parking management system for urban areas. Sensors (Switzerland) **16**(6), 1–16 (2016)
18. Gu, J., Zhang, Z., Yu, F., Liu, Q.: Design and implementation of a street parking system using wireless sensor networks. In: IEEE 10th International Conference on Industrial Informatics, Beijing, pp. 1212–1217 (2012)
19. Reve, S.V., Choudhri, S.: Management of car parking system using wireless sensor network. Int. J. Emerg. Technol. Adv. Eng. **2**(7), 262–268 (2012). www.ijetae.com
20. IrisNet: Internet-scale Resource-Intensive Sensor Network Service. http://www.intel-iris.net
21. Salpietro, R., Bedogni, L., Di Felice, M., Bononi, L.: Park Here! A smart parking system based on smartphones' embedded sensors and short range communication technologies. In: Proceedings of IEEE World Forum Internet of Things, WF-IoT 2015, pp. 18–23 (2016)
22. Stenneth, L., Wolfson, O., Xu, B., Yu, P.S.: PhonePark: street parking using mobile phones. In: Proceedings of 2012 IEEE 13th International Conference on Mobile Data Management, MDM 2012, pp. 278–279 (2012)
23. Lan, K.C., Shih, W.Y.: An intelligent driver location system for smart parking. Expert Syst. Appl. **41**(5), 2443–2456 (2014)

24. Le Nguyen, T., Zhang, Y., Griss, M.: ProbIN: probabilistic inertial navigation. In: 2010 IEEE 7th International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2010, pp. 650–657 (2010)
25. Biondi, S., Monteleone, S., La Torre, G., Catania, V.: A context-aware smart parking system. In: 2016 12th International Conference on Signal-Image Technology and Internet-Based Systems, pp. 450–454 (2016)
26. Nandugudi, A., Ki, T., Nuessle, C., Challen, G.: PocketParker: pocketsourcing parking lot availability. In: Proceedings of ACM International Joint Conference on Ubiquitous and Pervasive Computing, pp. 963–973 (2014)
27. Xing, S., Tong, H., Ji, P.: Activity recognition with smartphone sensors. Tsinghua Sci. Technol. **19**(3), 235–249 (2014)
28. Shucong, Y.: Technical Code of Maintenance for Urban Road CJJ36-2006. China Architecture & Building Press, Beijing (2006)
29. Tang, V.W.S., Zheng, Y., Cao, J.: An intelligent car park management system based on wireless sensor networks. In: Proceedings of the 1st International Symposium on Pervasive Computing and Applications, Urumchi, Xinjiang, China, pp. 65–70 (2006)
30. Nawaz, S., Efstratiou, C., Mascolo, C.: ParkSense: a smartphone based sensing system for on-street parking. In: MobiCom 2013, pp. 75–86 (2013)
31. Pullola, S., Atrey, P.K., El Saddik, A.: Towards an intelligent GPS-based vehicle navigation system for finding street parking lots. In: IEEE International Conference on Signal Processing and Communications, pp. 1251–1254, 24–27 November 2007
32. Stenneth, L., Wolfson, O., Yu, P.S., Xu, B.: Transportation mode detection using mobile phones and GIS information. In: Proceedings of 19th ACM SIGSPATIAL International Conference on Advanced Geographic Information System, p. 54 (2011)
33. Fraifer, M., Fernström, M.: Investigation of smart parking systems and their technologies. In: Thirty Seventh International Conference on Information System, IoT Smart City Challenges Applications (ISCA 2016), pp. 1–14 (2016)

Volatile Organic Compounds (VOCs) Feature Selection for Human Odor Classification

Ahmed Qusay Sabri[1(✉)] and Rayner Alfred[2]

[1] University of Sharjah, Sharjah, UAE
asabri@sharjah.ac.ae

[2] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
ralfred@ums.edu.my

Abstract. A problem of selecting appropriate human VOCs (Volatile Organic Compound) emitted from sweat for human odor classification is presented in this paper. In this paper, all gases emitted by human through sweat have been collected and detected using the latest technology (High resolution GCMS/TOF) Gas Chromatograph Mass Spectrometry/Time of Flight. Due to the limitations of the experimental conducted, only a total of four different persons are required to have the samples of odor collected for twenty different times. 198 VOCs have been detected and feature selection methods have been applied to determine which VOCs are suitable to be used to classify human odor. Two feature selection methods based on Entropy and Chi Square test have been used to determine and decide the best and acceptable VOCs. Based on the results obtained, a total of 17 stable VOCs are extracted from 198 VOCs. In addition to that, there are 10 gases that are detected having zero values for both the entropy and chi-square test and these gases are considered the strongest candidates that can be used for odor detection and classification. The results obtained from this work can be used to assist the task of classifying specific VOCs for human detection through odor.

Keywords: Odor classification · Volatile organic compound · Entropy
Chi-square test

1 Introduction

Conducted researches that are related to odors classification are very limited and yet the problem of recognition and classification of odors are important due to its application in security. This due to the facts that information related to human odors requires specific instrumental equipment in order to collect information related to odors. Odor is one of the five senses data and human being have used these five senses to enjoy comfortable human life with communication and mutual understanding. Artificial odor sensing and classification systems through electronic technology called an electronic nose have been developed according to various odor sensing systems and several classification methods [1–3].

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 252–261, 2018.
https://doi.org/10.1007/978-981-10-8276-4_24

In this paper, a problem of selecting appropriate human VOCs (Volatile Organic Compound) emitted from sweat for human odor classification is presented. In this paper, all gases emitted by human through sweat have been collected and detected using the latest technology (High resolution GCMS/TOF) Gas Chromatograph Mass Spectrometry/Time of Flight. Due to the limitations of the experimental conducted, only a total of four different persons are required to have the samples of odor collected. In this work, each person is tested 5 times on different days in order to ensure the stability of emitted VOCs. As a result, we managed to extract most stable, accurate and rigidity list of VOCs from human with specific Gas name. Compared to other related works in which most of other references have tested the person only one time to detect the emitted VOCs while in this research, we have repeated the test for each person 5 times with total number of 20 test have been applied for each Gas. The final result of our short listed of emitted VOCs list consists of 17 VOCs based on the two feature selection algorithms.

In this paper, a selection procedure system has been developed that can be used to select the appropriate Volatile Organic Compounds (VOCs) emitted from human's sweats that will be used for human odor classification, under different densities based on feature selection algorithms. This feature extraction is performed using the latest technology (High resolution GCMS/TOF) Gas Chromatograph Mass Spectrometry/ Time of Flight.

The rest of the paper follows, where Sect. 2 will highlight some related works on odors classification. Section 3 will describe the processes involved in extracting stable and consistent features for Odor Classification. Section 4 discusses and analyses the results obtained. Section 5 will conclude the paper by suggesting future works that can be conducted based on the results obtained in this work.

2 Related Works

Individuals are thought to have their own distinctive scents, analogous to a signature or fingerprint, the axillary region is of particular interest to play an important role in generating individual odor [3, 5]. The armpit is a skin region where a vast number of glands and bacteria cooperate to produce a strong smell which may give a unique pattern allowing identification of different persons [6]. Classifying the various human odors under different densities have been successfully conducted previously [4]. VOCs emanating from skin contribute to a person's body odor [5]. The human body generates VOCs that may provide information related to types of diseases, behavior, emotional state and health status of a person [6]. Also, body odor is one of the physical characteristics of human which can be used to identify people [6]. The human body odor contains different volatile organic compounds which can be used as biomarkers for diseases, experiments with special trained dogs showed that they are able to smell diseases as e.g. cancer, hypoglycemia or hyperglycemia [7].

Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded

methods. Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. Filter methods include the Chi-square test, information gain or entropy and correlation coefficient scores [8].

This paper applies Chi-square test and entropy in identifying relevant VOCs features used for classifying human's odor. In this paper, each person will be tested for 5 times in 5 different days to get VOCs from same person under different circumstances to ensure the rigidity of our result in feature selection techniques.

3 Extracting Odor Features for Classification

3.1 Sweat Collection

In this subsection, the sweat samples are collected and extracted by using the latest technology (High resolution GCMS/TOF) Gas Chromatograph Mass Spectrometry/ Time of Flight. Individuals will be tested in successive days of the week, in which different genders and variety of ages are also considered. Individuals will have a wash out phase for seven days before starting the test. In these seven days, the volunteers are only allowed to wash their under arms with neutral soap. All days in the morning the Individuals come in the lab and wash their under arms with neutral soap this is a standardized procedure. After that, they will put on a cotton t-shirt in which pads will be placed under the arm pits. The pads will be worn by these individuals for 4 h and during this time they do their normal day actions. In order to get better odor representation, individuals are asked to run stairs up and down for 3 times approximately 3 min per time point. After 4 h the individuals bring the pads to the lab.

A sweat sample is introduced into a Nalophan bag with 1.5 L nitrogen and heated to 90 °C for 30 min. Afterwards, the headspace is collected from the nalophan bag of each sample into a thermodesorption tube (Tenax/Carbograph) until a total volume of 1000 ml. Additionally, one thermodesorption tube without sample and under the same sampling conditions than real samples was put into nalophan bag used in sampling. These tubes were used as blanks and the objective of the conducted experiments is to identify all compounds in sweat samples.

Figure 1 shows the data collected for a single person from a single test of all VOCs emitted from human sweat. There are four volunteers who have been tested for five successive days each. In this experiment, different genders are considered for the volunteers and based on the result of peaks obtained for the same person, there are some differences in reading obtained for different days. In Fig. 1, X axis represents the duration of the sample placed inside the VOC detection device, while Y axis represents the level of concentration of VOC detected.

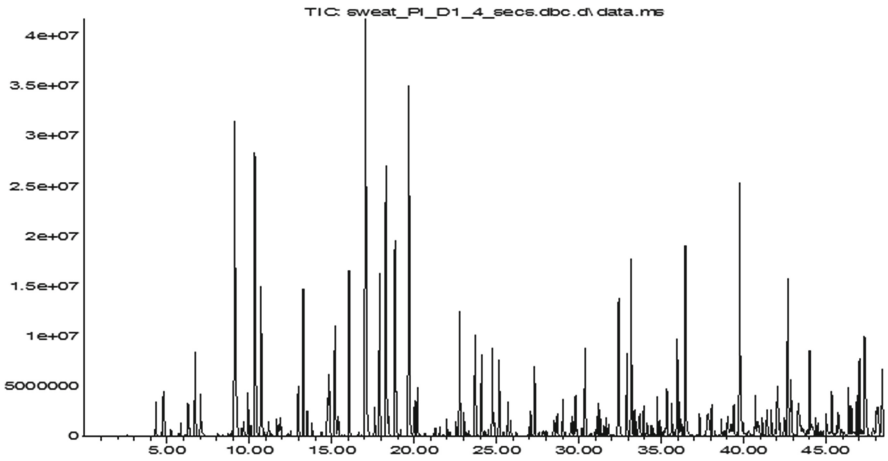


Fig. 1. Result of VOCs emitted from human sweat

Table 1 outlines a list of VOCs categories detected from all volunteers where each gas was tested 20 times to record the appearance. A total of 198 VOCs emitted from our volunteers are detected and recorded. A feature selection process is conducted in order to select relevant VOCs features that can be used for odor classification.

Table 1. Categorized VOCs with total number 198.

Chemical category	Number of detected VOCs
Alcohols	12
Aldehydes	19
Aliphatic Hydrocarbons	33
Aromatic Alcohol	2
Aromatic compounds	3
Cyclic Hydrocarbons	8
Esters	13
Ethers	14
Halogen-containing compounds	7
Furans	7
Ketones	32
Lactones	2
Nitrogen-containing compounds	11
Organic Acids	11
Sulfur-containing compounds	5
Terpenes	19
Total	198

3.2 Feature Selections Processes

Feature selection methods can be classified in a number of ways. The most common one is the classification into filters, wrappers, embedded, and hybrid methods [9, 12–14]. Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods are fast, scalable and can rank individual features or evaluate entire feature subsets, filter method is used in order to reduce the feature space dimension space, possibly obtaining several candidate subsets [9, 12]. Wrappers consider feature subsets by the quality of the performance on a modelling algorithm, which is taken as a black box evaluator. Wrappers are only feasible for greedy search strategies and fast modelling algorithms such as Naïve Bayes, linear SVM, and Extreme Learning Machines [9, 10]. Embedded methods perform feature selection during the modelling algorithm's execution. These methods are thus embedded in the algorithm either as its normal or extended functionality, Embedded methods usually work with linear classifiers (SVM) and induce penalties to features that do not contribute to the model [9].

3.3 Feature Selections Methods

Feature selection techniques have become an apparent need in many bioinformatics applications. In addition to the large pool of techniques that have already been developed in the machine learning and data mining fields, specific applications in bioinformatics have led to a wealth of newly proposed techniques [12, 13]. The choice of feature selection methods differs among various application areas. Feature selection methods, pertaining to several well-known application domains such as text mining, image processing, bioinformatics and Industrial applications. In our case for the Bioinformatics application, the best performing relevant feature selection are Chi Square and Information Gain [9, 12, 13].

Chi-Square Test X^2

The chi-square (X^2) goodness of fit test begins by hypothesizing that the distribution of a variable behaves in a particular manner, in which we assume the expected values based on observation and then decide to accept or reject the assumption or hypothesis [11, 15]. The Chi-square test (X^2) is a nonparametric test which does not assume a particular distribution of data. The Chi-square test [11, 15] equation can be presented as follows,

$$X^2 = \sum \frac{(O - E)^2}{E} \tag{1}$$

where O represents the observed values and E represents the expected values. The expected values equation can be presented as follows,

$$\mathrm{E(Y)} = \mathrm{Y_1P_1} + \mathrm{Y_2P_2} + \mathrm{Y_3P_3} + \ldots + \mathrm{Y_NP_N} \tag{2}$$

where Y represents the observed values (e.g., Observed values for VOC) and P represents the probability of Y. The **Chi-square** test is intended to test how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Entropy
The entropy of a discrete random variable X [13, 15–17] is defined in terms of probability of observing a particular value x of X as

$$\mathrm{E(X)} = -\sum_{x} P(X = x) * Log_2(P(X = x)) \tag{3}$$

The entropy is used to describe the diversity of a variable or vector. The more diverse a variable or vector is, the larger entropy they will have. Generally, vectors are more diverse than individual variables, hence have larger entropy. In this work, a filter based feature selection is used since this method is fast, scalable and independent of the classifier [12, 13], which is useful in our domain of application research since there is no dependencies between feature selection and classifier.

4 Results and Analysis

4.1 Sweat Collection

See Table 2.

Table 2. Sample of data collected that indicates gases detected on 4 persons

Gas code	Gas name	Number of appearance that gas detected			
		Person 1	Person 2	Person 3	Person 4
67-56-1	Methyl Alcohol (*)	0	5	5	2
64-17-5	Ethanol	3	5	4	4
67-63-0	Isopropyl Alcohol	5	0	0	0
71-23-8	1-Propanol	3	2	5	0
14898-79-4	2-Butanol, (R)-	2	2	2	0
78-83-1	1-Propanol, 2-methyl-	3	5	5	5
71-36-3	1-Butanol	2	2	1	0
71-41-0	1-Pentanol	3	0	3	1
111-27-3	1-Hexanol	3	2	1	1
104-76-7	1-Hexanol, 2-ethyl-	4	2	2	0
143-08-8	1-Nonanol	3	0	0	0
41.7	1-Decanol	3	0	0	0

4.2 Chi-Square Test

Chi test is applied to filter the accepted and rejected VOC based on number of appearance. Based on the values obtained in the Chi-Square test, 62 VOCs are considered as valid measurement based on the following parameters, degree of freedom = 1, critical value = 2.706. However, only 39 VOCs out of 198 are selected and accepted by Chi Test in which the number of gas appearances is more than the number of observations, and Table 3 shows all the VOCs.

Table 3. 39 gases are shortlisted based on the Chi-square test

No	Gas name	X^2	No	Gas Name	X^2
1	Ethanol	0.500	21	Furan, 2-methyl-	0.000
2	1-Propanol, 2-methyl-	0.667	22	Furan, 3-methyl-	1.588
3	Acetaldehyde (*)	0.000	23	Furan, 2-pentyl-	0.158
4	2-Propenal	0.000	24	Acetone	0.000
5	Propanal	2.077	25	Methyl vinyl ketone	2.571
6	Propanal, 2-methyl-	0.667	26	2,3-Butanedione	1.800
7	Methacrolein	0.000	27	2-Butanone	0.000
8	Butanal, 3-methyl-	0.667	28	3-Penten-2-one	0.000
9	2-Butenal, (E)-	0.000	29	2-Pentanone	0.000
10	Pentanal	0.733	30	3-Pentanone	1.800
11	2-Butenal, 3-methyl-	1.500	31	2-Butanone, 3,3-dimethyl-	0.000
12	Benzaldehyde	0.231	32	3-Hexen-2-one, 5-methyl-	0.667
13	Octanal	0.733	33	2-Heptanone	2.077
14	Hexane, 2-methyl-	0.273	34	Cyclopentanone, 2-methyl-	0.647
15	Phenol	0.667	35	5-Hepten-2-one, 6-methyl-	0.158
16	Cyclopropane, 1,1-dimethyl-	1.500	36	5,9-Undecadien-2-one, 6,10-dimethyl-, (Z)-	0.733
17	Acetic acid, methyl ester	2.455	37	Disulfide, dimethyl	0.176
18	Ethyl ether	1.588	38	Eucalyptol	0.000
19	1,3-Dioxolane, 2-methyl-	0.176	39	Formic acid, 1,1-dimethylethyl ester	1.727
20	Furan	0.000			

4.3 Filtering Using Entropy

The final VOCs selection criteria should be accepted by Chi-Square test and having entropy values less than 0.5 (in order to ensure only gases with high number of appearances are detected based on the entropy values). Table 4 lists all the final VOCs that are considered critical and can be used for odor classification.

Figure 2 shows the relationship between the Chi-square test and also the entropy for the gasses that are selected. Table 4 indicates that there are 10 gases that are detected having zero values for both the entropy and chi-square test and these gases are considered the strongest candidates that can be used for odour detection and classification.

Table 4. Final list of 17 accepted VOCs.

No	Gas name	Entropy	Chi-square test
1	Acetaldehyde (*)	0.000	0.000
2	2-Propenal	0.000	0.000
3	Methacrolein	0.000	0.000
4	2-Butenal, (E)-	0.000	0.000
5	Furan	0.000	0.000
6	Furan, 2-methyl-	0.000	0.000
7	Acetone	0.000	0.000
8	2-Butanone	0.000	0.000
9	3-Penten-2-one	0.000	0.000
10	2-Pentanone	0.000	0.000
11	Furan, 2-pentyl-	0.286	0.158
12	5-Hepten-2-one, 6-methyl-	0.286	0.158
13	1-Propanol, 2-methyl-	0.469	0.667
14	Propanal, 2-methyl-	0.469	0.667
15	Butanal, 3-methyl-	0.469	0.667
16	Phenol	0.469	0.667
17	3-Hexen-2-one, 5-methyl-	0.469	0.667

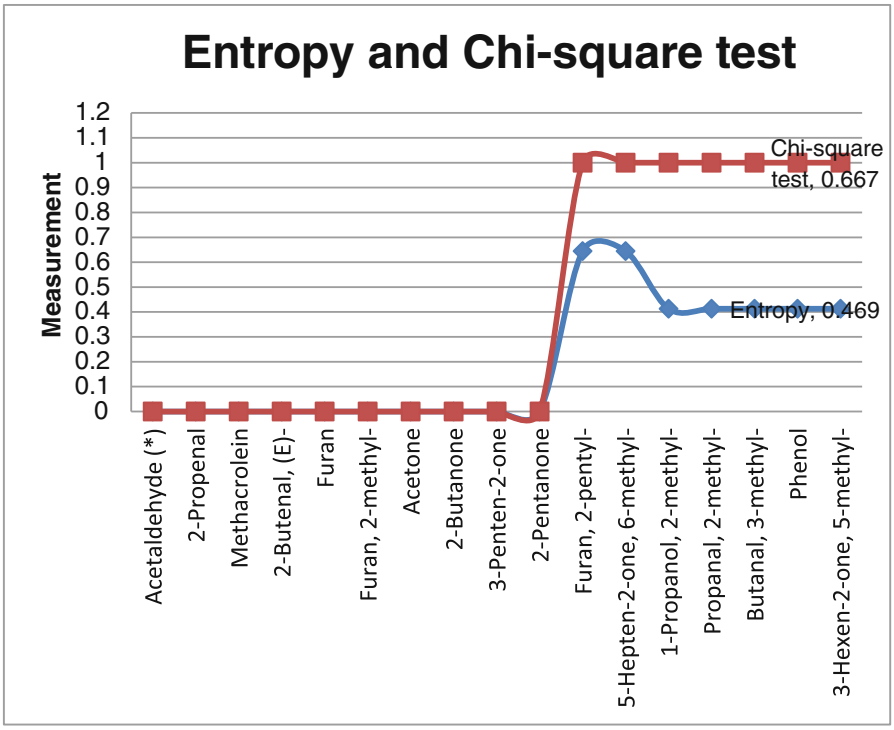


Fig. 2. Relationship between the entropy and Chi-square test values for all 17 gas

5 Conclusion

In this paper, feature selection methods are applied in order to select the appropriate features. Two feature selection methods, Chi-Square test and entropy, are chosen and applied on a list of VOCs emitted from human sweat. In this paper, the conducted experiments have successfully discovered relevant VOCs which can be used as stable VOCs emitted from human. A total of 17 stable VOCs are extracted from 198 VOCs. We have tested VOCs from human sweat for several days which give a complete list of VOCs emitted for the same person continuously. We examined all VOCs in two level of calculation using the Chi-square test and also the entropy.

The results of this paper can be applied for further investigation to perform studies related to human odor such as human detection from odor, gender detection from odor, age detection from odor, diseases detection from odor. Minimizing the 17 VOCs is one of the future target missions.

References

1. Kasap, B., Schmuker, M.: Improving odor classification through self-organized lateral inhibition in a spiking olfaction-inspired network. In: IEEE EMBS Conference on Neural Engineering (NER 2013) (2013)
2. Omato, S., Araki, H., Fujinaka, T., Yano, M.: Intelligent classification of odour data using neural networks. In: ADVCOMP 2012. Osaka Institute of Technology (2012)
3. Omato, S.: Odor classification by neural networks. In: The 7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Berlin, Germany, 12–14 September 2013
4. Gallagher, M., Wysocki, C.J., Leyden, J.J., Spielman, A.I., Sun, X., Preti, G.: Analyses of volatile organic compounds from human skin. Br. J. Dermatol. **159**, 780–791 (2009)
5. Chansri, C., Srinonchat, J.: Personal Shirt Odor Classification Using an Electronic Nose. Rajamangala University of Technology Thanyaburi (2014)
6. Wongchoosuk, C., Lutz, M., Puntheeranurak, T., Youngrod, T., Phetmung, H., Kerdcharoen, T.: Identification of people from armpit odor region using networked electronic nose. J. Name Stand., in press
7. Voss, A., Witt, K., Fischer, C., Reulecke, S., Poitz, W., Kechagias, V., Surber, R., Figulla, H.R.: Smelling heart failure from human skin odor with an electronic nose. In: 34th Annual International Conference of the IEEE EMBS San Diego, California USA, 28 August–1 September 2012. IEEE
8. Brownlee, J.: Introduction to feature selection. In: Machine Learning Process, 6 October 2014
9. Jović, A., Brkić, K., Bogunović, N.: A review of feature selection methods with applications. In: IEEE 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (2015)
10. Kasap, B., Schmuker, M.: Feature subset selection for cancer classification using weight local modularity, Sci. Rep. **6** (2016). Article ID 34759. https://doi.org/10.1038/srep34759
11. Balakrishnan, N., Voinov, V., Nikulin, M.S.: Chi-Squared Goodness of Fit Tests with Applications. eBook (2013). ISBN 9780123977830
12. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007). https://doi.org/10.1093/bioinformatics/btm344
13. Peng, Y., Wu, Z., Jiang, J.: A novel feature selection approach for biomedical data classification. J. Biomed. Inform. **43**, 15–23 (2010)
14. Haury, A.-C., Gestraud, P., Vert, J.-P.: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One **6**(12), e28210 (2011)
15. Novaković, J., Strbac, P., Bulatović, D.: Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav J. Oper. Res. **21**(1), 119–135 (2011)
16. Azhagusundari, B., Thanamani, A.S.: Feature selection based on information gain. Int. J. Innov. Technol. Exploring Eng. (IJITEE) **2**(2) (2013). ISSN 2278-3075
17. Adel, A., Omar, N., Al-Shabi, A.: A comparative study of combined feature selection methods for Arabic text classification. J. Comput. Sci. **10**(11), 2232–2239 (2014)

Combining Sampling and Ensemble Classifier for Multiclass Imbalance Data Learning

Mohd Shamrie Sainin[1(✉)], Rayner Alfred[1], Fairuz Adnan[2], and Faudziah Ahmad[2]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), 88400 Kota Kinabalu, Sabah, Malaysia
{shamrie, ralfred}@ums.edu.my
[2] Data Science Research Lab, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok Kedah, Malaysia
fudz@uum.edu.my

Abstract. The aim of this paper is to investigate the effects of combining various sampling and ensemble classifiers on the prediction performance in addressing the multiclass imbalance data learning. This research uses data obtained from the Malaysian medicinal leaf images shape data and three other large benchmark datasets in which seven ensemble methods from Weka machine learning tool were selected to perform the classification task. These ensemble methods include the AdaboostM1, Bagging, Decorate, END, MultiboostAB, RotationForest, and stacking methods. In addition to that, five base classifiers were used; Naïve Bayes, SMO, J48, Random Forest, and Random Tree in order to examine the performance of the ensemble methods. Two methods of combining the sampling and ensemble classifiers were used which are called the Resample with ensemble classifier and SMOTE with ensemble classifier. The results obtained from the experiments show that there is actually no single configuration that is "one design that fits all". However, it is proven that when using the sampling and ensemble classifier which is coupled with Random Forest, the prediction performance of the classification task can be improved on the multiclass imbalance dataset.

Keywords: Ensemble · Sampling · Multiclass · Imbalance · Random Forest

1 Introduction

In multiclass data, the performance of a classifier degrades when imbalance data exists. Imbalanced datasets occur when one of the classes has significantly less number of examples compared to the other classes. An ensemble approach, which combines several single classifiers, is one of the methods used to solve imbalanced multiclass classification tasks. However, it remains unknown how the performance accuracy of the ensemble methods is influenced by the selected base classifiers coupled with different sampling techniques used. An ensemble classifier model is proposed by combining the advantages data-level approach (sampling) and algorithm level approach

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 262–272, 2018.
https://doi.org/10.1007/978-981-10-8276-4_25

(ensemble classifiers). This may enable the enhancement of ensemble classifier performance in classifying multiclass imbalance data.

Thus, the proposed framework of the ensemble classifier model consists of two parts which are combined together. These two parts include various sampling techniques and various ensemble classifiers, including the various types of base classifiers used for the proposed ensemble classifier. The aim of this paper is to investigate the effects of combining various sampling and ensemble classifiers on the prediction performance in addressing the multiclass imbalance data learning. Thus, the objectives of this research include (1) Proposing and outlining the framework that combines various sampling techniques and ensemble classifiers; (2) Evaluating the prediction performance of the proposed combination framework. The study is divided into four phases. Phase 1 involves the activities for data acquisition. In phase 2, data will be sampled using the Resample and SMOTE. Weka will be used to perform the sampling process. In phase 3, each sampled data will then be used as training and testing datasets for the modelling process of the selected classifiers. In phase 4, the results will be analyzed.

This paper is organized as follows. Section 2 briefly discusses the related works in learning imbalanced datasets, followed by the research methodology described in Sect. 3. The experimental results are discussed in Sect. 4 and finally Sect. 5 concludes this paper.

2 Literature Review

Data is said to be imbalanced if there exists unequal distribution between its classes, to analyze and learn from such large data is a challenge. Researcher in [1] stated that if imbalance ratio in a general classification problem is no less than 19:1 with the size of minority class is only 5% of the entire size of the data, and then it is called as a highly imbalanced classification problem.

While the existence of imbalance is experienced in two-folds, in a case when one class is outclass the other reached out to numerous classes the impacts of imbalance are much more difficult. The imbalance deficiency experienced in multiclass is in this manner fascinating for two imperative reasons. In the first place, most learning algorithms do not manage the large varieties of challenges presented in the difficulty multi class. Second, various classifiers do not effectively reach out to the multi class space. Along these lines, there are few works conducted that address the imbalance issues in multiclass [2].

Class imbalance has been discovered as an important issue in the machine learning and data mining. This problem takes place when there is no even distribution of the training data among classes, whereby, when a classes is noticeably larger than the other, then, the data set becomes imbalanced. The majority class usually tends to overshadow the standard classifier by overlooking the minority class examples, which

is providing unwanted and also unsatisfactory classification performance. Therefore, the traditional algorithms need to be improved for a better handling of the imbalance data [3].

In addition, the imbalanced data distribution is often accompanied by high dimensionality in real-world data sets such as text classification and bioinformatics [4]. Therefore, the re-sampling methods or even cost sensitive learning extremely improves the classification performance on the imbalanced data to some degree [5]. With the existence of imbalance and multiclass data at the same time, traditional machine learning methods do not perform well since they generally assume relatively balanced distribution of data and significant class is of much more attention [6].

There are two approaches in handling this problem which are data-level and algorithm-level approach. In data-level approach, sampling methods including under-sampling and oversampling have their own drawback mainly loss in information and duplicating instances [7]. Thus combining them will balance the loss and duplication. In algorithm-level approach, modification of algorithm can be considered complex and not necessarily achieve high classification. Single classifier does not improve the classification accuracy in imbalance multiclass data problem. In order to overcome this problem, combining diverse classifier has widely been recognized. Although that combination of classifiers yield prediction accuracy improvement, however not all combining approaches are successful at producing multiple classifiers with good classification accuracy [8]. Thus, the best combination of sampling method and ensemble classifier would suggest an acceptable solution to the multiclass imbalance problem. Thus, further work in determining the ensembles is required.

3 Methodology

The methodology consists of Phase 1 (data acquisition), Phase 2 (sampling using Resample and SMOTE), Phase 3 (training and testing of the combination of sampling and ensemble classifiers), and Phase 4 (comparison). WEKA tool software program version 3.8 is adopted in order to implement the study in this paper.

3.1 Data Acquisition

In this paper, secondary data that is generated in [9] is used, whereby the dataset is obtained from villages situated in Perlis state, where 65 leaf pieces are randomly selected from specified leaf species. Table 1 illustrates the list of leaf species which are selected for this study and Table 2 illustrates the description of the data.

Table 1. Sample leaf species datasets

Class	Leaf	Name	Train	Test
1		Cemumar (CM)	11	4
2		Kapal Terbang (KP)	12	4
3		Kemumur Itik (KI)	11	4
4		Lakom (LK)	5	4
5		Mengkudu (MK)	6	4
		Total	45	20

Table 2. Medicinal leaf dataset information.

Description	Value #
#Examples	65
#Attributes	624
#Training	45
#Testing	20
#Majority	12
#Minority	5

3.2 Sampling and Classifier Algorithm

Two sampling methods were used in this step which are Resample and SMOTE. In Resample method, without replacement and with replacement options were investigated while SMOTE is applied using single and double SMOTE application. SMOTE with single application ensures that minority class is oversampled and SMOTE with double application runs with two SMOTE filters in order to balance the data. Seven ensemble classifiers were used in this project which is AdaboostM1, Bagging, Decorate, END, MultiBoostAB, RotationForest and Stacking. The base classifier for the ensemble methods are chosen from naïve Bayes (NB), Sequential Minimum Optimization (SMO), Decision Tree (J48), Random Forest (RF) and Random Tree (RT).

4 Experimental Results

There are three experiments conducted which are (1) experiments with training and testing data distribution, (2) 5-cross validation experiments using best base classifier identified from experiment 1, and (3) benchmark large data with multiclass imbalance.

4.1 Experiment with Training and Testing Data

In Table 3, Resample method (with replacement) and ensemble classifier shows that Decorate provides the best average performance among base classifiers in use where RF gives the best base classifier. Almost all ensemble classifiers are suitable with RF with 75% accuracy, except for Stacking method.

Resample method without replacement was then investigated and the results are shown in Table 4. Based on the results obtained, using Resample technique with ensemble classifier and with the option with replacement set to false, it can be observed that the prediction accuracy is higher as shown in Table 4. In this environment setup, MultiBoostAB performed better in average for all base classifiers combined with. In other words, MultiBoostAB is best combined with RandomForest as it produces higher performance result at 70% accuracy. It can be seen that the setup having Resample with replacement provided slightly lower prediction accuracy compared to the design having Resample without replacement. This is not quite suitable to be used as solution to the multiclass imbalance problem.

Table 3. Resample with ensemble classifier (with replacement).

	NB	SMO	J48	RF	RT	Average
AdaBoostM1	45	40	60	75	45	53
Bagging	45	50	60	75	55	57
Decorate	50	55	60	75	65	61
END	45	45	45	75	65	55
MultiBoostAB	45	50	55	70	45	53
RotationForest	45	50	60	75	65	59
Stacking	55	50	45	55	50	51
Average	47.14	48.57	55.00	71.43	55.71	

Table 4. Resample with ensemble classifier (without replacement).

	NB	SMO	J48	RF	RT	Average
AdaBoostM1	50	60	70	70	45	59
Bagging	60	50	50	70	60	58
Decorate	50	55	55	70	50	56
END	50	60	55	70	55	58
MultiBoostAB	65	60	65	70	60	64
RotationForest	60	60	55	70	50	59
Stacking	65	55	50	70	55	59
Average	57.14	57.14	57.14	70.00	53.57	

Setting the sample size of 150% (as shown in Table 5), the combination of Decorate and RotationForest, as the ensemble classifier and RandomForest (RF) as base classifier, produced the performance accuracy of 80%. Based on Table 5, RF is still the best base classifier for the sampling and ensemble classifier method where it produces an average of 71.43%.

Table 5. Resample with Ensemble Classifier + Sample Size 150% (with replacement).

	NB	SMO	J48	RF	RT	Average
AdaBoostM1	50	60	60	**70**	50	58
Bagging	50	40	60	**65**	60	55
Decorate	45	55	55	**80**	60	59
END	45	55	50	**75**	50	55
MultiBoostAB	50	60	50	**70**	50	56
RotationForest	**60**	**50**	**80**	**80**	**45**	**63**
Stacking	55	40	65	**60**	60	56
Average	50.71	51.43	60.00	**71.43**	53.57	

Table 6 outlines the results obtained the SMOTE sampling method is used on the multiclass datasets. In these experiments, the SMOTE sampling method is used with its default settings and only single SMOTE is applied. The ensemble classifier Stacking performed the best having the highest average accuracy of 61% when combined with the base classifier. However, as presented in the results, only END with RF produced the prediction accuracy at 75%. This shows that the SMOTE sampling is not quite suitable to be used as the sampling method combined with the ensemble for the data.

Table 6. SMOTE (Single) with ensemble classifier.

	NB	SMO	J48	RF	RT	Average
AdaBoostM1	40	55	60	**70**	35	52
Bagging	55	50	55	**70**	40	54
Decorate	55	60	55	**70**	60	60
END	50	55	55	**75**	40	55
MultiBoostAB	55	55	50	**65**	35	52
RotationForest	60	50	60	**60**	55	57
Stacking	65	65	55	**60**	60	61
Average	54.29	55.71	55.71	**67.14**	46.43	

Table 7 stipulates the results obtained for the tested data using the double SMOTE (DSMOTE) sampling technique. When compared to the results obtained using the single SMOTE shown in Table 6, the improvement of prediction accuracy percentage of the second classifier for J48, and RandomTree (RT) with increment of 2.14% and 2.71% respectively is still below than average of RF percentage in Table 6 at 66.43% as base classifier. Though the Decorate ensemble performed better in term of the average prediction accuracy per combination with second classifier at 60% better than END and RotationForest, the combination with other classifiers is lower than that of Table 6.

Table 7. DSMOTE (Double) with ensemble classifier.

	NB	SMO	J48	RF	RT	Average
AdaBoostM1	45	55	55	**65**	50	54
Bagging	60	40	60	**60**	40	52
Decorate	**55**	**50**	**60**	**75**	**60**	**60**
END	40	50	45	**75**	30	48
MultiBoostAB	50	55	50	45	50	50
RotationForest	65	55	60	**75**	35	58
Stacking	55	65	50	**70**	55	59
Average	52.86	52.86	54.29	**66.43**	45.71	

Table 8 shows the best possible combination of sampling technique with ensemble classifier. Although that the accuracy of the classifiers are almost similar, however class accuracy varies. Some may good at majority but poor performance on minority class as indicated by the f-measure values. Sampling combined with ensemble with replacement using 150% setting, where the ensemble classifier is Rotation Forest (Random Forest as base classifier) and Decorate (also with Random Forest base classifier) provide the overall good performance on all classes.

Table 8. Detailed comparison (f-measure) for sampling and ensemble classifier.

Sampling	Classifier	%	CM	KP	KI	LK	LK
Resample + Single	RandomForest	75	0.500	0.800	0.750	0.667	1.000
Ensemble + With Replacement	Decorate + RF	75	0.333	1.000	0.667	0.667	1.000
	Bagging + RF	75	0.400	0.800	0.727	0.667	1.000
	END + RF	75	0.400	0.800	0.727	0.667	1.000
	RotationForest + RF	75	0.400	0.800	0.727	0.667	1.000
Ensemble + With Replacement (150%)	END + RF	75	0.286	0.889	0.800	0.667	1.000
	RotationForest + RF	**80**	**0.571**	**0.800**	**0.889**	**0.667**	**1.000**
	Decorate + RF	**80**	**0.571**	**0.800**	**0.889**	**0.667**	**1.000**
Ensemble + SMOTE	END + RF	75	0.333	0.889	0.727	0.667	1.000

4.2 5-Cross Validation

Results indicated that combination of sampling with replacement and ensemble classifier is giving the best classification performance as shown in Table 8. In order to see another perspective of the results, Table 9 shows the performance results starting from original data and different Weka's sampling algorithms based on 5-cv.

Based on Table 9, resample method (with uniform and replacement setting) gives the best average classification accuracy. Specifically, Random Forest and followed by Bagging, Decorate and Rotation Forest with Random Forest as base classifier are best the combination of ensemble classifier sampling. This shows that sampling is one of the powerful methods to address the multiclass imbalance problem. Interesting finding is

Table 9. Performance of classifiers based on 5-cv.

Ensemble classifier	Original data	Class balancer	Resample (uniform)	SMOTE	DSMOTE	Average
Random Forest	73.85	69.83	**96.92**	74.32	78.57	78.70
AdaboostM1 + RF	70.77	75.08	93.85	77.03	77.38	78.82
Bagging + RF	73.85	78.92	95.38	77.03	83.33	81.70
Decorate + RF	75.38	80.25	95.38	81.08	83.33	83.09
END + RF	75.38	79.00	95.38	82.43	84.52	83.34
MultiBoostAB + RF	70.77	75.08	90.77	77.03	77.38	78.21
RotationForest + RF	75.38	75.00	95.38	78.38	82.14	81.26
Stacking + RF	64.62	67.36	90.77	70.27	78.57	74.32
Average	72.50	75.07	**94.23**	77.20	80.65	

that the random forest performs very well considering this algorithm is not implemented with specific meta classifier. However it is actually known that random forest is an ensemble algorithm where it is a meta estimator that fits certain number of decision trees. The result in these experiments may not address the overfitting, thus in the next experiments, other benchmark data are used in order to examine the performance of similar methods as presented in Table 9, in order to assess the performance of the methods.

4.3 Benchmark Dataset

In this phase, the experiments are conducted in order to compare the predictive accuracy of the selected methods on the benchmark datasets. The selected large datasets comprise of imbalance multiclass (high imbalance ratio) as listed in Table 10.

Table 10. Benchmark dataset with high imbalance ratio and large data.

Data	#S	#A	Attribute properties	#C	Min	Max	Ratio	Previous result
Statlog (Landsat)	6345	36	Integer	6	56	1072	1:19	89.3% [10]
PageBlocks	5473	10	Real, Integer	5	28	4913	1:175	97.28% [11]
Statlog (Shuttle)	58000	9	Integer	6	10	45586	1:4559	96.3% [12]

Note: #S: Number of samples, #A: Number of attributes, #C: Number of classes

First, the performances of ensemble classifiers are investigated and recorded on the original datasets. The results of the experiment are presented in Table 11. Based on the results obtained, almost all of the ensemble classifiers performed similar with improved accuracy compared to the previous results in the respective research. AdaBoostM1 + RF produced higher accuracy and fairly fast on Landsat and Shuttle datasets but drop 0.17% for PageBlock dataset. On the other hand, Decorate combined with RF is a good

ensemble design, however its processing time is too high which is not good for large dataset problem. Specifically in Shuttle dataset where the classifier could not finish the experiment due to a very long processing time and high memory load error, thus marked as 'x' as noted in Table 11. More experiments on down-sampling or updateable classifiers can be investigated on this large dataset in future.

Table 11. Classification performance of ensemble classifier on benchmark dataset.

	Landsat	Time	Shuttle	Time	PageBlocks	Time
Random Forest	95.31	6.98	99.99	7.55	**97.35**	**0.84**
AdaboostM1 + RF	**95.64**	**1.18**	**99.99**	**5.80**	97.11	5.25
Bagging + RF	95.36	10.88	99.99	55.95	97.11	3.90
Decorate + RF	95.20	132.67	x	x	97.53	148.60
MultiBoostAB + RF	95.64	1.40	99.99	7.34	96.97	6.06
RotationForest + RF	95.03	16.74	99.97	107.67	97.28	6.31

The next experiment involves steps that investigate the performance of the methods that combine sampling and ensemble classifier. Using this approach, it will reduce the amount of data for learning and also balance the distribution of the sample in each class and the results are presented in Table 12.

Table 12. The performance of combined sampling and ensemble classifier.

	Landsat	Time	Shuttle	Time	PageBlocks	Time
Random Forest	95.64	0.73	**99.92**	**1.72**	**99.12**	**0.33**
AdaboostM1 + RF	95.80	0.78	99.92	1.84	99.01	4.53
Bagging + RF	95.58	6.44	99.92	1.84	99.01	2.83
Decorate + RF	95.75	82.86	99.83	734.64	99.07	129.97
MultiBoostAB + RF	**98.33**	**0.75**	99.92	1.80	99.01	5.06
RotationForest + RF	98.02	0.73	99.93	31.05	99.05	4.36

Based on the result obtained, the methods have improved the accuracy performance and processing time on Landsat and PageBlocks, while there is a very small performance drop in Shuttle dataset. Although the performance of the method in Shuttle dataset is not improved compared to when using original data, however the processing time has improved. Ensemble using Random Forest performed good for Shuttle and PageBlocks while MultiBoostAB + RF has improved in Landsat dataset to 98.33% with processing time 0.75 s (from 95.64% and 1.40 s processing time). Again the ensemble method using Decorate is not a good design for large datasets due to the high processing time.

Finally, in most cases from the experiments, RF shows that it performed better as classifier or as a base classifier for another ensemble methods. As discussed in [13], RF is a combinations of several decision trees which constructed using random process.

RF is also known as an ensemble methods of random decision trees [14], combining the predictions of the individual trees. Previous research such as [15, 16] proved that RF is the better classifier in their analysis. Thus, the results of RF in this study is another support that it is a strong classifier even combined as a base classifier for other ensemble methods.

5 Conclusion

Data from medicinal leaf images (shape data) and large multiclass imbalance benchmark data were trained and tested using a combination of sampling techniques coupled with the ensemble classifiers. While the combinations tested in this study were performed almost similar, there is no single combination that best fit to the problem. However, if there is no modification to the data, ensemble classifier with Random Forest is able to perform better than any single classifier. Furthermore, when Random Forest is combined as meta classifier such as AdaBoostM1 or MultiBoostAB, and sampling (Resample or SMOTE), the performance can be further improved.

Acknowledgement. This work was supported by FRGS under Grant No. 13142 (2014).

References

1. Ding, Z.: Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics. Ph.D. thesis, Georgia State University (2011)
2. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**(4), 463–484 (2012)
3. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013)
4. Bekkar, M., Akrouf, A.T.: Imbalanced data learning approaches review. Int. J. Data Min. Knowl. Manag. Process (IJDKP) **03**(04), 15–33 (2013)
5. Krawczyk, B., Wozniak, M., Schaefer, G.: Learning from imbalanced data: open challenges and future directions. Appl. Soft Comput. **14**, 554–562 (2014)
6. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. In: Proceedings of the Third International Conference on Soft Computing for Problem Solving Advances in Intelligent Systems and Computing, vol. 258, pp. 589–600 (2014)
7. Qiang, W.: A hybrid sampling SVM approach to imbalanced data classification. Abstr. Appl. Anal. (2014). Article ID 972786
8. Abdullah, Ku Ruhana, K.-M.: Ant system-based feature set partitioning algorithm for k-NN and LDA ensembles construction. In: Proceedings of 5th International Conference on Computing and Informatics (ICOCI), pp. 326–332 (2015)
9. Sainin, M.S., Ghazali, T.K., Alfred, R.: Malaysian medicinal plant leaf shape identification and classification. In: Proceedings of Knowledge Management Conference and Exhibition (KMICE 2014), Langkawi, Malaysia (2014)

10. Ghosh, S., Biswas, S., Sarkar, D., Sarkar, P.: A tutorial on different classification techniques for remotely sensed imagery datasets. Smart Comput. Rev. **4**(1), 34–43 (2014)
11. Eschrich, S., Chawla, N.V., Hall, L.O.: Generalization methods in bioinformatics. In: Proceedings of the 2nd International Conference on Data Mining in Bioinformatics (BIOKDD 2002), pp. 25–32 (2002)
12. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 1553–1567 (2004)
13. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
14. Vens, C.: Random forest. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (eds.) Encyclopedia of Systems Biology, pp. 1812–1813. Springer, New York (2013)
15. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H.: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics **19**(13), 1636–1643 (2003)
16. Kamanksha, D. P., Sanjay, A.: A critical analysis of twitter data for movie reviews through random forest approach. In: Information and Communication Technology for Intelligent Systems (ICTIS 2017), vol. 2, pp. 454–460 (2017)

Utilizing Smartphone and Tablet for Appliances Mobile Controller System

Aslina Baharum[1(✉)], Nurul Hidayah Mat Zain[2], Ismassabah Ismail[2], Chew Yun Fai[1], Siti Hasnah Tanalol[1], and Muhammad Omar[3]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
aslinabaharum@gmail.com, danny_931117@hotmail.com, hasnah@ums.edu.my

[2] Faculty of Computer and Mathematical Sciences, Universiti Technology MARA (UiTM), Cawangan Melaka, Kampus Jasin, 77300 Merlimau, Melaka, Malaysia
{nurulmz,isma}@tmsk.uitm.edu.my

[3] Faculty of Business Management, Universiti Teknologi MARA Sarawak, Kota Samarahan Campus, Jalan Meranek, 94300 Kota Samarahan, Sarawak, Malaysia
muhammad@sarawak.uitm.edu.my

Abstract. Reducing the electricity consumption is better for the planet and help to reduce harmful greenhouse releases. Hence, this paper proposes house appliances controller apps that can control and calculate the power usage of house appliances. Besides, the apps can control and set the auto-time based such as how long the light must be in switch on. This is important because the house might be a target for a thief when the light continuously on. The objectives of this paper are to identify and develop features of the mobile controller app and for the house appliances. The expected outcome would be a fully functional mobile application and web-based system namely as HomeBot.

Keywords: Mobile web application · Smart home
Appliances mobile controller system

1 Introduction

Mobile technology such as smartphones and tablets has transformed our lifestyle. Basically, smartphones and tablets are mini computers that have of features to make our lifestyle become easier. Nowadays, people are attached with smartphone. Hence, this paper comes out with the idea using mobile device to communicate with house appliances.

Heavy household's electricity consumption gives a big effect to the environment. The air-conditioner is one of the most electricity consumption of house appliances. A survey about the Malaysian behavior for natural ventilation in terraced houses had been done in year 2005. One of the findings from the survey was, nearly 62% of the respondents have at least one air-conditioner in their house [1]. The result also showed that there

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 273–283, 2018.
https://doi.org/10.1007/978-981-10-8276-4_26

is a significant relationship between mean monthly household income and number of air-conditioners among households [1].

Moreover, the study indicated that the more monthly income they earn, the more air-conditioners they have [2]. Besides, the overall crime rate is designated as high for Malaysia. Petty crime against expatriates is common, while violent crime remains relatively uncommon [3]. There has been a noticeable increase in crime in Kuala Lumpur in 2014. As well as, several reported assaults and robberies, sometimes involving weapons. Petty theft, particularly purse snatching and pickpocketing, and residential burglaries are the most common crimes committed against foreigners. Other types of non-violent criminal activity included credit card fraud and automobile theft [3]. It can be stated that Malaysia's overall crime and safety situation is placed at high risk and the concern from Government about crime is valid since the crime rates in Malaysia have been recorded high.

In this paper, we discussed the design and implementation of house appliances controller web based system including the mobile app. A mobile app, allows user access the application through the mobile as well as using web browser to access with the web based system. The web based system supports different platform devices, allowed with a minimum of effort to reach the widest audience. At the same time, it can be built cheap, fast and easy, while the mobile app will be supporting in android only.

This paper is structured as follows. In Sect. 2, the related work has been presented. In the next section, Sect. 3, the design of proposed mobile app and web-based system are presented. In Sect. 4, the usability testing method and result is discussed. Finally, the paperwork is summarized, and future work is revealed in the last section.

2 Related Work

To achieve the objectives of the proposed conception, we need to understand the basic elements of the existing system.

a. Conventional Paper-Based Systems: Smart Home System is widely using in modern country which can brings all the information of the house current situation and control the house appliances with just using the smartphone or tablet as a controller. The system included on or off switch button for house appliances control, power consumption monitoring system, set-time system and authentication. It is easy yet bringing convenient to people especially the communities that are disabled can easily control house appliances without large movement.

b. Existing System: With the advancement of technology, a different type of systems was launched in markets. Below are the example systems that currently used:

The Home Depot: wink. The wink app and hub at the Home Depot – Smart Home. Wink Team is bringing all the smart products, which created, by the Home Depot together around the home with one simple app [4]. This is the simple app named wink, which created, by Wink Team. Access this app need to login username and password to control the house appliances. The available house appliances that created by Home Depot that

linked with wink are almost 70 products. The app basically can control the home blinds, CCTV cameras, house lights, house locks and others.

Samsung Smart Home, SmartThings. Samsung Smart Home enables home devices to connect through a single integrated platform and provides a foundation for an emerging ecosystem of connected home services. Its unique functionality enables users to control and manage their home devices through a single application by connecting personal and home devices - from refrigerators and washing machines to Smart TVs, digital cameras, smartphones and even the wearable device Galaxy Gear [5] — through an integrated platform and server. This app consists of a lot of features that too complicated for users to familiar with this app.

3 Proposed System

The existing systems, *wink* app and *Samsung SmartThings* have the same functionality which is using mobile devices in android OS or iOS to control house appliances. The most unique function among these systems is Samsung SmartThings, which has an extra device to communicate with house appliances – Samsung Gear. However, the difference compared between the existed systems and proposed system is the proposed system is going to be developed is the power consumption monitoring system. This is important when come to reduce inefficient electricity used and to reduce the greenhouse effect.

3.1 Methodology

The discovery, design, development and delivery of information system are often linked together in a process labeled as a System Development Life Cycle (SDLC). There are different models which follow basic steps like an example Waterfall, Prototyping, Spiral, Iterative and incremental development, agile development, Rapid, Rapid prototyping and Evolution Model [6]. Each process model follows a series of steps unique to its type, to ensure success in the process of software development [6].

The Software Development Life Cycle methodology for this project will be Evolutionary Prototyping since evolutionary prototyping is based on the idea of developing an initial version of the system, exposing this to the user, refining this through many stages, until an adequate system has been developed [7]. Through this methodology, a prototype is a smaller version of the system with a minimal number of features and this prototype with the web-based and mobile app system will be developed.

The system's prototype will then be evaluated and given suggestions by the users. During the system prototype testing, users may give feedback and opinions to improve the system to be more user-friendly. The system's prototype will then be maintained based on the feedback and opinions gained from users until it becomes a completed and acceptance by majority of the users. At the end, this product should be a well-functioning web-based and mobile application system, which can be proposed to global. In Fig. 1 it shows the phase of the Evolution Prototyping methodology.

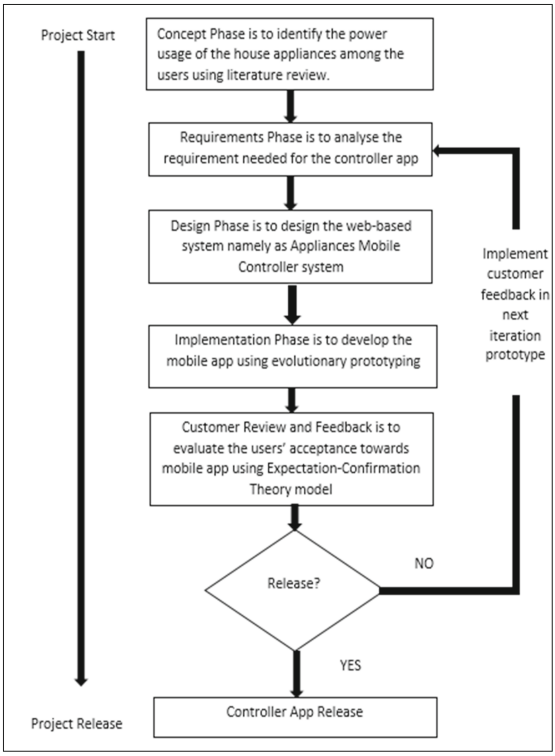


Fig. 1. The flow diagram for the evolutionary prototyping model for house appliances' control application.

3.2 System Architecture

The success or failure of a mobile application is depended on the different factors, including the device platform, application architecture, and the wireless connection. According to [8], different application models have the different features, making it suitable for the certain application, does not apply to the others.

In this section, we proposed the development of house appliances controller web based system and mobile app, integrated with the use of any platform devices or tablets that allow users' houses become smart home. The only needed from users is to have a web browser and Internet connection on their devices. For android platform devices, users must install the mobile app, HomeBot in their devices or open the web browser, and then they can access the house appliances controller system.

The proposed system follows typical client-server architecture. Advantages of this architecture include ease of deployment and support of thin client layer and the intrinsic scalability of the intermediate and data layers. The client-side is the mobile web browser installed on the devices while the server-side processes application requests and stores and retrieves data (Fig. 2).

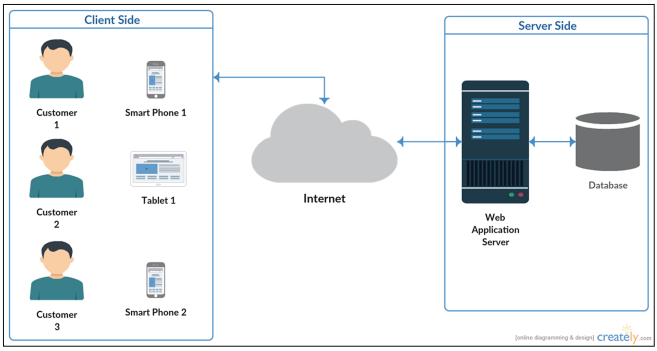


Fig. 2. House appliances controller system architecture

3.3 House Appliances Control Flow

The house appliances control flow of the app is illustrated in the activity diagram in Fig. 3. Each user has the default username and password without any registration needed. However, users can only change their passwords with the default usernames respectively after login into the system.

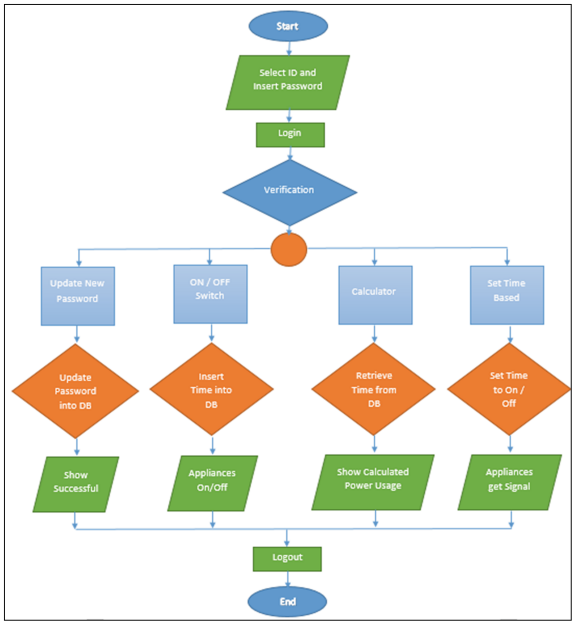


Fig. 3. Flow chart of house appliances controller system

After login into the system, the home page of the system consists of 4 modules, which are switch, set-time, calculator and admin. After entering switch page, users can select

which house appliance to switch on or off. While the set-time page is like the normal alarm system, users just need to enter the time and select the house appliance, and then on that specific time, the house appliance will either on or off depending on its condition. Meanwhile, calculator is the power consumption monitoring system, which will calculate the total power usage in kWh and total time duration in hours for the house appliances. Lastly, the admin page is for the users to only change passwords with the default usernames.

3.4 Functionalities for Customers

Figure 4 shows the user is required to login the password to access the mobile application. User only needs to enter password. Since the username is designed as drop-down list to be more user friendly without entering username and just need to select from the drop-down list.

Fig. 4. User login page

Fig. 5. Home page of HomeBot

Figure 5 shows the front page of the application being developed. The top of the front page has an image slider banner to aware users concern about the over consumption of electric of house appliances will cause the greenhouse effect. The most top of the page has a navigation bar. For user friendly, this navigation bar will always appear wherever page that users click into, users can still select the next page that wish to go through the

navigation bar without returning the previous page to select again. Then, at the bottom, there are 4 icons which are the main 4 modules about the HomeBot. Switch is for selecting the house appliances to switch on or off. Select the house appliance and turn it on or off (Fig. 6).

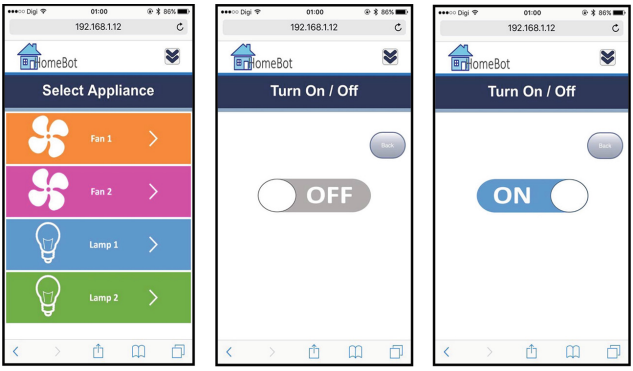


Fig. 6. Switch page

Set-time is for setting the time like alarm automatically on or off the house appliances (Fig. 7). First, users need to select the house appliance, and then select the time for the house appliance to turn on or off. After confirming the selection, press submit.

Calculator is the power usage monitoring system to monitor the electric consumption for every house appliances (Fig. 8). The top of the page has an image slider banner to show the current version of electricity rate from Tenaga Nasional Berhad (TNB main Malaysian energy provider). Then the bottom part is the power usage monitoring system. Once users select the house appliance and submit button is press, the details of the electric consumption of the house appliance will be shown.

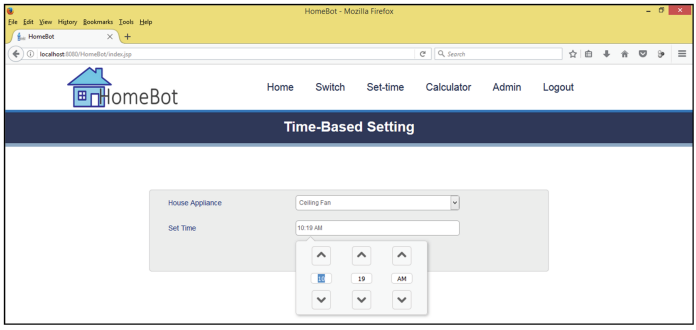


Fig. 7. Set-time page

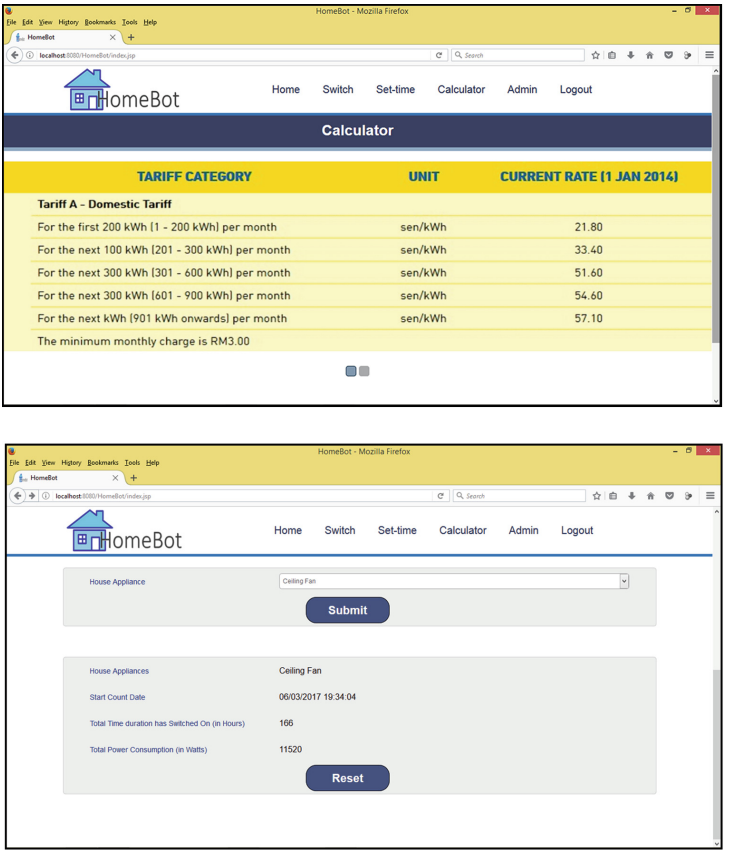


Fig. 8. Calculator page

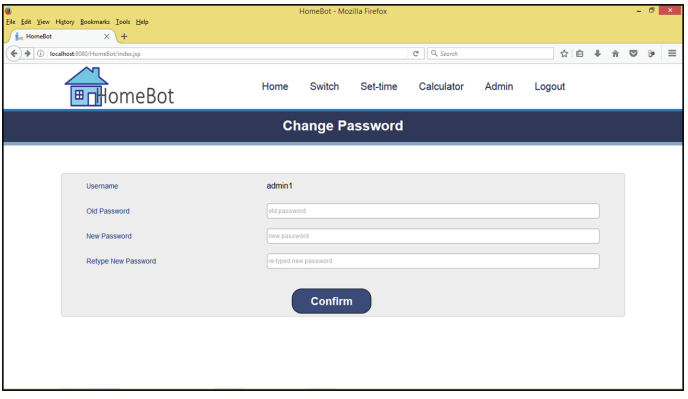


Fig. 9. Admin page

The last module is authentication which is for the users to change their passwords. Users are required to enter the old password before creating a new password (Fig. 9). After validation and verification is being done, users will be alerted by the system that the new password has been updated.

4 Usability Testing

When there is increasing number of users using technical devices to manage their everyday life, usability in software is important. Usability testing generally consists of evaluating the various software applications based on consideration of different field related to the application functionalities. It usually carried out in large user groups, and when the number of users increases, the probability of problem being solved increase.

4.1 Data Collection

Usability testing was conducted in Kota Kinabalu, Sabah, where 10 participants were chosen randomly. According to Krug and Nielson [9, 10], three to six users are sufficient to reveal the major usability issues. However, Nielson [11] stated that, 15 users are required to identify all of the issues. He pointed out, five users testing will reveal around 85% of the issues and is the most economical way. Yet, an individual test can unveil one-third usability problems, so a single test is better than not executing tests [9].

4.2 Method

Different usability testing methods have been developed to estimate the quality of user interfaces and to derive solutions for usability improvements. One of the usability testing methods is System Usability Scale (SUS) [12]. According to [12], SUS is an easy, reliable, ten-item scale, gives a total view of the availability of subjective evaluation. SUS alternated the questions in order to avoid response deviation. By take turns positive and negative statements, respondents will make the effort to read and think each statement, then decide whether agreed or disagreed with the statement [13]. In last 30 years, the SUS has been tested and investigated to use and has determined itself to be a reliable technique to evaluate the system usability [13]. According to [14], SUS is independent, does not relied on technology. It has been tested throughout the hardware, software, web technology and smartphone. The SUS has been cited by over 600 publications and it has proven that it can become an industry standard.

4.3 Results and Discussion

As mention previously, the SUS questions are alternating positive and negative statements. The odd-numbered questions are positive-framed questions and the even-numbered questions are negative-framed questions. Scale 1 to 5 is given to choose from based on the degree of agreement to the statement. 5 mean they strongly agree, 1 means they totally disagree. Figure 10 shows the total score for each question.

The SUS score of each participant had been calculated. Figure 11 shows the SUS score of the participants. The average SUS score is 89.5. As a conclusion, participants are acceptable to the mobile web app and will recommend it to their friends [15].

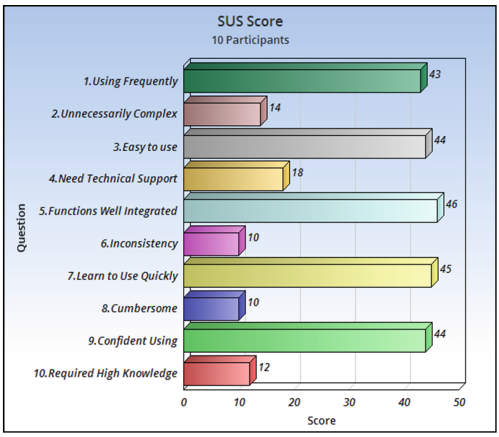


Fig. 10. SUS score of different question

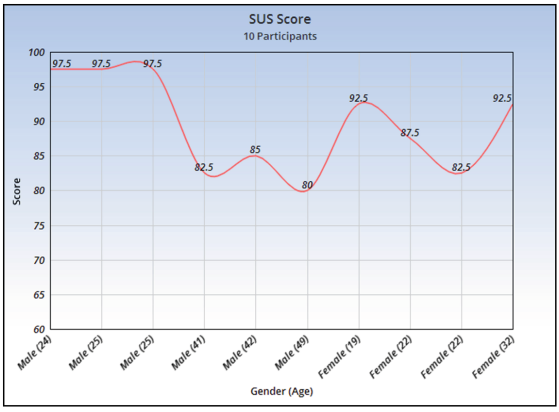


Fig. 11. SUS score of participants

5 Conclusion

Apart from calling and SMS, application in mobile phones nowadays is used to support daily activities of consumer. Developing an application to support cross-platform operating system or mobile device is not easy. The introduction of the mobile web application could solve this problem, where to design a single highly-branded responsive website or application that support by most of the smartphone and all operating system. The design of house appliances controller mobile app and web based system can help the

users especially the communities which are disabled can easily control the house appliances without any large movement. Besides, controlling the house appliances' power consumption is helping to reduce the greenhouse effect and users can control the house appliances through Internet connection from outstation to reduce the housebreaking cases. Future research will focus more on the usability testing. More data collection from diverse users may increase the accuracy of the utilization of mobile web application. The data collected can be analyzed by using IBM SPSS in future because the results are analyzed by the statistical approach, which is more robust.

Acknowledgements. Researchers are thankful to Universiti Malaysia Sabah (UMS) for the support of the resources and necessary facilities for the preparation of the research. This study is currently funded by a FRGS Grant from Ministry of Higher Education Malaysia (FRG0436-ICT-1/2016).

References

1. Kubota, T., Ahmad, S.: Energy efficient city in Malaysia wind flow in neighborhood areas. In: The 6th International Seminar on Sustainable Environment Architecture (SENVAR), pp. 1–10 (2005)
2. Gao, G.: A study of carbon emission right allocation under climate change. Adv. Clim. Chang. Res. **2**(6), 301–305 (2006)
3. OSAC. Malaysia 2015 Crime and Safety Report (2015)
4. Smart Home Open. The introduction of Wink App and Hub at the Home Depot, 22 September 2014. Retrieved from YouTube: https://www.youtube.com/watch?v=q2UsoMkn2HE
5. International Consumer Electronic Show. The Introduction of Samsung Smart Home, 9 January 2014. Retrieved from YouTube: https://www.youtube.com/watch?v=mEzSF29EBgI
6. Massey, V.: Comparing various SDLC models and the new proposed model on the basis of available methodology (2012)
7. Lenz, G., Moeller, T.: NET: A Complete Development Cycle. Addison-Wesley, Boston (2004)
8. Mallick, M.: Mobile and Wireless Design Essentials. John Wiley & Sons, New York (2003)
9. Krug, S.: Don't Make Me Think! A Common Sense Approach to Web Usability, 2nd edn. New Riders, Berkeley (2006)
10. Nielsen, J.: Usability 101: Introduction to Usability (2003). http://www.useit.com/alertbox/20030825.html. Accessed 2 May 2016
11. Nielsen, J.: Why You Only Need to Test With 5 Users (2000). http://www.useit.com/alertbox/20000319.html. Accessed 2 May 2016
12. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)
13. Brooke, J.: SUS: a retrospective. J. Usability Stud. **8**(2), 29–40 (2013)
14. Sauro, J.: A Practical Guide to the System Usability Scale: Background, Benchmarks, & Best Practices. Measuring Usability LLC, Denver (2011)
15. Thomas, N.: How to Use the System Usability Scale (SUS) to Evaluate the Suability of Your Website (2015). http://usabilitygeek.com/how-to-use-the-system-usability-scale-sus-to-evaluate-the-usability-of-your-website/. Accessed 9 May 2016

Dengue Fever Awareness Using Mobile Application: DeFever

Aslina Baharum[1(✉)], Siti Hasnah Tanalol[1], Jafhate Edward[1], Nordaliela Mohd. Rusli[1], Ismassabah Ismail[2], and Nurul Hidayah Mat Zain[2]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), 88400 Kota Kinabalu, Sabah, Malaysia
aslinabaharum@gmail.com, {hasnah,daliela}@ums.edu.my
[2] Universiti Teknologi MARA, Kampus Jasin, Merlimau, Melaka, Malaysia
{isma,nurulmz}@tmsk.uitm.edu.my

Abstract. Dengue Fever (DF) is the leading cause of illness and death around the world. In order to take action against this issue the involvement of the public and society is very important. However the public awareness towards DF is very low due to the lack of attention for this issue. Therefore, this research explores the feasibility to develop a mobile application, which is an informative mobile application that is designed to attract the public and society's attention and increase the awareness about DF. The mobile application, Dengue Fever Awareness using Mobile Application, DeFever, acts as an informative mobile application that can be used to provide more information about DF with some useful features. By using DeFever, all information regarding DF and infected areas will be accessible. The objectives of this research are to identify the level of public awareness based on the knowledge, attitudes and practices towards Dengue, to identify the suitable design features of DeFever and to develop the mobile app (DeFever) that helps to increase the public awareness towards DF. The method used in this research was by using the Appreciative Inquiry (AI) whereby the model used was the 5D Appreciative Inquiry Model for DeFever. AI is the cooperative, co-developmental search in finding the best in individuals of people, and the world around them. It includes deliberate revelation of what offers life to an association or a group when it is best and most skilled in financial, biological and human terms. Finally, hopefully DeFever can be used to assist efforts in increasing the public awareness level towards DF.

Keywords: Appreciative Inquiry · Mobile application · Awareness
Dengue Fever

1 Introduction

Keeping body fit and healthy all the time is very important. Dengue fever (DF) is the leading cause of illness and death in the tropics and subtropics, because one of four viruses transmitted by mosquitoes. Unfortunately, based on the current technology that we have now, there are still no vaccines exist to prevent infection with dengue virus. However, there are a lot of effective methods that we can use to avoid mosquitoes bites.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 284–293, 2018.
https://doi.org/10.1007/978-981-10-8276-4_27

Ever since its first appearance, dengue has emerged as a worldwide problem since 1950s, which took enormous human life [1]. The only problem that still becomes a downfall for us to apply and use the protective and precaution measures is awareness and the availability of knowledge about dengue.

Study showed that DF is on the rise and has become the viral infection of humans with an estimated 390 dengue infection occurs worldwide each year, with about 96 million reported in illness [2]. In comparison, Malaysia has reported estimation for about 90,000 number of dengue cases, approximately 20% higher than in 2014 as stated by the Malaysian Ministry of Health [3]. Sabah on the other hand, recorded for about 958 dengue fever cases and 662 cases in Kota Kinabalu, which includes 145 cases in Penampang and 38 cases in Putatan [4]. However, despite all of this warnings and vital statistics, it was not taken into concern by the society nowadays. This scenario showed that the lack of public and society awareness in DF. According to a report from Daily Express [5], it was shown that the public in Kota Kinabalu, Sabah is currently unaware of the current uprising of DF cases. Considering the victims of DF increased annually, the public is still not aware of these vital cases. This paper is organized as follows. In Sect. 2, the explanation of Knowledge, Attitude and Practice is discussed. In Sect. 3, Appreciative Inquiry (AI) model based on the AI is clarified. The results of this research is shown and compared in Sect. 4. There is an explanation of DeFever app according to their features in Sect. 5. Finally, this paper is summarized in the last section.

2 Knowledge, Attitude and Practice

The case study of Knowledge, Attitudes, Practices and Health Information-Seeking related to Malaria-Dengue Prevention in India [6], shows that the involvement of these three important factors can help to understand the pattern and also the behaviors of the societies. This is because these three factors can be used to understand the way people think about DF in term of their knowledge about the current issues, their attitudes about the issues and the practices to that issue. The findings of the case study show the significant and relevant results, which is crucial in their research. For example, as stated in their case study, what are the knowledge of malaria and dengue among Mumbai's resident? What are the attitudes towards preventive strategies related to Malaria and Dengue? What are the actual practices applied by the Resident of Mumbai? Therefore, as in this research, these three factors are used as it shows far greater and relevant data. In addition, a survey questionnaire is carried out using quantitative method based on these 3 factors; Knowledge, Attitude and Practice among public, specifically Kota Kinabalu, Sabah area towards DF.

3 Appreciative Inquiry Model

Appreciative Inquiry (AI) is a philosophical theory and a methodology used mostly by any organization in order to make their management much easier to manage. AI is also a procedure that asks into, distinguishes, and further adds to the best of "what is" in an organization and groups will always keep in mind the end goal in order to make a brighter

future [7]. In addition, it focuses more on the person's imaginative and innovative thinking in order to solve problems, instead of the person's negativity or criticism. Based on [8], "Appreciative Inquiry can get you much better results than seeking out and solving problems". AI can be used to get a solution faster than any other method like problem solving method. For example, when a group of researchers carries out a study about human problems and its incompatibility, the results was that both the harshness and number of these problems are growing. In the same way where the ideals of human, aspirations achievements, successes and accomplishment or any other kind of successes of a human tends to flourish. Therefore, AI can be recognized as a change efforts by purposely requesting that positive inquiries to make useful dialog and enlivened actions inside an associations or a communities.

Appreciate is defined as valuing, confirming over the past and present strengths, successes and potential. Inquiry on the other hand is defined as the demonstration of investigation, discovery and to be interested in seeing new potential and conceivable outcomes available. Therefore, it can come to sense and much more deeper meaning when both terms are combined, giving the term *Appreciative Inquiry* (AI) a multitude of meaning, such as AI being fundamentally agreed as a way to deal with the change that totally relinquishes on issue based management and thus in turn change most strategic planning, survey methods, social change and any measurement methods. Although, AI Model will be used, this mobile application can be used by the society (users) and for the Sabah Health Ministry. Thus, the mobile application is used to gain awareness towards the society and provides features whereby the user will be able to know whether or not he/she is infected by dengue virus by just using one of the useful feature-symptoms checkers. Furthermore, the mobile application could notify the user to immediately see a physician if the user is considered having a severe infection. This approach showed that smartphones can be used to its fullest extent in which it may not only provides information and knowledge about the diseases or illness caused by dengue, but may also save countless lives one day.

Organization's Evaluation System is an organization that deals with real system that uses AI as their philosophical methodology. This evaluation system was designed mostly using AI and the reason why it was used is because it has the most cost effective and potential approach for the organization to develop the evaluation system. According to [9], the evaluation system was developed using the four phases of AI that includes Inquire, Imagine, Innovate and finally Implement. In the first phase, "Inquire", the participants will be asked to answer a question and to interview by each other in pairs. In the second phase, "Imagine", the participants will then be asked to write on what is the system should be according to their visions. In the third phase, "Innovate", the participants are then asked to individually develop a provocative propositions statement that reflects about the organization. In the final phase, which is "Implement", where the participants will then be asked to develop two or three recommendations for what would be needed to happen and what are needed to make the provocative propositions from the previous third phase to become a reality. Besides, this research is improvising the 5D Appreciative Inquiry Model to be used and it will be explained in more detail on the next section.

This research is using the 5D AI Model by Cooperrider [7], in order to achieve the objective of this research and as well as to guide in the development of the mobile application itself. This model was originally called the 5D AI Model with 5 phases, namely; Define, Discovery, Dream, Design and Deploy. However, it was modified in order to suit the research in order to maximize the results. The phase and newly adopted process diagram are shown on Fig. 1.

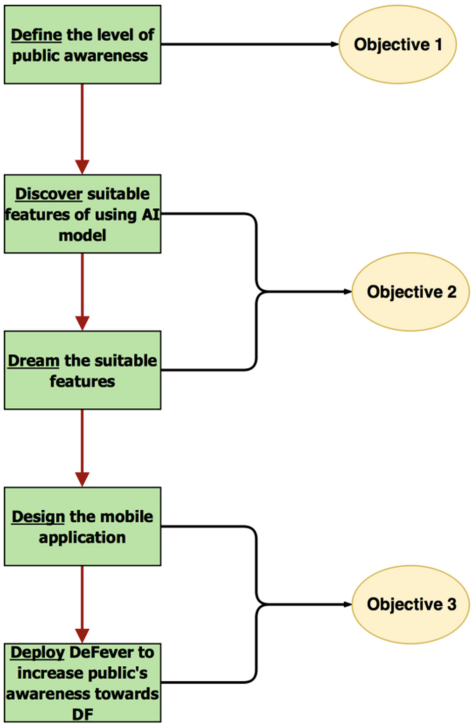


Fig. 1. AI 5D model

During the first phase of the AI process, the first thing that needs to be done is to identify and know what is the required inquiries or questions that needs to be solved in the research. The main task of this phase is mostly to define the questions that need to be used in the research. Therefore, it is an obligation to know how and what makes the public lack in terms of awareness towards DF. Since this issue comes from the public, then the next step in this phase is to conduct a research survey in a form of questionnaire among the public in three geographical areas; Urban, Sub-Urban, and Rural areas. The participants are asked to fill in these questionnaire which consist of 3 factors; knowledge, attitude and practice (KAP). By using a cluster sample, the data is collected from a total of 60 participants randomly from 3 areas (each area is having 20 participants consequently). The data then analyzed and studied in order to investigate whether the awareness level is sufficient or not. The reason why there are only 20 sample size needed to be taken in each geographical area is because that there are limitation of getting the

sample size of above 20 due to bias elimination between the three geographical areas. The publics in urban and sub-urban areas are more educated and have more knowledge on the issue about DF compared to individuals who stayed in rural places due to the low education level [10]. Therefore, it was recommended to have a balance of 20 samples for each area. In the second phase, in discovering the suitable features of using AI model, the inquiry and also the results gained from the previous phase was referred and based on that result the next step was making an analysis of discovering and finding what is best for the next step to be done. Taking the result from the previous phase and filtering it to find the best strength of it in a form of important key ideas to be brought into the next phase. The data from the collected survey questionnaire was referred to and from there the data was being analyzed in a form of statistics.

In the third phase, Dream, the main task is to identify and propose suitable features, the results and important information gain from the previous phase were then be used in this phase to think and visualize about the best solution before moving further into the research. If the result shows that the public has lack of awareness in DF risks, then in the dream phase, is to identify the suitable features to be used using the AI method. In the fourth phase, which is design the mobile application, the design of the DeFever interfaces were developed based on all the important results, ideas and key details gain from the dream phase including the first and second phase.

4 Experimental Results

This section explains and shows the results collected based on the survey questionnaire conducted with 60 sample sizes in 3 areas namely, Urban, Sub-Urban and Rural areas, with 20 samples for each area respectively. The data collected was based on the 3 factors of the public's knowledge, attitude and practice (KAP) towards DF. The selected districts are according to the geographical area in Kota Kinabalu which are Urban (Universiti Malaysia Sabah), Sub Urban (Penampang, Pekan, Donggongon), and Rural (Kampung Sugud, Jalan Tuaran, Penampang).

The data collected was categorized according to their areas; Urban, Sub-Urban and Rural, each area was analyzed and summarized respectively as shown in Fig. 2. It was clearly shown and stated that urban area has high awareness level, followed by sub-urban and the lowest awareness which was in rural area. These results show that the level of awareness in urban and sub-urban in Kota Kinabalu due to the fact that more educated individuals stayed in urban and sub-urban areas, compared to rural areas. Therefore, urban areas have high awareness level, sub-urban areas have a medium awareness level and rural areas show a low or lack of awareness level.

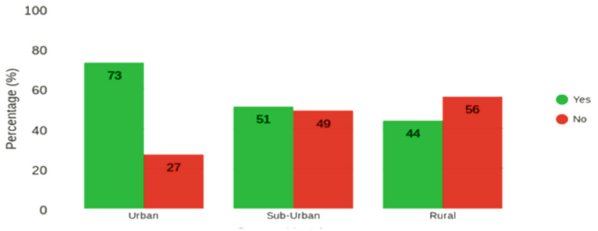


Fig. 2. Level of awareness in Kota Kinabalu

Figures 3, 4 and 5 show a clear statistic of the KAP towards DF. It has been shown that the knowledge factor in urban and sub-urban are much higher compared to rural area. However, the attitude towards DF shows the same high results in all three geographical area including rural areas. The practice towards the preventive measures of DF in urban shows a high practice, on the other side however, shows a balance practice results in both sub-urban and urban areas.

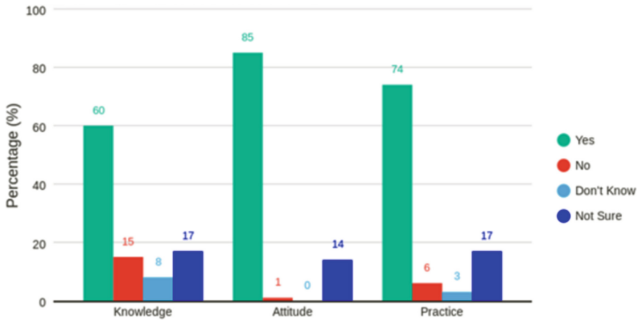


Fig. 3. The urban area according to KAP towards DF

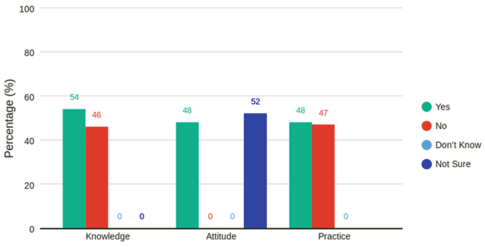


Fig. 4. The sub-urban area according to KAP towards DF

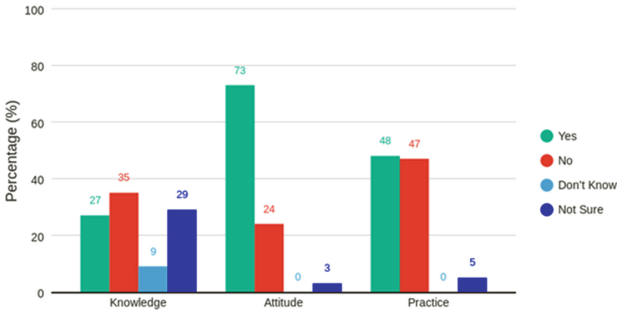


Fig. 5. The rural area according to KAP towards DF

5 Features of DeFever

There are 4 features available for users including information about the app and exit as shown in the main menu in Fig. 6. The 4 main features are dengue information, symptoms checker, dengue hotspot, and report dengue cases (Fig. 7). Admin will be able to gain access to all of this features in the main menu, which are similar to user's features but admin can view the reported cases that the user reported (Fig. 8). The view reported dengue case features would enable the admin to view, update and delete any cases.

Fig. 6. Main menu for users and admin

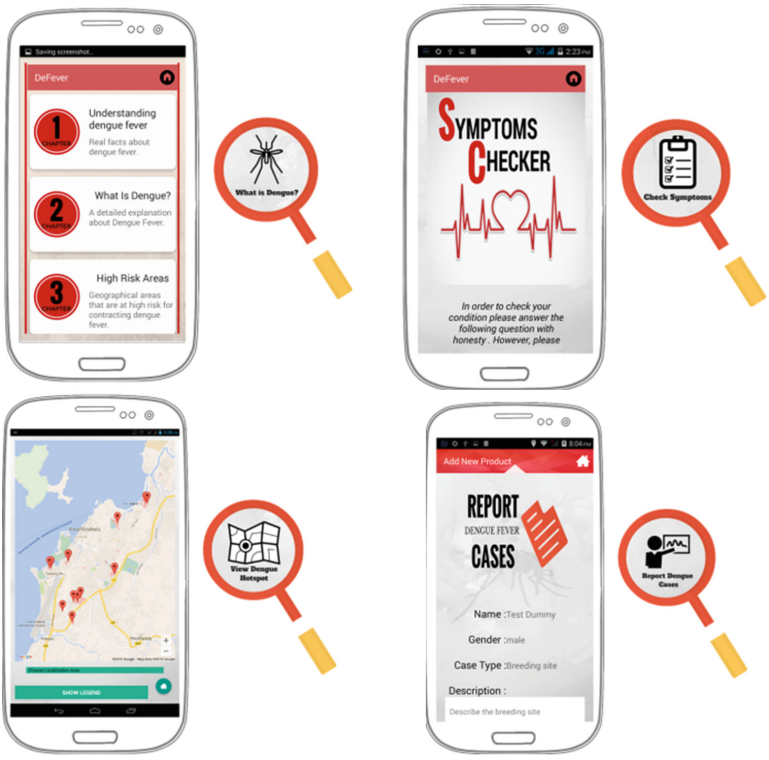


Fig. 7. Screenshots of DeFever's features

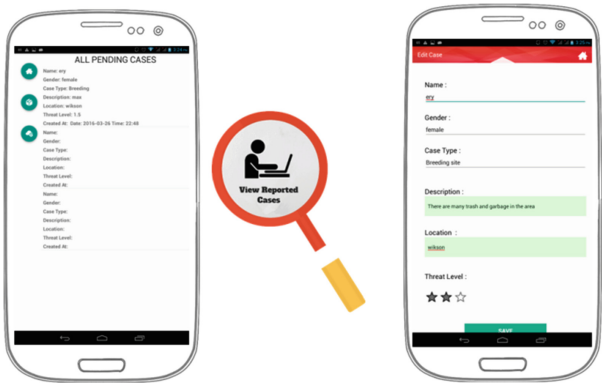


Fig. 8. Screenshots of DeFever's features for admin

Dengue information feature displays the overview of the chapter available containing information and facts about dengue fever. Once user choose a desired chapter a new screen is displayed containing the chapter's content. Symptoms checker feature is available for both user and admin privilege only. This feature enables user or admin to check

their current symptoms whether they are suspected with dengue fever or not. There are three suspected DF according to its respective symptoms, namely Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome. A series of symptoms will be shown, where the user or admin will be ask to check the checkbox if they have the selected symptoms.

Dengue hotspot feature is available to all users, which is enabling user to view dengue hotspot and hot areas. This feature integrates the use of Google map API and shows the dengue cases in hotspot areas even without Internet connectivity or GPS. This feature enables user to send in a report case of dengue breeding site by filling in the required information namely, description, location and threat level. The name and gender are based on the user's account, which will be automatically filled. The reported cases are then stored onto the database only when the user have Internet connectivity.

View reported cases feature is only available for admin. Whereas admin can view all the reported dengue cases send by users. Admin can update status, view and delete the cases once it is verified.

Figure 9 shows main menu for guest, which is unregistered user (public). Dengue information and dengue hotspot features are only shown and available to guest that does not have an account. However, guest can tap the button at the top right corner to go back to the login screen and register as a user.

Fig. 9. Guest main menu

6 Conclusion and Future Works

The risk of Dengue fever (DF) is increasing, therefore, the need of the public or the society to take action against this issue is highly necessary. This research was to investigate on the level of public awareness and to develop an informative mobile application namely, DeFever that hopes to solve the awareness issue. As a conclusion, developed DeFever may help to increase the level of public awareness towards DF. Therefore, regardless of the geographical area, it is an obligation to know more about the issues of DF in hand and what it has caused in this present world. Future works can involve the implementation of the DeFever app for larger scale not only in Kota Kinabalu but the

whole country of Malaysia and further research on usability testing to ensure the app works on multiple platforms for hundreds of users or even more. Feedback from the relevant authorities and users through the questionnaires for usability testing could be improved further.

References

1. Kuno, G.: Emergence of the severe syndrome and mortality associated with Dengue and Dengue-like illness: historical records (1890 to 1950) and their compatibility with current hypotheses on the shift of disease manifestation, pp. 186–201 (2009)
2. Lisa, B.: Dengue Fever (2015). http://www.webmd.com/a-to-z-guides/dengue-fever-reference
3. Robert, H.: Dengue in Asia (2015). http://outbreaknewstoday.com/dengue-in-asia-thailand-malaysia-vietnam-singapore-and-taiwan-67921/
4. Chok, S.: Zeroing in on five dengue hotspots. Borneo Post Online (2015). http://www.theborneopost.com/2015/03/20/zeroing-in-on-five-dengue-hotspots/
5. Edward, Y.: Support from community vital to prevent spread of dengue. Daily Express (2014). http://www.dailyexpress.com.my/news.cfm?NewsID=88409
6. Satosh, V., May, O.L., Theng, Y.L., Schubert, S.B., Gentatsu, L.: A Study of Knowledge, Attitudes, Practices and Health Information-Seeking Related to Malaria-Dengue Prevention in India, Singapore (2012). http://www.ntu.edu.sg/home/sfoo/publications/2012/2012-amhcr.pdf
7. Cooperrider, D.L., Whitney, D., Stavros, J.M.: Appreciative Inquiry Handbook. Lakeshore Publishers, Bedford Heights (2003)
8. Martinetz, C.F.: Appreciative inquiry as an organizational development tool. Perf. Improv. **41**, 34–39 (2002). https://doi.org/10.1002/pfi.4140410809
9. Preskill, H.: Building an Organization's Evaluation System: A Case Example of Using Appreciative Inquiry. Claremont Graduate University, California (2007). http://files.eric.ed.gov/fulltext/ED504356.pdf
10. McCracken, J.D., Barcinas, J.D.T.: Differences between rural and urban schools, student characteristics, and student aspirations in Ohio. J. Res. Rural Educ. **7**(2), 29–40 (1991)

A Model for Predicting and Determining the Best-Fit Programmers Using Prognostic Attributes

Sorada Prathan(✉) and Siew Hock Ow

Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
ida1111@siswa.um.edu.my, show@um.edu.my

Abstract. Different approaches have been used to determine, measure, and predict the performance of programmers to fit positions in the software team. In this study, we use a data mining approach to identify best-fit programmer to be appointed. A questionnaire was used to collect data from 470 programmers from different software companies. A best-fit programmer prediction model was developed to evaluate 10 performance attributes. This model incorporates the Bayes' Theorem and uses Artificial Neural Network (ANN) with Multilayer Perceptron (MLP) algorithm to predict the most suitable candidates for appointment as programmers. The results have shown that programmers who scored high in all or most of the attributes match the predicted values of the Bayes' probability values of the dataset. We conclude that the combination of the theorem and algorithm has proven to be effective in determining the best-fit programmers for appointment using the applied attributes.

Keywords: Data mining · Performance prediction · Programmers · ANN Multilayer Perceptron

1 Introduction

In complex software projects, programmers are assigned respective tasks and are evaluated based on their individual's achievements [1]. Within an organization, programmers are involved in multiple programming tasks, and their performance need to be evaluated from the viewpoint of cost effectiveness to the organization. The evaluation will help an organization to identify the best programmers among the existing staff as well as those to be recruited [1]. It also greatly facilitates the human resource department of any organization to make informed managerial decisions regarding their staff [2]. Better performing programmers will contribute positively to the organization through better software that also meet the set deadline.

Software firms seek to maintain and develop quality software products that also ensure high maintainability and reliability. Hence, different methodologies and techniques are used in project development to ensure successful product delivery [3]. In a study, Lehtinen et al. [4] predicted that most software failures are caused by the lack of focus of the people involved in the development of the software.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 294–301, 2018.
https://doi.org/10.1007/978-981-10-8276-4_28

In order to avoid such failures, it is important to choose good and competent programmers for the job. The recruitment or selection of programmers for a project team is one of the important processes in software project management as it will have effect on the human capital quality of a firm [5]. Hence, it is vital that software companies recruit appropriate and skilled talents. IT firms always face challenges when hiring new people for ongoing projects, due to the lack of an appropriate selection framework [6, 7]. The programming tasks are complex and determining the contribution from each programmer has always been a major challenge in software development [8]. The process of selecting and appointing the best programmers has always posed a challenge to researchers and academia [9, 10]. The rest of the paper is organized as follows: Sect. 2 discusses related works on selection of employee's based on performance prediction; Sect. 3 explains how the selection process is conducted; Sect. 4 discusses the results, and Sect. 5 concludes the research.

2 Related Work

Many studies have been conducted on the use of different approaches and techniques for performance appraisal. In all the studies, the main objective had always been in identifying the best way of selecting the best employee for a particular job. Some studies [9, 11, 12] have identified good and reliable methods for the selection of programmers to fit appropriate positions in the organization.

In their research, Satyanarayana and Nuckowski [5] employed Data Mining with Ensemble Classifiers to improve prediction of academic performance of students by eliminating unwanted instances, thereby, improving accuracy of the prediction. They identified the association rules that influence a student's result using a combination of rule-based techniques. The result obtained by using the technique is compared with that of the single model-based technique. The outcome reveals that the ensemble models not only produce better predictive accuracy of student performance, but also provide more appropriate rules for understanding the factors that influence the results.

Jantan [3] identified the pattern of employee performance through a decision tree classifier – tree C4.5 classification algorithm. The authors mined the hidden and valuable knowledge in the correlated databases and summed up the result in the form of a decision tree. The generated rules were then evaluated with an unseen data to estimate the accuracy of the prediction result. The techniques were chosen because they use the common techniques for classification and prediction in data mining. The decision tree is a divide-and–conquer approach from a set of independent instances. The study outcome proves that the C4.5 classifier has great potential for making performance prediction. It also clearly shows that knowledge is essential for assessing the performance of employees.

In another study on predicting employee performance, Al-Radaideh and Al Nagi [2] used data mining techniques to build a classification model. They adopted the CRISP-DM data mining technique to model the classification using a decision tree. To validate their model, several experiments were conducted using real data collected from many companies. The final outcome reveals that job title is the strongest factor, followed by the type of university an employee obtained his/her qualification.

In their research, Pal and Pal [8] evaluated teacher's performance using single data mining techniques. The authors collected information from teachers and students and developed rules using the Naïve Bayes Classifier technique. Their technique was compared with other related algorithms – ID3, CART, and LAD tree. The Naïve Bayes classifier was found to have the lowest average error as compared to the other two algorithms. The authors suggested that the Naïve Bayes algorithm is suitable for using data mining approach to evaluate performance.

In their studies on improving the selection process in the technology industries, Chien and Chen [13] developed an effective personnel selection method for finding the most suitable talents for organizations. They used data mining decision tree analysis to discover latent knowledge and extract the selection rules. They conducted an empirical study in a semiconductor company for different job functions. The final results showed the practical viability of their approach and identified some strategies in human resource management.

Ola and Sellappan [9, 10] modeled a technique for evaluating the performance of instructors. The authors established a few standards and criteria as a framework for evaluating an instructor's performance. They stated that if fully implemented, the framework will provide a basis for instructors' performance improvement that will, in turn, also optimize students' academic outcomes and improve the standard of education. Previous studies did not use prognostic attributes to identify and analyse the best-fit attributes that a programmer possesses. In this study, we incorporated Bayes' Theorem into Artificial Neural Network (ANN) to identify the prognostic attributes. A data mining technique was used to identify and analyze the attributes that the best-fit programmer should possess.

3 Performance Prediction Model

The main goal of this research is to establish a best-fit model to predict the performance of programmers. An ensemble data mining model was constructed using two established tools – Bayes' Theorem and Multilayer Perceptron (MLP) algorithm of the Artificial Neural Network (ANN), as shown in Fig. 1, to achieve the objective. A questionnaire was administered to 470 programmers, and the information collected was analyzed based on their work-pertinent attributes – attitude, creativity and innovation, performance, interpersonal skills, problem-solving skills, and teambuilding skills. Datasets formed from the returned questionnaires were fed into the model for prediction. The Bayes' Theorem was used to find the prognostic attributes that rank the relevant attributes that are most important pertaining to the work performance of programmers. The MLP algorithm was then applied to the result to refine the weightage of each attribute in order to shortlist the candidates who are suitable for the job. The recruitment process is heavily tied to a hypothesis in order to realize valid conclusions on independent and dependent variables based on their respective relationships. A manipulation process is integrated into the normal recruitment process to identify the most suitable programmers for the organization. This manipulation process affects the existing recruitment process by enforcing a new attribute behaviour. The

entire process hybridized the existing recruitment process to control the entire process without affecting the existing procedures.

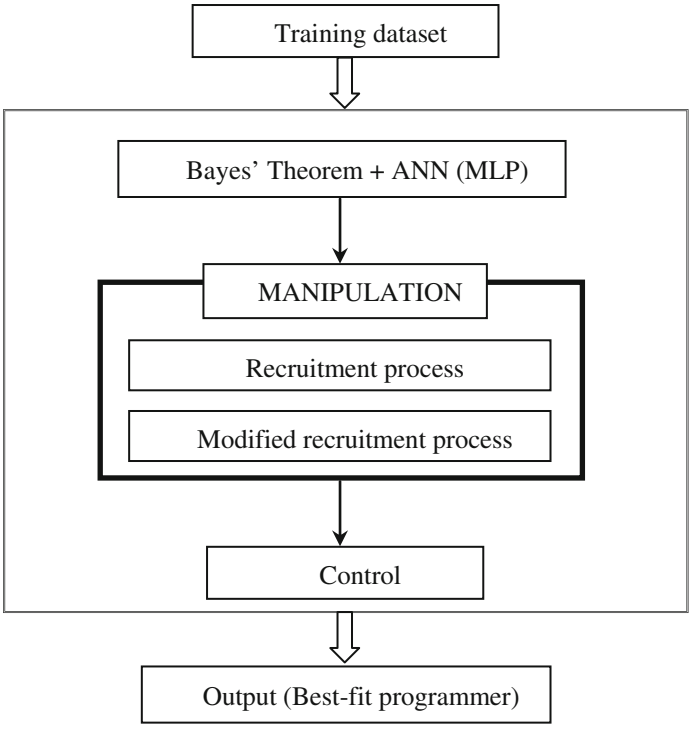


Fig. 1. Framework for determining the best-fit programmer

3.1 Multilayer Perception

MLP is the network containing a collection of sensory elements. These elements form the required input, hidden, and output layers. The algorithm performs the classification approximately with the vector attribute values.

Manipulation: The proposed manipulation process is integrated into the normal recruitment process where it identifies the most suitable programmers for the organization.

Control: The existing recruitment process remains stable, but changes are made to it by the inclusion of the proposed manipulation process. The proposed process is hybridized with the existing recruitment process to control the entire process without affecting the existing procedure.

3.2 Artificial Neural Network

ANN is an information processing system similar to the biological nervous system for information processing in the brain. ANN is used to solve a specific function in a network, combined with a large number of interconnected neurons. ANN is implemented via the learning process, with the respective weight of each input processed in the training algorithm, such as back-propagation. The weights are iteratively updated in the training to determine the most appropriate value in a specific function.

$$A(x) = V\left[\sum_j W_j C_j(x)\right] \tag{1}$$

where A(x) → Neuron Network function, $C_j(x)$ → (C_1, C_2,…..C_m) (Composition of Functions), W_j → Weight of Neurons, V → Activation Function.

4 Results and Discussion

Bayes' Theorem was applied on the dataset to find the programmers' attributes which are associated with high work performance. The probability applied is calculated as:

$$P(X|Ci) \tag{2}$$

where X refers to the scored attribute value, and Ci refers to the class i, where i = 1, 2 and 3. C1, C2 and C3 refer to Good, Average, and Poor, respectively. The result presented in Table 1 shows the attributes considered for performance evaluation. The probability is calculated based on the dataset, and the prognostic attributes were obtained after applying the Bayes' Theorem. However, programmers with attributes such as – "Writes Functionally Correct Code", "Writes Aesthetically Pleasing Code", "Performs Satisfactory Unit Testing", "Documents Code W", "Asks Questions When Needed", "Communication Skills" and "Corporate Responsibility" – are considered good and impactful programmers who are performing well in their job.

The Bayes' Theorem was applied to the dataset consisting of the 10 attributes, and the outcome shown in Table 1, indicates the acceptance of the performance indicators. Among the 10 attributes used, it is clear that if the candidate "Writes Functionally Correct Code" is rated Good, there is a high probability that the candidate would perform well in the job. Any attribute with a score greater than 0.5 also indicates a high performing or good programmer.

Justification of the score

In this study, the Bayes' Theorem was used to the dataset consisting of the 10 attributes. There are 7 attributes set out as the evaluation criteria which indicates high probability of those candidates that perform credibly well in the job. Furthermore, based on these attributes that we considered "Writes Functionally Correct Code, Writes Aesthetically Pleasing Code, Performs Satisfactory Unit Testing, Documents Code, Asks Questions When Needed, Communication Skills, Corporate Responsibility" and from the results obtained from the study, it shows that those candidates that score above 0.5 are high

performing or good programmers as revealed in Fig. 2. Also, we argued that if a programmer score greater than 0.5, he/she is considered "Good and impactful" programmer who is performing well in his/her jobs and could be best fit candidate for a company. This is shown in the result of our dataset (Table 1 and Fig. 2). We applied ANN with MLP on the result shown in Table 1 to predict the best-fit candidate. The result was applied to the attributes to identify the qualified programmers. Table 2 shows the outcome.

Table 1. Attributes considered for performance evaluation

Attribute	Attribute values	Probability $P(X\|Ci)$		
		Good	Average	Poor
Attribute 1 (Writes Functionally Correct Code)	Good	**0.77**	0.483	0.261
	Average	0.197	0.394	0.348
	Poor	0.033	0.123	0.391
Attribute 2 (Writes Aesthetically Pleasing Code)	Good	**0.64**	0.256	0.043
	Average	0.279	0.54	0.402
	Poor	0.082	0.205	0.554
Attribute 3 (Performs Satisfactory Unit Testing)	Good	**0.59**	0.306	0.228
	Average	0.344	0.565	0.598
	Poor	0.066	0.13	0.174
Attribute 4 (Documents Code)	Good	**0.738**	0.442	0.141
	Average	0.246	0.505	0.467
	Poor	0.016	0.054	0.391
Attribute 5 (Asks Questions When Needed)	Good	**0.738**	0.252	0.217
	Average	0.23	0.435	0.457
	Poor	0.033	0.312	0.326
Attribute 6 (Communication Skills)	Good	**0.705**	0.464	0.25
	Average	0.279	0.284	0.315
	Poor	0.016	0.252	0.435
Attribute 7 (Punctuality)	Good	0.23	0.284	0.24
	Average	0.393	0.363	0.402
	Poor	0.377	0.353	0.761
Attribute 8 (Corporate Responsibility)	Good	**0.525**	0.429	0.315
	Average	0.344	0.278	0.315
	Poor	0.131	0.293	0.37
Attribute 9 (Project Innovation)	Good	0.443	0.199	0.065
	Average	0.41	0.344	0.272
	Poor	0.148	0.457	0.663
Attribute 10 (Leadership)	Good	0.361	0.227	0.109
	Average	0.443	0.309	0.304
	Poor	0.197	0.464	0.587

	A	B	C	D	E	F	G	H	I	J	K	L
1	Writes Functionally Correct Code	Writes aesthetically Pleasng Code	Perfoms satisfactory unit testng	Documents Code Well	Asks questions when needed	Communication Skils	Punctualty	Corporate Responsbilty	Project Innovation	Leadership	Weightage Earned	
2	14	13	14	14	13	4	4	4	4	4	88	
3	14	14	14	13	14	5	4	4	4	3	89	

Fig. 2. Best scores for programmers fitting from the organization.

Table 2. MLP algorithm applied to attributes of the programmers' dataset.

Attribute	Performance		
	Good	Average	Poor
1	0.7705	NaN	NaN
2	0.6393	0.5394	0.5343
3	0.5902	0.5647	0.5978
4	0.7377	0.5047	NaN
5	0.7377	NaN	NaN
6	0.7049	NaN	NaN
7	NaN	NaN	NaN
8	0.5246	NaN	NaN
9	NaN	NaN	NaN
10	NaN	NaN	0.663

In evaluating the attributes using the model, the results show that two programmers also obtained high rating from the annual performance appraisal confirming them suitable for the organization. The results shown in Fig. 2 indicate that candidates 2 and 3 are rated highly with scores of 88 and 89, respectively. This indicates that they scored highly for all the attributes and therefore considered as the best-fit programmers.

Good programmers are very much sought after, but they are difficult to be assessed [14]. In this study, applying both the Bayes' Theorem and MLP algorithm to the attributes helps in determining the best-fit programmers for an organization. High scores obtained for an attribute always indicate best-fit, in other words, the attributes effectively reveal the best candidate to be selected as programmers. If a candidate's attribute is ranked lowly and/or if the Bayes' Theorem probability is below 0.5, it is more likely that the candidate will perform poorly as a programmer. By extension, as more attributes of a selected programmer are evaluated using the algorithm, the more of his/her potential can be determined.

5 Conclusion

This study models a technique by using Bayes' Theorem, and ANN with MLP algorithm to identify potentially good programmers. Some attributes evaluated using the model has helped in identifying the right or best-fit programmers. When the proposed hybrid model was used for evaluation, the results show a probability above 0.5 which indicates best-fit programmers for the organization. The model will help the in determining best programmers. The model was applied on a dataset of 470 programmers but this can also

be extended to larger number of programmers, without compromising its accuracy. Employing the best-fit programmers will enhance their productivity and consequently be beneficial for the organization.

Acknowledgment. This research was funded by the University of Malaya under the Postgraduate Research Grant (PPP), Account Number: PO032-2015A.

References

1. Ajay Prakash, B.V., Ashoka, D.V., Manjunah Aradhya, V.N.: Application of data mining techniques for software reuse. Procedia Technol. **4**, 384–389 (2012)
2. Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employees performance. Int. J. Adv. Comput. Sci. Appl. **3**(2), 144–151 (2012)
3. Jantan, H., et al.: Human talent prediction in HRM using C4.5 classification algorithm. Int. J. Comput. Sci. Eng. **2**(8), 2526–2534 (2010)
4. Lehtinen, T.O., Mäntylä, M.V., Vanhanen, J., Itkonen, J., Lassenius, C.: Perceived causes of software project failures–an analysis of their relationships. Inf. Softw. Technol. **56**(6), 623–643 (2014)
5. Satyanarayana, A., Nuckowski, M.: Data mining using ensemble classifiers for improved prediction of student academic performance. In: American Society for Engineering Education (ASEE) Mid-Atlantic Section Spring 2016 Conference, pp. 17–24. City University of New York Academic Works, New York (2017)
6. Agrawal, N.M., Khatri, N., Srinivasan, R.: Managing growth: human resource management challenges facing the Indian software industry. J. World Bus. **47**, 159–166 (2012)
7. Patih, R.S., Patil, V., Waje, P.: Human resource challenges & practices in IT industry. In: Proceedings of the 5th National Conference, INDIACom-2011, New Delhi, pp. 10–11 (2011)
8. Pal, A.K., Pal, S.: Evaluation of teacher's performance: a data mining approach. Int. J. Comput. Sci. Mob. Comput. **2**(12), 359–369 (2013)
9. Ola, A.F., Sellapan, P.: A data mining model for evaluation of instructors' performance in higher institutions of learning using machine learning algorithms. Int. J. Concept. Comput. Inf. Technol. **1**(2), 17–22 (2013)
10. Ola, A.F., Sellapan, P.: A framework of an improved model for evaluation of instructors' performance in higher institutions of learning. IOSR J. Res. Method Educ. **3**(2), 64–69 (2013)
11. Marimuthu, M., Arokiasamy, L., Ismail, M.: Human capital development and its impact on firm performance: evidence from developmental economics. J. Int. Soc. Res. **2**(8), 265–272 (2009)
12. Tourangeau, R., Smith, T.W.: Asking sensitive questions the impact of data collection mode, question format, and question context. Public Opin. Q. **60**(2), 275–304 (1996)
13. Chien, C.F., Chen, L.F.: Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. Expert Syst. Appl. **34**(1), 280–290 (2008)
14. Cooper, D., Robertson, I.T., Tinline, G.: Recruitment and Selection: A Framework for Success, 1st edn. Cengage Learning EMEA, London (2003)

Design and Development of Novel Android 3D 3rd Person Shooting Game

Kim On Chin[1(✉)], Syukri Majdi Hamdan[1], and Tan Tse Guan[2]

[1] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
kimonchin@ums.edu.my, stamif@gmail.com

[2] Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Kota Bharu, Malaysia
tan.tg@umk.edu.my

Abstract. In recent years, smartphones have grown exponentially in popularity, taking over the world by storm. While the mobile gaming industry has grown large due to the development, there is a deluge of mediocre games in the Android market which replicate the same formulas of popular games. Thus, it motivates us to try at a new game development and design. The Android game is developed using the jMonkeyEngine game engine for Android 2.2 versions and above. It utilizes simple 3D graphics, and use the accelerometer and touch screen for input. It is single one on one battle against an AI enemy, which uses a simple finite state machine and viewed using a third-person camera that stays behind the player's character and follows the enemy's movements. We believes that the virtual joystick, a popular method of control in current games is inadequate for use as a control method to control the player character in a 3D third-person game. As such the accelerometer is being used as the game's method of control, which is not only more sensitive and removes unneeded elements on the screen, but also to increase the game's uniqueness. The project is developed using a test-driven process model. The system could be divided into three sections. A section of the code handles the generation of the character models and others, while an input handling system take care of mapping the player's inputs to the correct responses. A 3D world is created, its members positioned correctly to form the ground, walls, and the player and enemy characters. Next, physics was added to the 3D objects so that they may interact with each other. Next came the implementation of the accelerometer to move the character around, and the touch screen to produce actions. Finally, a state machine is built for the game AI and suitable actions for each state is coded in. The game is tested via test runs as each functionality was implemented.

Keywords: Gaming AI · Android game development · 3rd person shooting game Artificial Intelligence

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 302–313, 2018.
https://doi.org/10.1007/978-981-10-8276-4_29

1 Introduction

In recent years, smartphones have become a staple within society, and while many adore the features in these tiny computers for having made their life so simple, smartphones are also viable as portable entertainment and social systems [1]. Current devices are able to snap high-definition pictures that can rival professional photos, store hundreds of music that would have needed thousands of CDs, and play movies that last for few hours, depends on the battery capacity [2] and naturally, when one speaks of entertainment, games always come to mind.

Mobile gaming has become quite a large industry due to the abundance of smartphones [3]. There are several outlets where developers can publish their smartphone applications depending on what are device they are developing for, and for smartphones that contain the Android platform, there is Google's Android Market [4]. While certain restrictions are imposed on developers by other application stores such as Apple's iTunes Store [5], the Android Market has no such limitations. This allows for a low level-of-entry, and lets developers ease into the industry, encouraging a great deal of creativity to developers and increasing the variety of the games being published in the Android Market.

While the mobile gaming industry has grown large due to the development, there is a deluge of mediocre games in the Android market which replicate the same formulas of popular games [6]. As an example, while a smartphone has many features besides the touch-screen, many developers use a virtual joypad which uses the familiar directional and action buttons layout, even though it lacks the same feel that one has as they grip an actual controller [7, 8].

It is common now for console games to have high-definition graphics and 3D models, very few games on the Android platform have high-end 3D graphics [9]. This is partly due to the range of machines that use the Android platform as not all are capable of running a 3D game smoothly. Thus, the first and biggest challenge would be the 3D game development itself. Next, there is a noticeable lack of games that use the smartphone's accelerometer, making it a suitable feature to have in a unique game. Thus the next challenge would be to implement the accelerometer to control characters inside the game. Finally, there must be an enemy character for the player to battle with, which would be controlled using Artificial Intelligence (AI). There are numerous AI methods could be implemented in gaming contexts such as Differential Evolution, Genetic Programming, Genetic Algorithm, as named a few [10–14], however, AI in simple games commonly use state machines for decision-making [15, 16]. Thus, the final challenge is to implement a finite state machine for the game AI.

Upon seeing the current selection of games in the Android Market, it was decided that this project would aim to develop a game that implements various techniques that are uncommonly used, in order to create a unique game that feels entirely different than most current games available in the Android Market. Hence, this project has the following goals:

1. To create a 3D third-person view game on the Android platform.
2. To successfully implement the accelerometer as a control device.
3. To successfully implement a finite state machine for the game AI.

The remaining parts of this paper are organized as follows. In Sect. 2, a brief discussion of mobile games is included. This includes the discussions of common game genre, genre choice, current controller used in most of the 3D third person mobile games, and benefits of using accelerometer. In Sect. 3, the proposed method is presented. This section also briefly described the jMonkeyEngine3 used, design process model, and the game AI design. The research result is discussed in Sect. 4, and finally the future work is presented in Sect. 5.

2 Mobile Games

This section covers a brief discussion of common game genre, genre choice involved in this project, current controller used in most of the 3D third person mobile games, and benefits of using accelerometer.

2.1 Common Game Genre

First-person shooters (FPS), Role-playing games (RPG), Fighting games, Platform games (Side-scrollers), Racing games, Strategy games and Rhythm games are the most common genres available in the mobile gaming industry [4, 5]. There are others, such as simulation type games (The Sims) and puzzle games which are increasing in popularity (Cut the Rope); then there exists combinations of genres, such as first-person RPG (Skyrim) and flight-simulation games (Ace Combat series) [4, 5]. These seven genres however are usually the most easily identifiable. The fact that the game industry cannot be summed up by only seven genres serves as a testament to how large the gaming industry is now when compared to the past.

2.2 Genre Choice

This project will be of the third-person shooter genre. This decision was made due to the lack of popular games in that genre on the Android platform [4]. Currently, the Android Market is dominated by puzzle and strategy games [4]. This is understandable as both game types are usually short and easy to play during breaks. Another plausible reason is that 3D games are difficult to develop for smartphones, as most devices are ill-equipped to handle 3D graphics. There are also other reasons why such games are favored when compared to third-person view games, which will be explained in the next section. It is due to these reasons that it is very hard to find a third-person view game whose quality is comparable to those of puzzle games. This design choice was made since there is a lack of 3D third-person shooter games, games in that genre would be more unique.

2.3 Current Controls for Third-Person Games

Third-person games are called as such because the game is viewed from behind the player character. This allows players to admire the world without the motion-sickness-inducing movement of first-person games. Players are also able to view their character in full; a sense of fulfillment can be achieved when they see their character decked out in the strongest and coolest looking armor.

The current traditional control method for third-person games are virtual joysticks on the screen [4]. There are several issues with this control method. On tab devices, the large screen allows the joystick to be placed at the devices' corners. On a smartphone however, the joystick's position are often such that the player's fingers will block the screen, limiting their view, and this is especially true for smaller devices. This is quite unavoidable, as joysticks need a large portion of the screen to be sensitive enough to detect the slight movements done to move the character on screen. An example of games that use this control method and how it looks on a smartphone device is presented in Fig. 1 below. The joystick and buttons themselves cover 20% to 40% of the screen, and this is still not counting the player's fingers which will be touching the screen as they play.

Fig. 1. Screenshots of games that utilize traditional control methods

Even with tab devices, joystick controls will never be as sensitive as their console counterparts. It is easy for a player's finger to slip from the joystick, causing players to lose control of their characters, which can be fatal especially when the game they are playing need high precision controls. There are attempts at avoiding such situations such as having the joystick appear wherever the finger is touching, but there are still issues with that as well (wanting a new joystick to appear at a new area on the screen, but the new area is on the old joystick's 'move right' button, thus accidentally moving the player to the right).

Despite the inadequacy of virtual joysticks for playing a third-person games, the control system is still being implemented frequently in such games, alongside virtual buttons for actions. Cumbersome controls are the bane of any game, and this is evident in the general unpopularity of third-person game. It should be realized that virtual joysticks can never offer the feel of pushing the console controller's joysticks or the feedback received from spring-loaded buttons, and find alternatives to movement and actions for touch-screen devices.

2.4 Benefits of Using Accelerometer

Accelerometers are built-in sensors now common in smartphones that can detect the angle at which the phone is tilting. It is still a relatively new feature that few apps manage to implement well. It is quite commonly used with racing games, as the tilting movements allow players to feel as though they are actually steering a vehicle [17].

There are few games other than racing games and self-scrolling/moving games in which the player character's movements can be controlled by the accelerometer. A game that implements the accelerometer well can improve controls significantly, and this will be especially true for third-person games. The sensors allows for a great level of sensitivity in movement, and can surely help in giving precision control to the player.

Since accelerometers are built-in, no additional controllers would be necessary on screen, thus improving visibility as the player no longer needs to press the screen as often as with virtual joystick controls. Furthermore, even in controller-based games on the console, players playing fast-paced action games will often sway their bodies according to their actions on-screen. Due to the tilting motion of accelerometers, plus the added visibility on the player's screens, the game could be more immersive, adding to the enjoyment of the players.

The accelerometer offers great control sensitivity, add to that the benefit of being able to view the phone screen in full and since usage of the accelerometer is still quite novel, accelerometers could be the most suitable control method for third-person games.

3 Methodology

3.1 jMonkeyEngine3'

The project is developed on the Android platform, as it has a low level-of-entry and the Internet provides easy access to learning materials and tutorials [18]. The game will be using a free 3D game engine known as the jMonkeyEngine, as it has relatively good documentation and uses the familiar Java programming language. It also allows for Android development [18]. The jMonkeyEngine utilizes OpenGL 2.0 for graphics, thus limiting the project to Android versions 2.2 and above.

jMonkeyEngine3 (jME3) is a game engine aimed towards 3D game development, programmed entirely in Java and uses the Lightweight Java Game Library (LWJGL) as a renderer. The libraries for this engine is released under the BSD license, thus making the engine widely available for use. Numerous tutorials are also available at the jME site for beginners [18].

The jMonkeyEngine SDK is a preconfigured Software Development Kit (SDK) based on the NetBeans Platform, which combines the jME3 libraries and adding unique functions such as project and file wizards, asset pack management, a scene viewer and others that help make the game development process easier. The SDK was introduced to provide a ready-to-use complete game development environment to developers, and while the SDK is recommended, developers can still use other IDEs for game development, and use the SDK only to manage assets and jME3 binaries.

With regards to the game itself, a single stage of a one-vs-one battle between player and enemy characters is created. The camera is of third-person view, and will follow the enemy's movement throughout a battle. Simple 3D models is used for the stage, player and enemy models. Player input consists of the touch screen, phone buttons, and accelerometer. The AI used a simple finite state machine.

3.2 Process Model

The development of this project uses a test-driven development cycle. The test-driven development starts when the developer designs a test case which the product will ultimately fail to do, either because the test implements a new function not yet coded in, or because of the code needs to be improved. The developer will then add in code until the product can pass the test. The development cycle is short, thus allowing for further improvements with each cycle until the product is complete.

This process model was chosen due to the benefits that it has to us. Due to the inexperience, exhaustive research needs to be done as development continues, and since the test-driven model allows for small and incremental improvements to the project due to its short lifecycle, it lets us test out the techniques that we learned. Designing test-cases also produces a clear short-term goal that we can achieve, which can guide us on what to research next. Tests in test-driven development is usually automated, the tests in this project is done manually, in order to save time and remove the need for writing code for the testing phase.

jMonkeyEngine3 Terminology
In jME3, certain terminologies are used that are necessary to explain the system design, which are scene graph, spatials, nodes, geometry, transformations, meshes, materials and textures, and Nifty GUI:

World Creation
In jMonkeyEngine, scenes can be added automatically after being created in a 3D computer graphics software. This is also true for character models. Game physics can also be handled by the engine automatically. Thus world creation is only a matter of adding scenes and models, and then adding physics into them.

- Add the scene - The scene is created and added onto the scene graph by attaching it to the rootNode. The scene consists of four wall Geometries and one ground Geometry.
- Add player and enemy models - player and enemy models are attached to their respective Nodes before being attached to the rootNode.
- Add the camera - In this project, the camera being used is a type of camera available in jME3 known as the CameraNode. It is attached to a relevant Node and follows the movement of a target. In this case, the camera is attached to the playerNode and looks at the enemy model.
- Add physics - jME3 allows for easy implementation of game physics by using the special classes it has prepared, such as RigidBodyControl for stationary objects, and CharacterControl for moving objects. RigidBodyControl is added to the scene, while

CharacterControl is added to the enemy and character models. Collision detection is handled automatically by the engine.
- Add GUI - A pre-created XML file which contains the configurations of the GUI is loaded onto the phone screen. The GUI displays information such as the HP bar of the player and the enemy.

And with that, the world has been created. Note that while actions such as loading models, adding physics and animations, etc are listed here, these actions are usually placed within the relevant object's classes, such as adding physics to the player model in the player class.

Player Input

The main input methods in the project are the accelerometer and touch screen. Any viable input entered will be processed and the game world is modified accordingly, the process of which is detailed below.

- Listen for input - Once everything has been started up, the game will then begin to listen for input. This is done continuously throughout the game, with the appropriate action done for different inputs.
- Accelerometer input received - The accelerometer handles player model movement. Thus when input is received, modifications will be done to the player model's location within the game world, with the appropriate animations being played. The player model can move forward, backwards, to the left and to the right according to the direction the phone is tilted towards, and will continue to move for as long as the phone is tilted.
- Touch screen input received - Player actions are done via the touch screen. In this project, actions is limited to attacks. The touch screen is divided in two, signifying the left and right hands. Input done within the left side of the screen will trigger actions using the left hand, and the opposite will be true for the right side of the screen. The left hand will hold a sword, and attacks can be triggered by tapping the left side of the screen. By tapping continuously, attacks with the sword can be done repeatedly. The right hand will hold a gun, which can be fired by tapping the right side of the screen. By pressing and holding the screen continuously, the gun can fire at set intervals. A dragging direction along any direction on any part of the screen will make the player jump. Long-pressing on the left side of the screen will turn on the AI and start the game.
- Phone buttons input - Most Android smartphones have several common buttons between them. Proper responses will be done when the buttons are pressed. The Home button exits to the phone's Home screen without closing the game. Ideally, the game is paused when this happens. The Back button is set by default to close the game upon confirmation by the player.

3.3 Game AI Techniques

There are several game AI techniques that developers usually use to use in their games, as these techniques are often well documented and easy to implement into different types

of games. [19] documented several common techniques; some techniques that are related to the project are:

- The A* pathfinding algorithm is used to find the cheapest (or shortest) path to a goal, giving a set of points which can be used by the AI to head for the goal.
- Dead reckoning uses predictions to aim and shoot at a player using variables such as position, velocity, and acceleration.
- Emergent behaviour is not explicitly programmed into the game's AI, but emerges as the result of several more basic behaviours.
- The level-of-detail AI technique is based off of a common 3D graphics optimization technique where further or useable objects use fewer resources than visible ones. Similarly, in the AI version of the concept, complex decision making are only made when the changes caused by the decision can be experienced by the player, using simpler techniques otherwise.

Any Non-Player Character (NPC) at any given time in a game is usually in a certain 'state', whether they are walking, talking, standing still, attacking, etc. They can enter a different state whenever certain conditions are met. A state machine contains all this and also defines the conditions needed to be met to enter the state. The simplest form of state machine implementation is using the machine to keep track of what the AI is currently doing, and changing the states as necessary. A more complex use of state machines is to define the NPC's current 'state of mind' with it, which will then influence what the AI chooses to do. For example, an 'aggressive' state will make the AI choose to attack more.

The game AI in this project will be a simple one, using state machines to control how the AI will behave. The AI is updated according to the current conditions it is in, which is done by running through the functions it has. The following are functions that the game AI function would have:

- Update the AI - This function is to begin a checking of all the conditions that may affect the AI's behaviour.
- Check current state - The current state of the AI is checked and executed accordingly. The AI will stay in a particular state until certain conditions are fulfilled. The conditions are commonly to execute code until a certain amount of time has passed. It will then move to the next state.
- 'Find' state - This state finds and keeps track of the player location in this state. This is done by simply getting the player model's current coordinates and keeping it in a variable, before going to the next state.
- 'Look at the player' state - This state simply turns the enemy model to look at the player. This is done by getting the direction at which the player is looking at and making the enemy look at the opposite direction.
- Attack state - This state executes code that corresponds to the current attack state. The code animates the enemy model to move according to the state. There are three attack states: a thrust attack which moves the enemy directly towards the player, a jump attack where the enemy jumps above the player, and an attack which throws objects at the player. The attack is continued for a certain amount of time before continuing to the next state.

- No action period - After each action, a small waiting period will be waited out before executing another action. This is to ensure that the AI's back-to-back attacks will not overwhelm the player, and to give the player a chance to deal more damage. Some actions may have a shorter or longer waiting period than others, so the player can capitalize the chance.

This cycle repeated for as long as the AI's Health Performance (HP) still remains. Once it reaches zero, the AI is defeated and the player has won. Included in Fig. 2 is a simplified diagram of the state machine design. Attack states include the three different attacks that the AI can do.

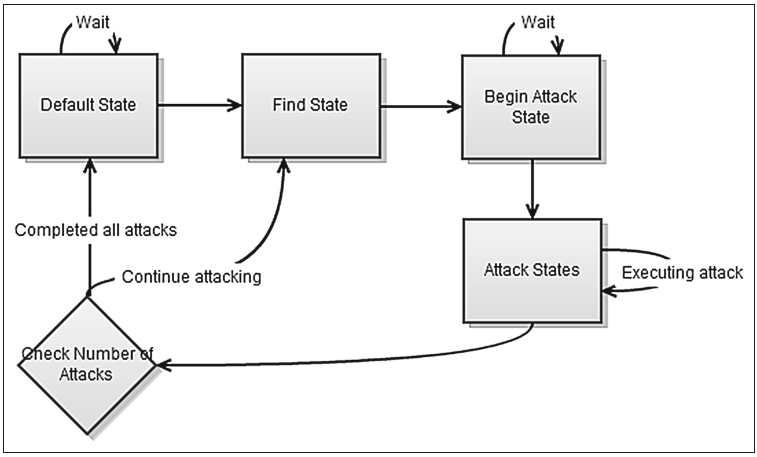


Fig. 2. State machine design

4 Implementation and Testing

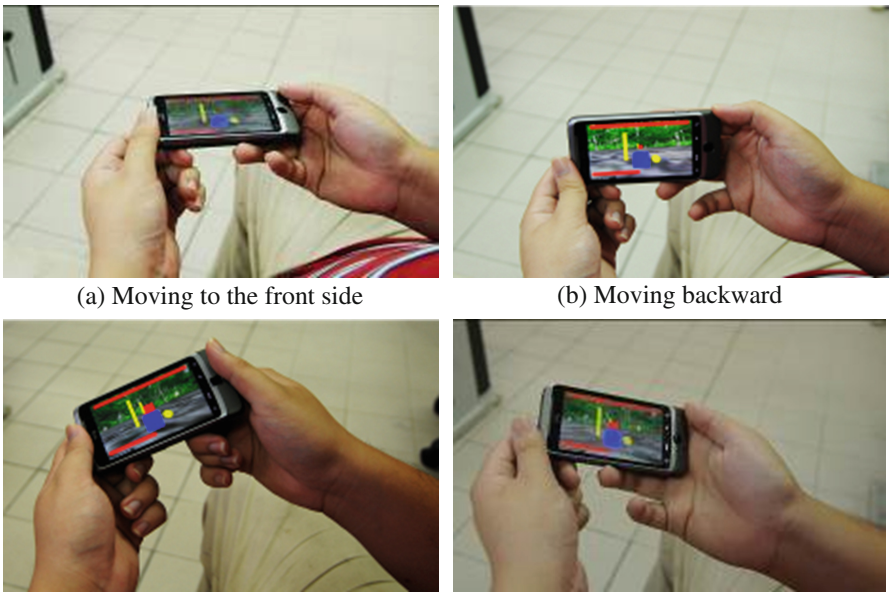
The games from the Monster Hunter and Armored Core series have influenced the game design aspect of this project greatly. From Monster Hunter comes the inclusion of tenacious AI enemies which attack at every chance they get, and get the player to think how to attack instead of rushing in. Armored Core influenced the use of a constant lock-on camera, as well as the use of a single large stage.

The interface was designed to be as minimalistic as possible, having only the player and enemy characters, as well the player's HP bar in view, allowing the screen to be free of unneeded elements. The system itself could be divided into three sections. A section of the code handles the generation of the character models and others, while an input handling system take care of mapping the player's inputs to the correct responses. The game AI uses a state machine to keep track of its actions and influence its decisions. The game loops constantly after initialization, updating the world according to the values that change.

First, a 3D world was created, its members positioned correctly to form the ground, walls, and the player and enemy characters. Next, physics was added to the 3D objects

so that they may interact with each other, and GUI was used to display certain information. Next came the implementation of the accelerometer to move the player around, and the touch screen to produce actions. Finally, a state machine was built for the game AI and suitable actions for each state was coded in.

The game is tested via test runs as each functionality was implemented. Implementation began with the creation of 3D objects, adding the third-person, adding physics and adding the GUI. Next, player input was implemented via accelerometer and touch screen. A game AI using state machines utilizing three attacks was then implemented, before finally adding in damage calculation to complete the game. Upon completion, some volunteers were found to test and play through the game, finding bugs to be dealt with. Figure 3 shows some snapshots during testing phase.



(a) Moving to the front side (b) Moving backward

(c) Moving to the left side (d) Moving to the right side

Fig. 3. Testing by volunteer

4.1 Project Limitations

A limitation of the project was the way the system was designed. In order for the game to start, the player must press the left side of the screen for a period of time. However, the left side of the screen also handles a different game action, thus making it possible to accidentally affect game operations in a negative way. This could have been solved by constructing a proper start screen containing a main menu and buttons.

Another limitation was that despite taking inspiration from Monster Hunter, which necessitates strategy in order to win against monsters, this system was not able reproduce the feeling of needing to strategize in order to win. This was due to

certain design flaws during development such as the damage dealt by the enemy, or utilizing more sophisticated AI.

5 Conclusions and Future Works

Various parts of the system itself can be improved on, such as adding a main menu or fine-tuning the player's movements using the accelerometer. The system could also be redesigned in order to develop a need for strategy to win. Research-wise, a study of how well accelerometer can improve the experience of playing third-person games could be conducted. A study of comparison between using the accelerometer and using a virtual joystick to see if players have any preference in control methods could also be done.

Acknowledgments. This project is partially supported by Artificial Intelligence Research Unit (AiRU).

References

1. Ling, S.: The Paradox of Technology on Social System (2012). http://blog.nus.edu.sg/glomerateapple/2012/11/18/our-complicated-relationship-with-technology/. Accessed 10 Oct 2013
2. Jung, B.: What Are the Benefits of Smartphone Technology? (2012). http://smallbusiness.chron.com/benefits-smartphone-technology-57037.html. Accessed 23 Nov 2013
3. Pearson, D.: Report: Mobile to become gaming's biggest market by 2015 (2014). http://www.gamesindustry.biz/articles/2014-10-22-report-mobile-to-become-gamings-biggest-market-by-2015. Accessed 10 Jun 2015
4. Google: Google Play Market (2015). https://play.google.com/store/apps?hl=en. Accessed 11 Jun 2015
5. Apple Inc.: App Store Review Guidelines (2010). https://developer.apple.com/app-store/review/guidelines/. Accessed 23 Jun 2012
6. Softonic Internacional: Android Low Quality Games Free (1997). http://en.softonic.com/s/low-quality-games-free-download:android. Accessed 9 Jan 2015
7. Unity Game Engine (2012). http://unity3d.com/unity/. Accessed 16 Feb 2013
8. Unreal Engine Unreal Development Kit Homepage (2012). http://www.unrealengine.com/en/udk/. Accessed 16 Feb 2013
9. TheTechPanda: Top Ten Must-Have Android 3D Racing Games with Awesome Graphics (2012). http://thetechpanda.com/2013/10/25/top-ten-must-android-3d-racing-games-awesome-graphics/#.VbShF_mUI1Y. Accessed 1 Jan 2015
10. Shi, J.L., Tan, T.G., Teo, J., Chin, K.O., Alfred, R., Anthony, P.: Evolving controllers for simulated car racing using differential evolution. Asia Pac. J. Inf. Technol. Multimed. **2**(1), 57–68 (2013)
11. Tan, T.G., Teo, J., Chin, K.O., Alfred, R.: A coevolutionary multiobjective evolutionary algorithm for game artificial intelligence. Asia Pac. J. Inf. Technol. Multimed. **2**(2), 53–61 (2013)
12. Leow, C.L., Gan, K.S., Tan, T.G., Chin, K.O., Alfred, R., Anthony, P.: Self-synthesized controllers for tower defense game using genetic programming. In: IEEE 2013 International Conference on Control System, Computing and Engineering, pp. 487–492 (2013)

13. Gan, K.S., Tan, T.G., Chin, K.O., Alfred, R., Anthony, P.: A comparison on the performance of crossover techniques in video game. In: IEEE 2013 International Conference on Control System, Computing and Engineering, Penang, Malaysia, pp. 493–498 (2013)
14. Guan, T.T., Nan, Y.Y., Chin, K.O., Teo, J., Alfred, R.: Automated evaluation for AI controllers in tower defense game using genetic algorithm. CCIS, vol. 378, Kuala Lumpur, Malaysia, pp. 135–146 (2013)
15. State Machines (2012). http://lazyfoo.net/articles/article06/index.php. Accessed 26 Nov 2012
16. Champandard, A.J.: Common Ways to Implement Finite State Machines in Games (2007). http://aigamedev.com/open/article/fsm-implementation/. Accessed 26 Nov 2012
17. Mob.org.: Free mobile downloads: android games (2012). http://play.mob.org/genre/gonki/. Accessed 1 Jan 2015
18. jMonkeyEngine: jMonkeyEngine Tutorials and Documentation (2007). http://jmonkeyengine.org/wiki/doku.php/jme3. Accessed 26 Nov 2012
19. Rabin, S.: Common game AI techniques. In: Rabin, S. (ed.) AI Game Programming Wisdom 2, pp. 3–14. Charles River Media Inc., Hingham (2004)

An Exploratory Study on Latent-Dirichlet Allocation Models for Aspect Identification on Short Sentences

Ameer Abu Bakar, Lay-Ki Soon(✉), and Hui-Ngo Goh

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
ameerasyrafab@gmail.com, {lksoon,hngoh}@mmu.edu.my

Abstract. This paper reports an exploratory study conducted to investigate the performance of topic modelling algorithms in aspect identification. Aspect identification is an important step in aspect-based sentiment analysis. Latent-Dirichlet Allocation model serves as the baseline of topic models in the experiments. One of the variations of LDA, namely Phrase-LDA was experimented to benchmark its performance against the original LDA. Although it was reported that PLDA performs better compared to LDA in aspect-based sentiment analysis, our experimental results indicate that LDA works better on dataset with short sentences. A new PLDA model was also proposed by using different types of dependencies to extract the phrases.

Keywords: Latent-Dirichlet Allocation · Topic modeling · Aspect identification Sentiment analysis

1 Introduction

In sentiment analysis, aspect identification is the process of identifying aspects of opinion target, which could be a product or an entity [5]. For instance, a restaurant being the opinion target may have been evaluated from different aspects, such as the food, the pricing, the ambience, its location or service qualities. These aspects can be topics from a set of documents – a corpus. Topic modelling is a model used to derive topics from a corpus; it is a form of text mining. Topic modelling algorithms group the words in the documents from the corpus into topics. In terms of identifying aspects from a corpus, topic modelling algorithms have been applied to identifying words which are frequently used to express opinions on specific aspects [1].

In this paper, topic modelling algorithms are explored for identifying aspects in sentiment analysis. To accomplish the task, several topic modelling algorithms are studied and investigated. Latent Dirichlet Allocation (LDA) serves as the main topic modelling algorithm is selected for this exploratory study. LDA was first published by Blei et al. [2]. Subsequently, many new variations were proposed, either by extending it; or creating improved or new versions to the original LDA model. LDA is a simple topic model that is used to gather topics from their corpus. LDA outputs a bag-of-words

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 314–323, 2018.
https://doi.org/10.1007/978-981-10-8276-4_30

for each topic and the user will have to label each bag-of-words for the lists to specific topic.

The goal of this exploratory study is to compare the performance of the chosen topic modelling algorithms in identifying the aspects from a set of online reviews. The online reviews are labelled with aspects and they are about restaurant (as the opinion target or entity). The list of topics generated by the topic modelling algorithms can then be used to derive the aspects discussed in the online reviews.

2 Literature Review

2.1 Background Study

LDA is a topic modelling algorithm used to discover the topics of a corpus by using probabilistic models; the probabilistic models used in LDA are the multinomial distribution and the Dirichlet distribution. Each document in a corpus will have a mixture of topics. Each document in the corpus is assigned to several topics using probabilities. The input here would be the corpus and the output would be a bag of words for each topic. Each bag of words will have to be labeled by the user at the end [2]. To better explain LDA, consider the three sentences given below, where each sentence is considered as a document:

- Sentence A: The restaurant was serving uncooked food.
- Sentence B: The waiter served chicken and fish.
- Sentence C: Chickens and cows are farm animals.

Step 1: Choose how many topics there are to be found from the documents. There are likely to be two bag-of-words for the topics of animal and food.

Step 2: Every word in the documents, except stop-words (useless words), will be assigned to a temporary topic using a Dirichlet distribution. This step will be iterated for as long as the user thinks is necessary to get an accurate generated output.

Step 3: The topic will then attempt to generate the words itself by using the multinomial distribution.

Output for the bag of words:

- Topic A: food, uncooked, chicken, restaurant, served…
- Topic B: chicken, fish, animal, cow, farm…

From this output, topic A is about food while topic B is about animals. Hence, sentence A can be labelled as topic A, sentence B can be labelled as topic A and sentence C can be labelled as topic B.

2.2 Related Work

Aspect identification is one of the crucial processes in aspect-based opinion mining. Aspect-based topic modelling aims to identify aspects given the topics produced by LDA. In this section, different variations of LDA discussed by Moghaddam and Ester [6] are briefly explained in Table 1.

Table 1. Description of different variations of LDAs

LDA variation	Data input format	Main differences
LDA	Bag of words	Only one latent variable needed to assign a topic to a word
S-LDA [3]	Bag of words	Two separate variables: rating and aspect instead of the one variable. Both the variables are sampled independently
D-LDA [4]	Bag of words	Like S-LDA but there is a dependency between rating and aspect. The sampling of rating takes into account of the sampled aspect variable
PLDA	Bag of phrases	Unlike LDA, PLDA uses a bag of phrases model. Instead of one variable 'word' for the review, there is two variables: a head term and a modifier
S-PLDA	Bag of phrases	S-PLDA is similar to S-LDA with its two observed variables: aspect and rating but its difference is that it uses the head term and modifier variables. The head term depends on the aspect variable while the modifier only depends on the rating variable
D-PLDA	Bag of phrases	It similar to S-PLDA except there are dependencies between aspect and rating. The modifier will now depend on both the aspect and rating unlike in S-PLDA

LDA is a probabilistic model that is able to retrieve topics from many documents (corpus). This is the original topic modelling algorithm. It learns the topics of review using all the words of the review; it uses a bag of words model. There is only one variable (z variable) for the topic as opposed to two variables in S-LDA [3]. The result of LDA would be lists of words where each belongs to a certain topic [2].

The θ value is first sampled by using the Dirichlet probability on α value. After the θ value is sampled, for each word, the topic is then sampled with $P(zn \mid \theta)$; the zn variable is the variable for topics. The word is then sampled again using $P(wn \mid z\neg n, \beta)$.

Given the descriptions and experimental results presented by Moghaddam and Ester [6], both the original LDA and its variations – PLDA were implemented and compared in this project. PLDA is chosen because it is a direct variation of LDA which is predicted to be much faster and better. Instead of the bag-of-words model in LDA, PLDA uses the bag-of-phrases model.

3 Experimental Design

3.1 Experimental Setup for LDA

The dataset used in the experiments contain English reviews. For the comparison study purpose, the reviews were experimented in both stemmed and un-stemmed formats. The LDA model are experimented on both stemmed and un-stemmed reviews. Both begin with the conversion of the XML restaurant data-set into a text-file. Next, the text-file

will have to be converted into the correct form (Term-Document Matrix) using the text-mining package in python. The term-document matrix represents the number of occurrences of terms of the corpus (text-file). After the conversion to the TDM form, the LDA package in Python can finally begin with the term-document matrix as input. The LDA package will then output ten bag-of-words for twenty topics. The LDA package will automatically assign each document (each line in the text-file) a topic (bag-of-words). The user can then have to manually label the bag-of-words – there are six possible labels (aspects): restaurant, food, location, drinks, ambience and service. These aspects are selected from the experimental dataset.

3.2 Experimental Dataset

The dataset contains restaurant reviews in XML form. The XML file is displayed in Fig. 1. The text has an opinion; an opinion consists of a target, category, polarity (positive or negative). The category refers to aspect of the target. Each sentence in the data-set is counted as a review; therefore all the reviews consist of a short sentence. The review in Fig. 1 is about the place or location of the restaurant.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Reviews>
    <Review rid="1004293">
        <sentences>
            <sentence id="1004293:0">
                <text>Judging from previous posts this used to be a good place, but not any longer.</text>
                <Opinions>
                    <Opinion target="place" category="RESTAURANT#GENERAL" polarity="negative" from="51" to="56"/>
                </Opinions>
```

Fig. 1. Restaurant reviews in XML format, with labelled opinion target and category (aspect)

The LDA package in python requires the dataset to be in three different forms: LDA-C, tokens and titles. The LDA-C file is in the form of a document-term matrix, where each line represents a document and each number in the line is in the form A:B. A represents the first index of the word and B represents the number of occurrences. The tokens file consists of all the words in the documents excluding the stop-words and the titles file consists of all the titles of the documents.

3.3 Data Pre-processing

To preprocess the data, the data was extracted from the XML file. This was done using the *xml.etree.ElementTree* package for Python 3.4. After using this package, the words were all changed to lower-case. After this, the stop-words were then removed. To get the LDA-C, tokens and titles files, the text mining package in python was then used. Some stop-words were: a, am, and I. These were removed because they do not carry much meaning. Hence, it is not necessary to assign these stop-words to a topic. There were two different variations of LDA used in the package: LDA and PLDA. There are also two versions of LDA and PLDA: stemmed LDA and un-stemmed LDA for the LDA variation and PLDA with original dependencies and PLDA with new dependencies for the PLDA variation.

For LDA stemmed: the stemming of the dataset was conducted after the stop-words were removed. This is to make sure that there is only one variation for a word. For example: if there were "price" and "prices" in the data-set, the stemmer would change it to "price". This can be seen in Fig. 2, the word "judging" changed to "judg" after stemming.

```
judg previous post good place longer
arriv noon place staff act impos rude
brought complimentari noodl repeat request sugar threw dish table
```

Fig. 2. The stemmed dataset after pre-processing

For LDA un-stemmed: after the removal of the stop-words, the file is left as it is so that it is possible to compare the difference between stemmed and un-stemmed. This version of the LDA input file has more words than the LDA stemmed input file. This can be seen in Fig. 3, the stop-words like "from" were removed from the file.

```
judging previous posts good place longer
arrived noon place staff acted imposing rude
brought complimentary noodles repeated requests sugar threw dishes table
```

Fig. 3. The un-stemmed dataset after pre-processing

For PLDA with the original dependencies: after the removal of the stop-words, the un-stemmed dataset was then transformed to the PLDA version based on the nine different dependencies proposed by the article. The dependencies were extracted from the un-stemmed data-set using the Stanford parser. Table 2 shows the dependencies used in PLDA with examples. Table 3 explains the terms used in the dependencies. The un-stemmed data-set was transformed to the PLDA version by combining the words with the original dependencies, as listed in Table 2. The result of this is shown in Fig. 4, terms like "*posts~previous*" were created from the original dependencies in the article.

Table 2. Nine different dependencies used for PLDA

First dependency	Second dependency	Resulting dependency	Example of the results
amod (N, A)	none	<N, A>	<restaurant, favourite>
acomp (V, A)	nsubj (V, N)	<N, A>	<people, friendly>
cop (A, V)	nsubj (A, N)	<N, A>	<food, excellent>
dobj (V, N)	nsubj (V, N′)	<N, V>	<dessert, recommended>
<h1, m>	conj and(h1, h2)	<h2, m>	<service, good>
<h, m1>	conj and(m1, m2)	<h, m2>	<list, impressive>
<h, m>	neg(m, not)	<h, not + m>	<place, not.recommend>
<h, m>	nn(h, N)	<N + h, m>	<noodle.soup, good>
<h, m>	nn(N, h)	<h + N, m>	<pad.thai, delicious>

Table 3. Terms used in dependencies

Term	Meaning
N	Noun
A	Adjective
V	Verb
h	Head-term
m	Modifier

```
posts~previous place~good
place~empty they~rude
noodles~complimentary requests~repeated
```

Fig. 4. The PLDA with original dependencies dataset after pre-processing

For PLDA with new dependencies: after the removal of the stop-words, the un-stemmed data-set was then transformed to the PLDA version based five of the dependencies: amod (adjectival modifier), acomp (adjectival complement), nsubj (nominal subject), cop (copula) and dobj (direct object).

3.4 Evaluation Metrics

A confusion matrix is used to evaluate the results [7]. A confusion matrix shows the number of correct and incorrect predictions made by the topic modelling algorithms compared to the actual results labelled in the dataset (Table 4).

Table 4. Confusion matrix

	Predicted – Positive	Predicted – Negative
Actual – Positive	True Positive	False Negative
Actual – Negative	False Positive	True Negative

By using the figures in confusion matrix, accuracy of LDA and PLDA in predicting or identifying the aspects in the opinion can be calculated as:

$$(\text{True Positive} + \text{True Negative})/(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}) \quad (1)$$

4 Results and Findings

4.1 Experimental Results

There were two input files for LDA: stemmed and un-stemmed. The stemmed LDA has an alpha value of 0.0071 and a beta of 0.0037 while the un-stemmed data-set has an alpha value of 0.0071 and a beta of 0.0035. Figures 5 and 6 show the output of the LDA

program with its stemmed and un-stemmed input files. Each topic has 10 words (bag-of-words) and the user has to manually label the topic by determining what the topic of the words are. There were also two input files for PLDA: with original dependencies and with new dependencies. They both had the same values for alpha and beta: 0.0071 and 0.0035 respectively. The phrases consists of a head-term and a modifier. For example, for the phrase "food~good", the head-term is "food" and the modifier is "good".

```
×Topic 0
- great servic place friend excel service nice dinner ambienc pizza
×Topic 1
- famili year world food serv italian size amaz portion 5
×Topic 2
- thing dish small appet night chicken waiter order dessert late
×Topic 3
- good food servic bad service work time peopl staff rude
×Topic 4
- food fish delici sushi delicious fresh menu excellent rice thai
```

Fig. 5. An output of the program using LDA stemmed

The outputs of the files were then labelled using crowd-sourcing website – Crowdflower. Crowdflower was used to help determine the categories for each list of words. Crowdflower has an online workforce of users that help to clean and label data. For the laptop dataset, three different data were given to the online workforce. Each different data had different iterations: 10 iterations, 20 iterations and 30 iterations. For the restaurant dataset, there were 2 different data: 10 iterations and 20 iterations.

There were 6 different pre-defined topics extracted from the benchmarked restaurant dataset: restaurant, ambience, food, service, location, and drinks. The users in Crowdflower will then have to label each list of words to one of the pre-defined topics. The outputs from LDA and PLDA were then compared with the extracted aspects from the XML file.

An output of the LDA program using a stemmed text-file input is shown in Fig. 5. An example of a stemmed word is "servic", it is derived from the words service and services. Every topic in the figure has 10 words associated with it; the user will have to label each topic by looking at the words. For example, topic 3 in Fig. 5 has more words ("servic", "bad", "service, "time", "people", "staff", "rude" are all words that could belong to service) associating with the topic service therefore it can be labelled as service.

There are more duplicates of a word in the un-stemmed text-file. Therefore there are more unique words. There are 10 words associated with each topic. Each topic will have to be labelled by the user manually. For the output of the PLDA model using the original dependencies (dependencies from the article), each phrase in a topic consists of a head-term and modifier. For example: "service~slow" is a phrase, service is the head-term and slow is the modifier. A head-term is the aspect and the modifier is the rating (polarity). The output of the PLDA program using new dependencies show that there are a lot more unique words in this data-set compared to the previous PLDA program.

Each topic has 10 phrases and each phrase consists of a head-term and modifier. The user will have to manually label each topic.

4.2 Findings

The results presented here are the accuracy of predicting aspects in the review using LDA and PLDA, against the labelled aspects (categories) in the original XML dataset. During the experiments, the alpha and beta values were tweaked with and changed; they were not kept at the default values. There was a noticeable difference after changing the values. The iterations were also changed, to 1000 iterations, 1250 iterations, 1500 iterations, 1750 iterations, and 2000 iterations. Changing the iterations was needed to determine what the optimal iteration number was. In addition to the PLDA presented in Moghaddam and Ester [6], a new PLDA model was also proposed by using different types of dependencies to extract the phrases. There were more phrases in the new PLDA with new dependencies and more stop-words in the phrases. Tables 5, 6, 7 and 8 show the results of different variations of LDA, PLDA and new PLDA in different iterations. For fair comparisons, three trials were performed for each iterations.

Table 5. Results of stemmed LDA in three trials

Iterations	Trial 1	Trial 2	Trial 3
1000	0.49	0.49	0.59
1250	0.55	0.49	0.52
1500	0.58	0.60	0.59
1750	0.59	0.53	0.62
2000	0.60	0.55	0.56

Table 6. Results of un-stemmed LDA in three trials

Iterations	Trial 1	Trial 2	Trial 3
1000	0.54	0.61	0.58
1250	0.58	0.67	0.56
1500	0.65	0.58	0.60
1750	0.57	0.58	0.61
2000	0.60	0.61	0.57

Table 7. Results of PLDA with original dependencies in three trials

Iterations	Trial 1	Trial 2	Trial 3
1000	0.40	0.43	0.41
1250	0.50	0.45	0.40
1500	0.44	0.47	0.44
1750	0.46	0.51	0.44
2000	0.46	0.46	0.46

Table 8. Results of PLDA with new dependencies in three trials

Iterations	Trial 1	Trial 2	Trial 3
1000	0.37	0.37	0.36
1250	0.36	0.35	0.41
1500	0.38	0.41	0.46
1750	0.40	0.42	0.42
2000	0.49	0.46	0.47

For PLDA, having more stop-words, it was harder to determine what some of the topics were. For example, it is hard to label topic 0 in Fig. 8 because the phrase "ok~is" does not belong to one of the pre-defined 6 titles. This led to worse results as seen in Table 8 (a table showing the results of PLDA with new dependencies) compared to Table 7 (a table showing the results of PLDA with the original dependencies).

From Table 9, it can be deduced that the un-stemmed LDA model was the best model when using this particular restaurant data-set; it has a percentage of 59.5%. The dataset is too small for the PLDA models because for some reviews there were only few phrases. The un-stemmed LDA model had the most unique words in this small data-set, therefore it could calculate and assign the topics more accurately. The stemmed model had the second most words in its text-file compared to the other models so it would logically make sense that it had the second most accurate average percentage of 55.7%.

Table 9. Comparison of the four models

Iterations	Stem	Unstem	PLDA	newPLDA
1000	0.52	0.58	0.41	0.37
1250	0.52	0.60	0.45	0.37
1500	0.59	0.61	0.45	0.42
1750	0.58	0.59	0.47	0.41
2000	0.57	0.59	0.46	0.47
Average	**0.56**	**0.59**	**0.45**	**0.41**

5 Conclusion

In this paper, LDA and its variation – PLDA were explored for aspect identification from restaurant reviews. PLDA is a new variation that uses a bag-of-phrases model with a head-term and a modifier instead of just one variable that is used in LDA. From the main reference, PLDA model was reported to have better accuracy, recall, precision, perplexity and MSE than the LDA model. In addition, a new PLDA was also experimented where it contains more dependencies. The performance of the four variations that were implemented were compared and the best variation for this data-set is the un-stemmed LDA because of the small dataset. To our surprise, our experimental results show that LDA performed better in identifying aspects in the restaurant dataset, instead of PLDA. From the experimental results, it can be concluded that PLDA requires data

in longer sentences to perform well while LDA is more robust as it works well with short sentences.

References

1. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Lakkaraju, H., Bhattacharyya, C., Bhattacharya, Y., Merugu, S.: Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: SDM 2011 (2011)
4. Li, F., Huang, M., Zhu, X.: Sentiment analysis with global topics and local dependency. In: AAAI 2010 (2010)
5. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, San Rafael (2012)
6. Moghaddam, S., Ester, M.: On the design of LDA models for aspect-based opinion mining. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 803–812. ACM, October 2012
7. Markham, K.: Simple guide to confusion matrix terminology, 26 March 2014. http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Evaluation of Artificial Neural Network in Classifying Human Gender Based on Odour

Ahmed Qusay Sabri[1(✉)] and Rayner Alfred[2(✉)]

[1] University of Sharjah, Sharjah, UAE
asabri@sharjah.ac.ae

[2] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
ralfred@ums.edu.my

Abstract. Biometrics is an advanced way of person recognition as it establishes more direct and explicit link with humans than passwords, since biometrics use measurable physiological and behavioral features of a person. In this paper, a gender recognition framework is proposed based on human odour. 20 samples of human odour from male and female are collected and only 16 out of 198 Volatile Organic Compounds (VOCs) are selected using the Chi-square test and entropy for gender detection and classification using artificial neural networks. In this paper, several different neural network activation functions were tested (e.g., Levenberg-Marquardt backpropagation, Gradient descent backpropagation and Resilient backpropagation) and several different neural network topologies are also tested with variety of hidden layers and number of neurons. It is also found that with 2 hidden layers having more number of neurons in the hidden layers (16 and 16 neurons in which hidden layer) was able to produce greater performance accuracy. The best learning algorithm that can be applied in gender detection shown in paper is the Gradient Descent learning algorithm. Also, it is notable that 8 out of 9 cases where all male samples are able to be detected or classified correctly compared to the 3 out of 9 cases in which all females are correctly detected or classified.

Keywords: Levenberg-Marquardt backpropagation
Gradient descent backpropagation · Resilient backpropagation · VOC · Odour

1 Introduction

Gender is one of the demographic attributes of human being; apart from gender there are various other demographic attributes like age, ethnicity, which can be identified by computer vision and can be applied to many applications such as human computer interaction, surveillance, biometrics and demographic studies [3, 7, 18]. In recent years, identification of demographic attribute using computer vision is becoming a great challenge. Gender recognition can be done using different approaches like using face features, audio signal frequency [15, 18]. Gender classification problem is an active research area which has attracted a great deal of attention recently [5]. Many techniques for solving the problem of gender identification use facial features [4, 5]. Among many

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 324–337, 2018.
https://doi.org/10.1007/978-981-10-8276-4_31

trustworthy biometric traits, such as face is a very popular one and it have this reputation for accessibility. But unfortunately, the malicious environments, enabling attackers to easily create photocopies and spoof face recognition systems. Spoofing is an attempt to gain authentication through a biometric system by presenting a forgery evidence of a valid user. This vulnerability of face has evoked significant attention in the biometric community and numerous papers have been published in countermeasure studies [4]. Gender identification based on the voice of a speaker consists of detecting if a speech signal is uttered by a male or a female. Automatically detecting the gender of a speaker has several potential applications [3]. The highest gender identification rate that can be achieved by using voice of a speaker is 83.3% [3].

Classification of soft biometric traits in terms of gender, ethnicity, dour and age still far from being considered as a solved problem for the case of difficult exposure conditions [13, 19]. The advantage to a biometric is that it doesn't change or lose. Many body parts, personal characteristics and imaging methods have been used for biometric systems such as fingers, hands, feet, eyes, ears teeth, veins voices, signatures, typing styles and gaits. Each biometric has its own strength and limitations and accordingly each biometric is used in Identification (authentication) applications [2]. In our research GC/MS test (Gas Chromatograph Mass Spectrometry) applied on Male and Female human on different days to ensure the stability of emitted VOCs (Volatile Organic Compounds), we reach the most stable, accurate and rigidity list of VOCs from human with specific Gas name, getting the result of our short list of 16 VOCs to be used for the gender detection process.

The aim of this paper is to investigate the performance of the Artificial Neural Network (ANN) under different settings in classifying human into male and female classes based on the VOCs emitted by human. These settings include different activation functions and different topologies that will be applied to the target of gender detection in which 17 VOCs are collected and considered as input to neural networks. These VOCs emitted from human are fed into the ANN in which three different activation functions are tested (e.g., Levenberg-Marquardt, Gradient descent and resilient backpropagations) and several different neural network topologies are also tested with variety of hidden layers and number of neurons.

The rest of the paper follows, where Sect. 2 will highlight some related works on genders classification. Section 3 will describe the processes involved in gender detection for gender Classification. Section 4 discusses and analyses the results obtained. Finally, the conclusion of the research is presented in Sect. 5 by suggesting future works that can be conducted based on the results obtained in this work.

2 Related Works

Individuals are thought to have their own distinctive scents, analogous to a signature or fingerprint, the axillary region is of particular interest to play an important role in generating individual odor [20, 21]. Gender recognitions from human gait in image sequence have been successfully investigated [6]. Gender recognition system have achieved recognition performance of 93.4%, 94.6%, and 94.7% with 2 layers/20

neurons, 3 layers/30 neurons and 4 layers/30 neurons respectively [6]. One of key features of (initially anonymous) interacting user is her/his gender. Automatic gender recognition can be considered as a method studied in the domain of biometry. Artificial neural network classifier assures high accuracy of gender recognition [16]. Works on the classification of gender into male or female classes have been accomplished by using well-known neural network architecture named multi-layered perceptron (MLP) with back-propagation training algorithm [17]. Back Propagation Artificial Neural Network is used for classifying the gender, gives 100% accurate results in identifying male and female [13]. Biometric face and fingerprint recognition using neural network system for identity verification produces good accurate result with high efficiency [2, 8, 11, 12]. Biometric hands and palm print recognition using feed forward back propagation neural network with Levenberg-Marquardt training algorithm is used in Palm print Biometrics classification, and the obtained results show that the best classification accuracy can be achieved with 99.998% for two hidden layers [9, 10]. As a result, in this work, the Levenberg-Marquardt training algorithm is used as one of the three neural networks training algorithms.

The use of artificial neural networks in biological informatics is very effective, neural network consists of three layers input, hidden and output layers trained using momentum back propagation learning method with gradient descent used for person identification to classify subjects with a classification accuracy of 97% [22]. A multilayer neural network structure with LevenbergMarquardt algorithm was used as training algorithm for the weights update of the neural network for hepatitis diseases diagnosis, results obtained a classification accuracy of 91.87% [23]. Classification of brain and heart biometrics using artificial neural network classifier results accuracy of 92.4% to 95.1% [24].

3 Gender Detection Process from Human Odour

3.1 Sweat Collection and GC/MS (Gas Chromatograph Mass Spectrometry) Test

Individuals will be tested in successive days of the week, in which different genders and variety of ages are also considered. Individuals will have a wash out phase for seven days before starting the test program. In these seven days, the volunteers are only allowed to wash their under arms with neutral soap. All days in the morning the Individuals come in the lab and wash their under arms with neutral soap this is a standardized procedure. After that they will put on a cotton t-shirt in which pads will be placed under the arm pits. The pads will be worn by these individuals for 4 h and during this time they do their normal day actions. In order to get better odor representation, individuals are asked to run stairs up and down for 3 times approximately 3 min per time point. After 4 h the individuals bring the pads to the lab.

A sweat sample is introduced into a Nalophan bag with 1.5 L nitrogen and heated to 90 °C for 30 min. Afterwards, the headspace is collected from the nalophan bag of each sample into a thermodesorption tube (Tenax/Carbograph) until a total volume of 1000 ml. Additionally, one thermodesorption tube without sample and under the same sampling conditions than real samples was put into nalophan bag used in sampling.

These tubes were used as blanks, the objective is Identify all compounds in sweat samples. Figure 1 shows one test result for one person for all VOCs emitted from human sweat, 4 volunteers have been tested 5 successive days each, different genders are considered in volunteers, results of peaks have some differences for same person on different days, X axis in Fig. 1 represents the number of minutes required to the sample inside the VOC detection device while Y axis represents the concentration of VOC detected. Table 1 outlines a list of 16 VOCs emitted from human and will be applied as an input to the artificial neural network for the gender detection target. These 16 VOCs are selected and shortlisted by using the Chi-square test and also the entropy.

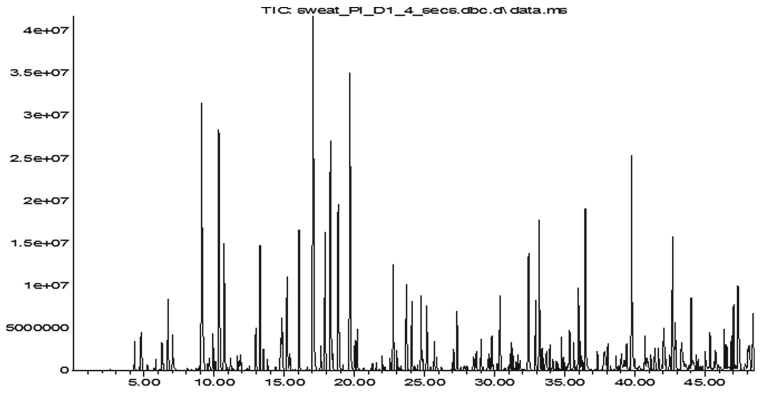


Fig. 1. Result of VOCs emitted from human sweat

Table 1. List of 16 VOCs emitted from individuals

No	VOCs
1	1-Propanol, 2-methyl-
2	Acetaldehyde (*)
3	2-Propenal
4	Propanal, 2-methyl-
5	Methacrolein
6	Butanal, 3-methyl-
7	2-Butenal, (E)-
8	Phenol
9	Furan
10	Furan, 2-methyl-
11	Furan, 2-pentyl-
12	Acetone
13	2-Butanone
14	3-Penten-2-one
15	2-Pentanone
16	3-Hexen-2-one, 5-methyl-

3.2 Artificial Neural Networks

Classification is one of the most important task in application areas of artificial neural networks (ANN) [1]. There are number of batch training algorithms which can be used to train a network, three types of training algorithms have been evaluated for classification [1]. Gradient Descent backpropagation algorithm is a gradient descent local search procedure. It measures the output error, calculates the gradient of the error by adjusting the weights in the descending gradient direction [1]. Resilient backpropagation training algorithm eliminates the effects of the magnitudes of the partial derivatives. In this sign of the derivative is used to determine the direction of the weight update and the magnitude of the derivative have no effect on the weight update. The size of the weight change is determined by a separate update value. The update value for each weight and bias is increased by a factor whenever the derivative of the performance function with respect to that weight has the same sign for two successive iterations [1]. Levenberg–Marquardt backpropagation algorithm locates the minimum of a multivariate function that can be expressed as the sum of squares of non-linear real-valued functions. It is an iterative technique that works in such a way that performance function will always be reduced in every iteration of the algorithm [1]. The backpropagation neural network learns by calculating the errors of the output layer to find the errors in the hidden layers. This qualitative ability makes it highly suitable to be applied on problems in which no relationship is found between the output and the inputs. The ANN techniques very appealing in application domains for solving highly nonlinear phenomena [14]. Due to its high rate of plasticity and learning capabilities, it has been successfully implemented in wide range of applications [25, 29]. The relationship between the neural network output Y and the input X is given by [29]:

$$\gamma_t = \omega_o + \sum_{j=1}^{q} \omega_j \cdot f\left(\omega_{0,j} + \sum_{i=1}^{p} \omega_{i,j} \cdot x_t\right) \tag{1}$$

where $w_{i,j}(\mathrm{i} = 0, 1, 2, \ldots, \mathrm{p}; \mathrm{j} = 1, 2, \ldots, \mathrm{q})$ and $w_j(\mathrm{j} = 0, 1, 2, \ldots, \mathrm{q})$ are the connection weights, p is the number of input nodes, q is the number of hidden nodes, and f is a nonlinear activation function that enables the system to learn nonlinear features. The most widely used activation function for the output layer are the sigmoid and hyperbolic functions [29].

In this paper, since the aim of this paper is to investigate the effects of varying the parameters settings used for gender detection, three settings are used as shown in Fig. 2 through Fig. 4, in which the output neuron is either 0 to indicate a female or 1 to indicate a male. These settings include the following topologies

- 16 inputs, 16 neurons hidden layer and 1 neuron output layer (Fig. 2)
- 16 inputs, 2 hidden layers (16 neurons and 10 neurons hidden layers) and 1 neuron output layer (Fig. 3)
- 16 inputs, 2 hidden layers (16 neurons and 16 neurons hidden layers) and 1 neuron output layer (Fig. 4)

For each topology, there are also three learning schemes that will be used in the ANN and these learning algorithms are Levenberg-Marquardt backpropagation, Gradient descent backpropagation and Resilient backpropagation.

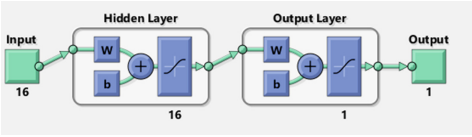


Fig. 2. 16 inputs, 16 neurons hidden layer and 1 neuron output layer.

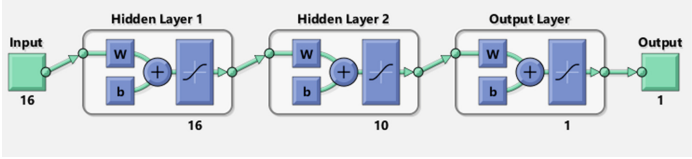


Fig. 3. 16 inputs, 16 neurons hidden layer 1, 10 neurons hidden layer 2 and 1 neuron output layer

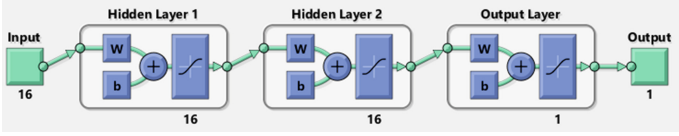


Fig. 4. 16 inputs, 16 neurons hidden layer 1, 16 neurons hidden layer 2 and 1 neuron output layer.

Levenberg-Marquardt Backpropagation

The Levenberg-Marquardt algorithm is a variation of Newton's method that was designed for minimizing functions that are sums of squares of other nonlinear functions. This is very well suited to neural network training where the performance index is the mean squared error [23, 27]. The Levenberg Marquardt (LM) algorithm is an approximation to the Newton method used for training ANNs. This optimization technique is more powerful than standard Backpropagation Neural Network (BPNN). LM algorithm is very efficient and fast having also a quite good global convergence property [28]. LM algorithm is shown as follows [28]:

Initialize Weights
While not Stop Criterion DO
Calculate $e^P(\omega)$ for each pattern
Calculate $e1 = \sum_{p=1}^{P}[e^P(\omega)]^2$
Calculate $J^p(\omega)$

Repeat:
Calculate $\Delta\omega$

$$\Delta\omega = \left[\mu I + \sum_{p=1}^{P} J^p(\omega)^T J^p(\omega)\right]^{-1} \nabla E(\omega)$$

$$e2 = \sum_{p=1}^{P} e^p(\omega + \Delta\omega)^T e^p(\omega + \Delta\omega)$$

If e1 ≤ e2
Then μ = μ ∗ β
Until (e2 < e1)
μ = μ/β
$\omega = \omega + \Delta\omega$

where:

$J^p(\omega)$	: Jacobian matrix of derivatives of each error to each weight.
μ	: A scalar.
$e^P(\omega)$	: The vector error of pattern p.
I	: The identity Matrix.
β	: A factor.
$\Delta\omega$	: Equation for update ANN weights.

Gradient Descent Backpropagation

The gradient descent used to gradually, but consistently, decrease the output error by adjusting the weights. The weights and biases are updated in the direction of the negative gradient of the performance function, Backpropagation is used to calculate derivatives of performance with respect to the weight and bias variables. For multi-layer networks, the relationship between the errors and any weights in the network needs to be calculated.

$$\Delta\omega_k = -\alpha_k \cdot g_k \tag{2}$$

where, $\Delta\omega_k$ is a vector of weights changes, g_k is the current gradient, α_k is the learning rate that determines the length of the weight update [29]. This involves propagating the error at the output nodes backwards through the network, one layer at a time, at each layer error is computed for each node. Gradient descent backpropagation computations shown below [29]:

- Calculated the feed-forward signals from the input to the output
- Calculate output error e:

$$e = \frac{1}{2}\sum_k \left(T_k - O_k\right)^2 \tag{3}$$

where: T_k: The output vector of the network and O_k: The desired output vector

- Backpropagate the error signals by weighting it by the weights in previous layers and the gradients of the associated activation functions
- Calculating the gradients for the parameters based on the backpropagated error signal and the feedforward signals from the inputs
- Update the parameters using the calculated gradients

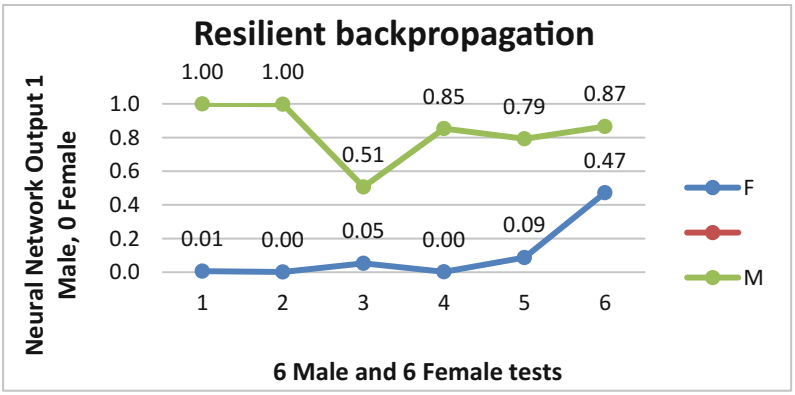
Resilient Backpropagation

Resilient backpropagation is considered the best algorithm, measured in terms of convergence speed, accuracy and robustness with respect to training parameters, resilient backpropagation algorithm offers faster convergence and is usually more capable of escaping from local minima [26, 29]. Only the sign of the partial derivative is considered to determine the direction of the weight update multiplied by the step size. Weights update in resilient backpropagation is updated as follows [29]:

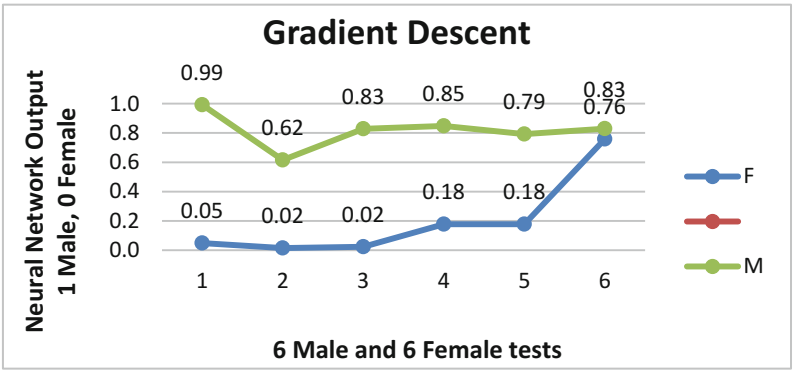
$$\Delta\omega_k = -sign\left(\frac{\Delta E_k}{\Delta\omega_k}\right) \cdot \Delta_k \tag{4}$$

4 Results and Analysis

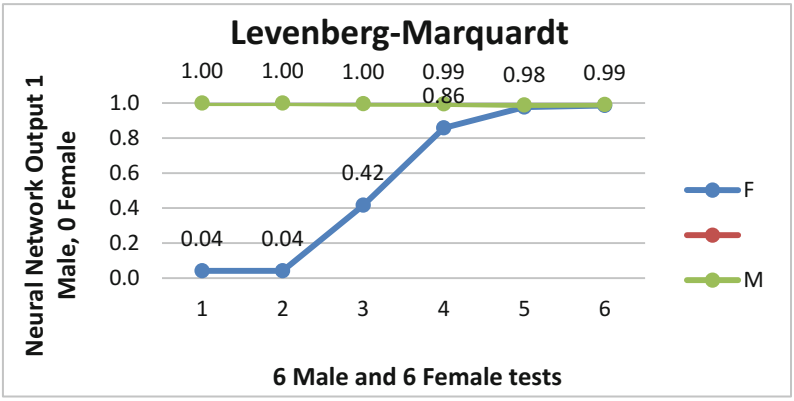
For the setting that consists of 16 inputs, 16 neurons hidden layer and 1 neuron output layer (Fig. 2), using the Resilient backpropagation as the learning algorithm, based on the results shown in Fig. 5(a), it shows that there are 5 out of 6 male detected perfectly, and 5 out of 6 female detected accurately with 83.3% performance accuracy. Next, the Gradient Descent backpropagation is then applied and produced results as shown in Fig. 5(b). The results showed that all male detected with values above 0.6, but 1 female classified as male. Thus the accuracy performance is 91.7%. Finally, the Levenberg-Marquardt backpropagation is applied to the topology described in Fig. 2. Based on the produced results shown in Fig. 5(c), 6 Male and only 4 Female samples are detected correctly. The accuracy performance obtained when using the Levenberg-Marquardt backpropagation is 83.3%.



(a) Resilient backpropagation

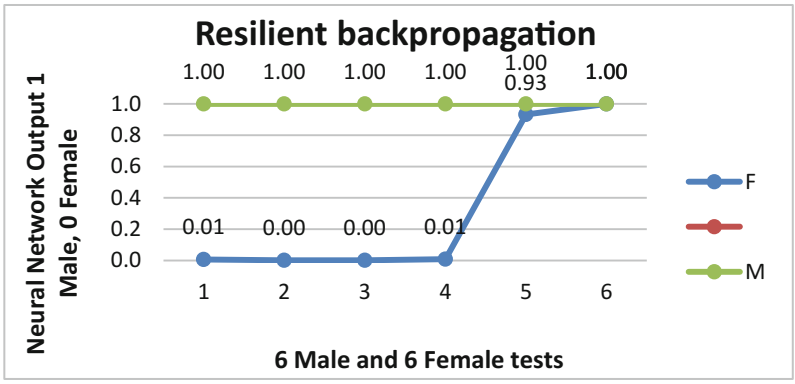


(b) Gradient Descent backpropagation

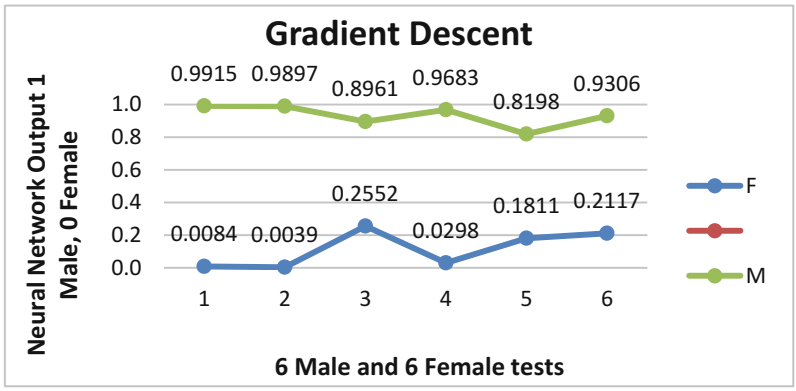


(c) Levenberg-Marquardt backpropagation

Fig. 5. Results obtained when using 16 inputs, 16 neurons hidden layer and 1 neuron output layer using different learning algorithms

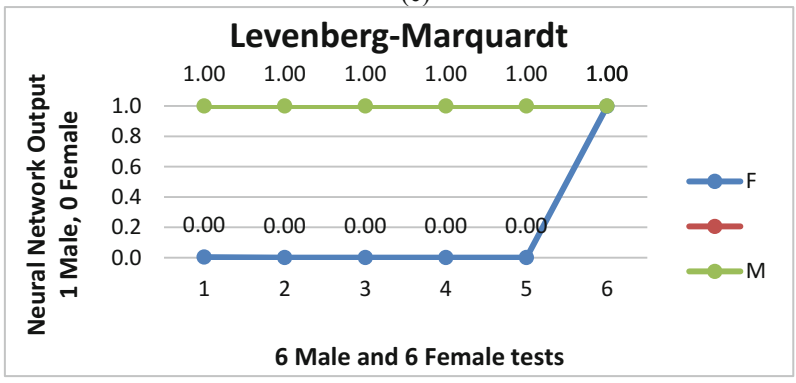


(a) Resilient backpropagation



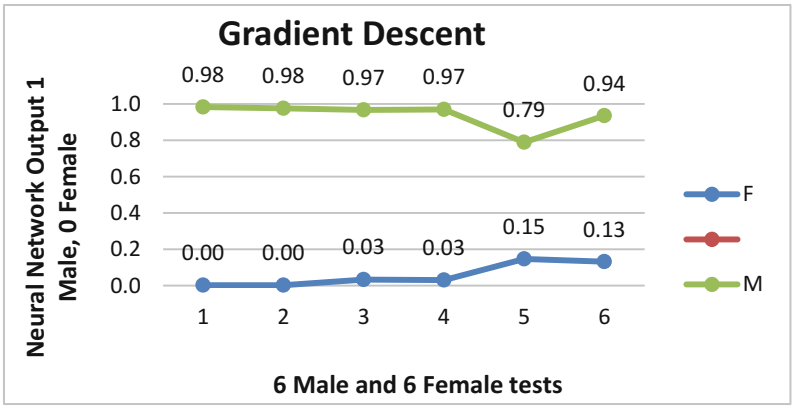
(b) Gradient Descent backpropagation

(c)

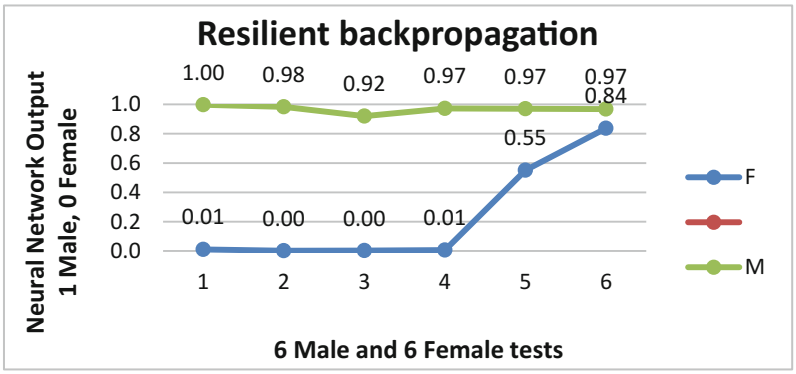


(c) Levenberg-Marquardt backpropagation

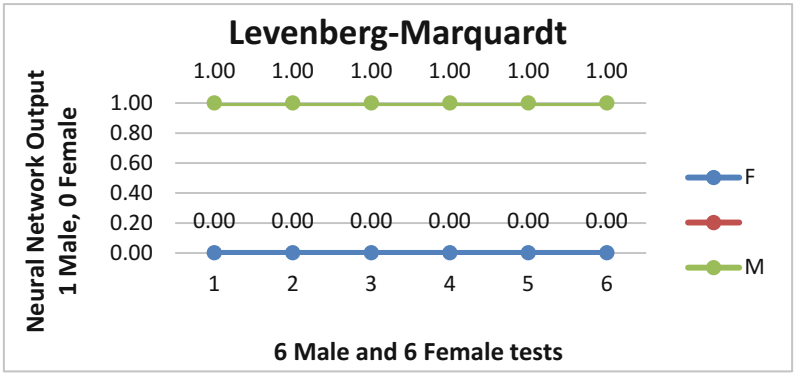
Fig. 6. Results obtained when using 16 inputs, 16 neurons hidden layer 1, 10 neurons hidden layer 2 and 1 neuron output using layer using different learning algorithms



(a) Resilient backpropagation



(b) Gradient Descent backpropagation



(c) Levenberg-Marquardt backpropagation

Fig. 7. Results obtained when using 16 inputs, 16 neurons hidden layer 1, 16 neurons hidden layer 2 and 1 neuron output using layer using different learning algorithms

For the second setting that consists of 16 inputs, 2 hidden layers (16 neurons and 10 neurons hidden layers) and 1 neuron output layer (as shown Fig. 3), based on the results shown in Fig. 6(a), the accuracy performance obtained is 83.3% when using the Resilient backpropagation, in which 2 female samples are incorrectly classified. Next, based on the results shown in Fig. 6(b), the accuracy performance obtained is 100% when using the Gradient Descent backpropagation, in which all male and female samples are correctly classified. On the other hand, based on the results shown in Fig. 6(c), the accuracy performance obtained is 91.7% when using the Levenberg-Marquardt backpropagation, in which 1 female sample is incorrectly classified.

Finally, for the third setting that consists of 16 inputs, 2 hidden layers (16 neurons and 16 neurons hidden layers) and 1 neuron output layer (as shown Fig. 4), based on the results shown in Fig. 7(a) and (c), the accuracy performance obtained is 100% when using the Resilient backpropagation and the Levenberg-Marquardt backpropagations. On the other hand, based on the results shown in Fig. 7(b), the accuracy performance obtained is 91.7% when using the Gradient Descent backpropagation, in which 1 female sample is incorrectly classified.

Table 2 summarizes all the findings obtained for three different learning algorithms for three different neural network topologies. Based on this summary, a neural network with topology having 2 hidden layers produced higher accuracy performance. It is also found that with 2 hidden layers having more number of neurons in the hidden layers (16 and 16 neurons in which hidden layer) was able to produce greater performance accuracy. The best learning algorithm that can be applied in gender detection shown in paper is the Gradient Descent learning algorithm. Also, it is notable that 8 out of 9 cases where all male samples are able to be detected or classified correctly compared to the 3 out of 9 cases in which all females are correctly detected or classified.

Table 2. Summary of accuracy performances obtained for three different learning algorithms for three different neural network topologies.

Learning algorithms	Performance accuracy (%)								
	16 neurons hidden layer			2 hidden layers (16 & 10 neurons)			2 hidden layers (16 & 16 neurons)		
	M	F	All	M	F	All	M	F	All
Resilient	83.3	83.3	83.3	100.0	66.7	83.3	100.0	100.0	100.0
Gradient Descent	100.0	83.3	91.7	100.0	100.0	100.0	100.0	83.3	91.7
Levenberg-Marquardt	100.0	66.7	83.3	100.0	83.3	91.7	100.0	100.0	100.0

5 Conclusion

In this paper, an investigation has been conducted to assess the performance of the artificial neural network (ANN) classifier in gender detection based on three different learning algorithms using three different neural network topologies. The best classification results are obtained when the Gradient Descent backpropagation and the worse classification results are obtained from Resilient backpropagation in some of the topologies. Higher number of neurons and higher number of hidden layers have caused higher performance accuracy. The results also described that the sweat collection based on the

16 VOCs are more beneficial to detect a male sample rather than detecting a female sample. More detailed experiments can be conducted by using several more settings. The best approach will be applying a genetic algorithm in order to find the best setting in gender detection or classification.

References

1. Sharma, B., Venugopalan, K.: Comparison of neural network training functions for hematoma classification in brain CT images. IOSR J. Comput. Eng. **16**, 31–35 (2014)
2. Nayak, P.K., Narayan, D.: Multimodal biometric face and fingerprint recognition using neural network. Int. J. Eng. Res. Technol. **2**, 313–321 (2012)
3. Kashyap, K., Yadav, M.: Fingerprint matching using neural network training. Int. J. Eng. Comput. Sci. **2**, 2041–2044 (2013)
4. Divya, B., Savetha, V.: Spoofing face and gender recognition classification using neural network with 3D masks. In: International Conference on Engineering Trends and Science & Humanities (2015)
5. Jaswante, A., Khan, A.U., Gour, B.: Gender classification technique based on facial features using neural network. Int. J. Comput. Sci. Inf. Technol. **4**, 839–843 (2013)
6. Shukla, R., Shukla, R., Shukla, A., Sharma, S., Tiwari, N.: Gender identification in human gait using neural network. Int. J. Modern Educ. Comput. Sci. **4**, 70 (2012)
7. Michelsanti, D., Guichi, Y., Ene, A.-D., Stef, R., Nasrollahi, K., Moeslund, T.B.: Fast fingerprint classification with deep neural network. In: International Conference on Computer Vision Theory and Applications, VISAPP (2017)
8. Marak, P., Hambalík, A.: Fingerprint recognition system using artificial neural network as feature extractor: design and performance evaluation. Mathematical Institute, Slovak Academy of Sciences (2016)
9. Shrivas, R., Verma, N., Singh, V.: Palm print biometrics using feed forward back propagation neural network. Int. J. Comput. Appl. **74**, 45–49 (2013)
10. Ramirez-Cortes, J.-M., Gomez-Gil, P., Alarcon-Aquino, V., Baez-Lopez, D., Enriquez-Caldera, R.: A biometric system based on neural networks and SVM using morphological feature extraction from hand-shape images. Vilnius Univ (2011)
11. Wang, R., Han, C., Wu, Y., Guo, T.: Fingerprint Classification Based on Depth Neural Network (2014)
12. Pati, S.R., Suralkar, S.R.: Fingerprint classification using artificial neural network. Int. J. Emerg. Technol. Adv. Eng. **2**, 513–517 (2012)
13. Narang, N., Bourlai, T.: Gender and ethnicity classification using deep learning in heterogeneous face recognition. IEEE (2016)
14. Kohail, S.N.: Using artificial neural network for human age estimation based on facial images. In: International Conference on Innovations in Information Technology (2012)
15. Kalansuriya, T.R., Dharmaratne, A.T.: Neural network based age and gender classification for facial images. Int. J. Adv. ICT Emerg. Regions **7**(2) (2014)
16. Sas, J., Sas, A.: Gender recognition using neural networks and ASR techniques. J. Med. Inform. Technol. **22**, 179–187 (2013)
17. Khanum, S., Sora, M.: Speech based gender identification using feed forward neural networks. Int. J. Comput. Appl. (0975–8887). National Conference on Recent Trends in Information Technology (NCIT 2015) (2015)

18. Roy, S., Bandyopadhyay, S.K.: Gender recognition using Self Organizing Map (SOM) - an unsupervised ANN approach. Int. J. Emerg. Res. Manag. Technol. **3**(8) (2010). ISSN: 2278 - 9359
19. Ramesha, K., Raja, K.B., Venugopal, K.R., Patnaik, L.M.: Feature extraction based face recognition, gender and age classification. Int. J. Comput. Sci. Eng. (IJCSE) 2(01S), 14–23 (2010)
20. Omatu, S.: Odor classification by neural networks. IEEE (2013)
21. Chansri, C., Srinonchat, J.: Personal shirt odor classification using an electronic nose. IEEE (2010)
22. Bassiouni, M., Khalefa, W., El-Dahshan, E.A., Salem, A.-B.M.: A machine learning technique for person identification using ECG signals. Int. J. Appl. Phys. **1**, 37–41 (2016)
23. Bascil, M.S., Temurtas, F.: A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm. J. Med. Syst. **35**, 433–436 (2011)
24. Rehman, M.Z., Nawi, N.M.: The effect of adaptive momentum in improving the accuracy of gradient descent back propagation algorithm on classification problems. Springer, Heidelberg (2011)
25. Rehman, M.Z., Nawi, N.M.: Studying the effect of adaptive momentum in improving the accuracy of gradient descent back propagation algorithm on classification problems. In: International Conference Mathematical and Computational Biology (2011)
26. Chen, C.-S., Su, S.-L.: Resilient back-propagation neural network for approximation 2-D GDOP. In: International Multi Conference of Engineering's and Computer Scientists (2010)
27. Nawi, N.M., Khan, A., Rehman, M.Z.: A new Levenberg Marquardt based back propagation algorithm trained with Cuckoo search. In: The 4th International Conference on Electrical Engineering and Informatics (2013)
28. Muaidi, H.: Levenberg-Marquardt learning neural network for part-of-speech tagging of Arabic sentences. WSEAS Trans. Comput. **13**, 300–309 (2014)
29. Lahmiri, S.: A comparative study of backpropagation algorithms in financial prediction. Int. J. Comput. Sci. Eng. Appl. **1**, 15–21 (2011)

Application of Social Media Among Medical Practitioner for Sharing Tacit Knowledge: A Pilot Study

Asra Amidi, Yusmadi Yah Jusoh(✉), Mar Yah Said, Marzanah A. Jabar, and Rusli Haji Abdullah

Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 Serdang, Malaysia
asra.amidi@gmail.com,
{yusmadi,maryah,marzanah,rusli}@upm.edu.my

Abstract. Tacit knowledge is perceived as the most strategically important resource of competitiveness. The rise of web-based applications such as social media also gives rise to the question on whether these applications can facilitate tacit knowledge sharing in a collaborative work environment. Previous studies have indicated that such notion is indeed possible, but there is still a lack of understanding of how social media could facilitate tacit knowledge sharing as well as the condition that is most effective in transferring this type of knowledge. Hence, it is crucial to understand the individual and technical characteristics involved in tacit knowledge sharing using social media. This research attempts to bridge this gap as there is a need to develop a holistic tacit knowledge sharing model. Towards this end, before the model is developed, the conceptual model and its instruments are validated by three field experts. A pilot study is conducted to determine the reliability and validity of the measurement indicators as well as an analysis using SPSS. The findings of the pilot study are hence presented in this paper. The results confirmed the validity of the proposed model as well as the validity and reliability of the instrument. This pilot study investigated on whether the proposed research model is viable for further research, or whether pertinent changes to the model or the methodology need to be done before the model can be used on a larger sample. Recommendations for a follow-up study concludes the paper.

Keywords: Tacit knowledge · Social media · Knowledge sharing

1 Introduction

The collaboration and sharing of information, particularly in the healthcare sector, has been made available through the use of social media networks which have emerged as a powerful tool towards this end. In the past six years, many researchers have studied explicit knowledge sharing over social media and attempted to discover whether social media can be utilized as a pioneering model for successful KM strategies and frameworks. Conclusions made so far have generally been positive. Several researches

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 338–346, 2018.
https://doi.org/10.1007/978-981-10-8276-4_32

concluded that organizational KM can be leveraged by social media which is naturally compatible with the process of tacit knowledge sharing [1, 2]. It becomes even more challenging when it comes to knowledge sharing in the healthcare sector as this sector is highly tacit knowledge-based [3] and makes up as much as 80% of the total vital knowledge [4]. The main challenges are to nurture the sharing of tacit knowledge and the codification of tacit knowledge. Prior research that focused exclusively on the contribution of social web to tacit knowledge sharing is minimal. The role of social web in tacit knowledge sharing is currently undetermined and many dimensions of tacit knowledge have not been examined [5].

This study thus complements previous researches and integrates two major aspects from different theories to propose a new research model in the field. Although there has been many discussions regarding the two mechanisms of band sharing which have been disseminated across various research areas, there is yet any study conducted on a model that explores both (socio-personal and technical) mechanisms. Therefore, this study is set to answer this research question: what are the technical and socio personal factors behind the intention of an individual to share tacit knowledge on social media? In this research, a model is proposed and tested based on the self-determination theory [6], social cognitive theory [7], and social capital theory [8] as well as the technology acceptance model [9] These theories provided reasonable justifications as to why and how people conduct knowledge sharing and interpersonal helping. This paper is organized as such: Sect. 2 provides a brief review of literature on knowledge sharing and social media as well as the research model proposed. Section 3 explains the research methodology used in this paper. Section 4 presents the results of content validity and results of the pilot study. Section 5 presents the findings of the study, and finally Sect. 6 presents the conclusion and future implications of the study.

2 Related Works

Various disciplines and scholars have conceptualized tacit knowledge differently, causing spirited debates with regards to these differing viewpoints. The transmission of TK (Tacit Knowledge) is crucial for organizations to ensure that TK is passed throughout the organization, rather than being retained by a single employee. Explicit knowledge can be transmitted orally or in written form, is usually impersonal and formal, and is often presented in the form of documents, reports, “white papers”, catalogues, presentations, patents, formulas, etc. [10]. On the other hand, tacit knowledge (expressed in the form of abilities, developed skills, experience, undocumented processes, “gut-feelings”, etc.) is extremely personal and difficult to be expressed in written form. An individual’s experience and values make up the basis of tacit knowledge [11]. Hence, the nature of tacit knowledge limits its ability to be transmitted [12, 13]. With the growing development of technology, particularly in information and communication technologies (ICT), several studies such as those by Alajmi [14], Li [15], Ma and Yuen [16], Tamjidyamcholo et al. [17], Zhao [18], Papadopoulos et al. [19], Razmerita et al. [20], as well as Tamjidyamcholo et al. [21] had distinguished individual and human factors that influence knowledge sharing in social media. These researchers only discussed issues relating to general knowledge

sharing without any specific emphasis on the motivational factors influencing tacit knowledge sharing. However, several technological factors have also been discovered by several other researchers such as Matschke et al. [22], Nielsen and Razmerita [23], Vuori and Okkonen [24], as well as Yuan et al. [25], which indicated the influence of these variables on knowledge sharing in social media platform. This group of researches offers insight into the affordance of specific factors that affect knowledge sharing in social media, but does not specifically seek to understand these affordances when social media tools are used to share tacit knowledge. The researches either focus merely on human or technical factors, whilst the effectiveness of those factors in disseminating tacit knowledge through social media platform was neglected.

In interpreting the results of the previous studies, it can be summarized that almost all the research models are similar in terms of how they explore knowledge sharing behavior, and how the different components were investigated. Knowledge sharing is facilitated by a variety of social and technical enablers that complement each other in shaping knowledge management efforts. This complementary interaction needs to take into consideration aspects such as sociability, usability, and the fit between social and technical factors. The focus of socio-technical studies is usually on the continuous interactions between IT and the people during the stages of designing, implementing and utilizing the IT systems. A holistic approach is adopted which can highlight the interplay of social and technical factors in the way people work. Hence, both factors need to be considered by organizations because dynamic business processes require the synergy of both factors; such synergy is important as it has been found to produce better performance. In a similar vein, this paper proposes a research model. This model does not delineate all the inherent enablers; instead, it highlights several key factors that justify most of the variance in knowledge sharing. Therefore, this model is more likely to aid in the investigation of the various enablers to decipher tacit knowledge sharing activities. This model theorizes that knowledge enablers influence the intention and behavior in tacit knowledge sharing. The basis of this analysis is established on the individual level as the understanding of knowledge sharing between individuals can result in a better understanding of knowledge sharing in an organizational level.

In general, all models examine the relationship between factors and knowledge sharing. Only a few studies had identified the type of knowledge in the process of knowledge sharing and the influential factors that may have an effect in an online environment. Therefore, many dimensions of tacit knowledge sharing as an effective factor in social web environments have yet been examined. This study focuses on developing a holistic model of knowledge sharing for the healthcare sector based on an integrated approach. This approach combines factors that influence knowledge sharing which includes the individual, organization and technology. Towards this end, the literature review in this study indicated two major requirements that are pertinent in tacit knowledge sharing namely technical factors and socio personal factors.

3 Methodology

The research methodology undertaken in this study is pilot study. The first phase involves the identification of the Major variables (independent, dependent, mediator and moderator variables). Subsequently, the questionnaire was conceptualized where the constructs were developed based on previous researches to ensure that the scales selected are reliable and valid. A five-point Likert scale was used for each item in the questionnaire where 1 represents "strongly disagree" and 5 represents "strongly agree".

The Second phase involves the validation of the proposed model and instrument by three experts in information systems and knowledge management. The assessment for content validity ensures that: (1) the items and the construct's theoretical domain are consistent; (2) the items represent the constructs that they are supposed to measure, and (3) the items are neither difficult, ambiguous, nor double-barreled statements. The experts were asked to review the model to rate the level of relevancy of the constructs to the hypothesis model. The rating for the construct validity ranges from 1 = not very relevant, 2 = somewhat relevant, 3 = quite relevant, 4 = very relevant, and 5 = finally asked experts to give any comments. Kappa analysis was performed to analyze the experts' validation. Content Validity Index (CVI) was employed to statistically analyze the content validity of the developed instrument, which involved three experts.

The Third phase involves reliability test. The consistency between research variables is measured using Cronbach's coefficient alpha. The random error in measurement is referred to as reliability, which indicates the accuracy of the measuring instrument as an addition to the validity of the instrument. A pilot test was performed on the instrument to ensure readability and understandability. This test was also conducted to determine the performance of data collection protocols under realistic settings [26]. Zikmund [27] noted that a satisfactory Cronbach's alpha value should be at least 0.70 to be able to determine the correlation between the items that are being measured. The Corrected Item-Total Correlation was also examined whereby items with Corrected Item-Total Correlation of less than 0.3 are removed from the scale. The pilot was analyzed using SPSS version 22. To ensure that the questionnaire does not contain any potential errors, a reliability test was then conducted on a sample size of 25 respondents. Table 1 shows the Cronbach's Alpha for the variables consisting of Sharing Tacit Knowledge (STK), Frequency of Social Media Usage (FSMU), Work Experience (WE), Commitment (COM), Trust (Trust), Knowledge-Sharing Self-Efficacy (KSSE), Motivation (MOT), Performance Expectation (Per), Perceived Enjoyment (PE), Perceived Usefulness (PU), Perceived Ease of Use (PEU), and Attitude Toward Social Media Usage (AT).

4 Results

4.1 Expert Review Results

The results of kappa analysis indicated that among all the instrument items, two items with a CVI score lower than 0.70 were eliminated and two items with a CVI of 0.67 were modified (modifications of items were performed based on the recommendation of the expert).

Table 1. Analysis of the reliability for the pilot study

Measure	Item	Num	A	B	C	D	E	Summary
Sharing tacit knowledge	STK 1	25	6.0000	1.000	.604	.673	.760	Acceptable
	STK 2	25	6.2800	1.460	.586	.721		Acceptable
	STK 3	25	6.3600	.990	.642	.620		Acceptable
Frequency social media usage	FSMU 1	25	6.8000	2.583	.758	.581	.794	Acceptable
	FSMU 2	25	6.6800	3.227	.605	.754		Acceptable
	FSMU 3	25	7.0800	3.077	.561	.802		Acceptable
Work experience	WE 1	25	5.760	2.107	.694	.560	.741	Acceptable
	WE 2	25	5.720	2.043	.596	.630		Acceptable
	WE 3	25	6.200	1.583	.493	.817		Acceptable
Commitment	Com 1	25	8.760	5.023	.668	.863	.869	Acceptable
	Com 2	25	8.920	5.160	.709	.839		Acceptable
	Com 3	25	9.000	5.667	.807	.808		Acceptable
	Com 4	25	9.080	5.660	.751	.824		Acceptable
Trust	Trust 1	25	11.920	5.827	-.083	.748	.748	Non-acceptable
	Trust 2	25	12.040	3.207	.526	.467		Consider to modify
	Trust 3	25	11.520	4.593	.373	.566		Acceptable
	Trust 4	25	11.840	3.473	.605	.425		Acceptable
	Trust 5	25	11.560	4.090	.532	.490		Acceptable
Knowledge sharing self- efficacy	KSSE 1	25	20.520	29.593	.623	.924	.925	Acceptable
	KSSE 2	25	20.720	29.460	.665	.922		Acceptable
	KSSE 3	25	21.120	26.527	.727	.917		Acceptable
	KSSE 4	25	20.720	25.543	.890	.903		Acceptable
	KSSE 5	25	20.880	25.527	.884	.903		Acceptable
	KSSE 6	25	21.120	26.193	.805	.910		Acceptable
	KSSE 7	25	21.160	26.557	.755	.915		Acceptable
	KSSE 8	25	20.640	28.657	.627	.924		Acceptable
Motivation	MOT 1	25	16.680	5.727	.401	.707	.726	Acceptable
	MOT 2	25	16.880	4.943	.662	.621		Acceptable
	MOT 3	25	16.520	4.927	.553	.659		Acceptable
	MOT 4	25	16.400	5.917	.451	.692		Acceptable
	MOT 5	25	16.560	6.507	.225	.752		Acceptable
	MOT 6	25	16.360	6.157	.537	.679		Acceptable
Performance expectation	PER 1	25	10.120	4.193	.708	.692	.795	Acceptable
	PER 2	25	9.920	4.243	.742	.678		Acceptable
	PER 3	25	10.000	4.000	.765	.660		Acceptable
	PER 4	25	10.400	5.417	.280	.898		Acceptable
Perceived enjoyment	PE 1	25	10.320	5.810	.872	.930	.946	Acceptable
	PE 2	25	10.320	5.310	.893	.924		Acceptable
	PE 3	25	10.280	5.543	.874	.929		Acceptable
	PE 4	25	10.000	5.833	.852	.936		Acceptable

(*continued*)

Table 1. (*continued*)

Measure	Item	Num	A	B	C	D	E	Summary
Perceived usefulness	PU 1	25	14.600	6.667	.581	.801	.819	Acceptable
	PU 2	25	14.800	5.333	.779	.734		Acceptable
	PU 3	25	14.840	5.390	.817	.726		Acceptable
	PU 4	25	14.880	4.777	.787	.725		Acceptable
	PU 5	25	15.440	6.507	.265	.899		Acceptable
Perceived ease of use	PEU 1	25	9.360	4.157	.773	.705	.816	Acceptable
	PEU 2	25	9.320	4.310	.711	.734		Acceptable
	PEU 3	25	9.800	4.833	.670	.762		Acceptable
	PEU 4	25	9.920	4.410	.465	.871		Acceptable
Attitude toward social media usage	AT 1	25	28.920	34.493	.795	.876	.903	Acceptable
	AT 2	25	29.720	32.877	.623	.918		Acceptable
	AT 3	25	29.200	32.917	.826	.867		Acceptable
	AT 4	25	28.480	33.260	.769	.879		Acceptable
	AT 5	25	28.800	32.667	.822	.868		Acceptable

Note: A: Scale mean if item deleted, B: Scale variance if item deleted, C: Correlated item-Total correlation, D: Cronbach's alpha if item deleted, E: Cronbach's alpha

4.2 Reliability Test Result

The internal consistency of the proposed constructs was assessed using Cronbach's alpha. Consequently, only four measurement items were used for this construct after measurement item Trust 4 was omitted. Measurement item Trust 1 was considered for modification. The loading ranges and value for each construct are summarized in Table 2, where the ranges are between 726 and .946. The reliability of each question item was measured using item-total correlation. Any value that is less than 0.3 indicates that, as a whole, the item is measuring something different from the scale. The item total correlations were found to be within the acceptable threshold values. Table 1 shows the analysis of the reliability for pilot study.

Based on the result of the pilot study, each of the Cronbach's alpha is higher than 0.7, which shows that the internal consistency of each scale is considered acceptable as shown in Table 1 with only a few minor corrections. Table 2 Summarized the Cronbach's alpha measure for all constructs.

5 Discussions and Implications

This study examines an integrated model of tacit knowledge sharing in social media based on ideas drawn from knowledge management and information system theories. The proposed model was validated by three experts in the field after which an instrument that is valid and reliable was tested for all the adapted variables. Based on statistical findings, the validity and reliability tests had produced results with sufficient confidence. There were no significant issues encountered in this study concerning

content validity as the survey instruments were primarily adopted from extensive literature and validated by experts and the pilot study. All the Cronbach alpha coefficients were well above the recommended 0.70 cutoff, which shows that the internal consistency of each scale is considered acceptable. Two broad factors, namely Socio Personal and Technical, were examined to assess their impact on tacit knowledge sharing. Effective tacit knowledge sharing has been persistently problematic, but the results of this study have interestingly suggested that social media is able to promote effective tacit knowledge sharing. Although the current work is too preliminary for actionable conclusions, the proposed model provides a deeper understanding of the influencing factors to this phenomenon, as well as highlighting the social media support given to the user via interaction in the healthcare domain. These results shed light on at least two theoretical issues worth pursuing. Firstly, the effectiveness of the combination of information technologies and personal factors may be higher than the use of each type individually. Secondly, the findings are potentially useful for organizations and health departments that are responsible for providing relevant and accessible information (via social media platforms).

Table 2. Summary of Cronbach's alpha for measurement in the pilot study.

Construct	Cronbach alpha	Summary
Sharing tacit knowledge	.760	Acceptable
Frequency social media usage	.794	Acceptable
Work experience	.741	Acceptable
Commitment	.869	Very good internal consistency
Trust	.748	Good but it need modification
Knowledge sharing self-efficacy	.925	Very good internal consistency
Motivation	.726	Acceptable
Performance expectation	.795	Acceptable
Perceived enjoyment	.946	Very good internal consistency
Perceived usefulness	.819	Very good internal consistency
Perceived ease of use	.816	Very good internal consistency

6 Conclusion and Future Work

The findings of this pilot study are exploratory in nature; thus, they are not intended to be used as a generalization for a larger population. Further research employing a larger and more diverse sample would be needed to firmly establish the categories and to construct a theoretical model. This research has the potential to inform organizations of their social media platforms so that they can provide relevant and accessible and health departments in the development. This study examines the appropriation of social media platforms in assessing the sharing of tacit knowledge and in examining how technologies contribute to this process as enablers. This ongoing research aims to put in place a holistic model that encloses the appropriation of socio-technical constructs that may influence tacit knowledge sharing among medical practitioners. In validating the

research model, future studies may further examine the various variables by using the Structured Equation Modeling (SEM), which provides the model fit and enables the simultaneous running of the study's tested links.

Acknowledgements. The authors gratefully acknowledge the financial assistance from the Fundamental Research Grant Scheme, Project Code: 08-02-13-1368FR, Ministry of Higher Education.

References

1. Amidi, A., et al.: An overview on leveraging social media technology for uncovering tacit knowledge sharing in an organizational context. In: 2015 9th Malaysian Software Engineering Conference (MySEC). IEEE (2015)
2. Levy, M.: WEB 2.0 implications on knowledge management. J. Knowl. Manag. **13**(1), 120–134 (2009)
3. Panahi, S., Watson, J., Partridge, H.: Towards tacit knowledge sharing over social web tools. J. Knowl. Manag. **17**(3), 379–397 (2013)
4. Callahan, S.: Want to manage tacit knowledge? Communities of practice offer a versatile solution. White Paper Anecdote (2006). http://www.anecdote.com.au/papers/Want_to_manage_tacit_knowledge.pdf
5. Panahi, S., Watson, J., Partridge, H.: Information encountering on social media and tacit knowledge sharing. J. Inf. Sci. **42**(4), 539–550 (2015)
6. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. **55**(1), 68 (2000)
7. Bandura, A.: Human agency in social cognitive theory. Am. Psychol. **44**(9), 1175 (1989)
8. Nahapiet, J., Ghoshal, S.: Social capital, intellectual capital, and the organizational advantage. Acad. Manag. Rev. **23**(2), 242–266 (1998)
9. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**, 319–340 (1989)
10. Nonaka, I.: A dynamic theory of organizational knowledge creation. Organ. Sci. **5**(1), 14–37 (1994)
11. Nonaka, I., Von Krogh, G.: Perspective-tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. Organ. Sci. **20**(3), 635–652 (2009)
12. Lucas, L.M.: The impact of trust and reputation on the transfer of best practices. J. Knowl. Manag. **9**(4), 87–101 (2005)
13. Nezakati, H., et al.: Review of social media potential on knowledge sharing and collaboration in tourism industry. Proc.-Soc. Behav. Sci. **172**, 120–125 (2015)
14. Alajmi, B.M.: The intention to share: professionals' knowledge sharing behaviors in online communities. Rutgers University-Graduate School-New Brunswick (2011)
15. Li, W.: Virtual knowledge sharing in a cross-cultural context. J. Knowl. Manag. **14**(1), 38–50 (2010)
16. Ma, W.W., Yuen, A.H.: Understanding online knowledge sharing: an interpersonal relationship perspective. Comput. Educ. **56**(1), 210–219 (2011)
17. Tamjidyamcholo, A., Baba, M.S.B., Shuib, N.L.M., Rohani, V.A.: Evaluation model for knowledge sharing in information security professional virtual community. Comput. Secur. **43**, 19–34 (2014)

18. Zhao, L.: Sharing knowledge in virtual communities: Factors affecting a member's intention to share (2010)
19. Papadopoulos, T., Stamati, T., Nopparuch, P.: Exploring the determinants of knowledge sharing via employee weblogs. Int. J. Inf. Manag. **33**(1), 133–146 (2013)
20. Razmerita, L., Phillips-Wren, G., Jain, L.C.: Advances in knowledge management: an overview. Innovations in Knowledge Management, pp. 3–18. Springer (2016)
21. Tamjidyamcholo, A., Baba, M.S.B., Tamjid, H., Gholipour, R.: Information security–professional perceptions of knowledge-sharing intention under self-efficacy, trust, reciprocity, and shared-language. Comput. Educ. **68**, 223–232 (2013)
22. Matschke, C., Moskaliuk, J., Bokhorst, F., Schümmer, T., Cress, U.: Motivational factors of information exchange in social information spaces. Comput. Hum. Behav. **36**, 549–558 (2014)
23. Nielsen, P., Razmerita, L.: Motivation and knowledge sharing through social media within danish organizations creating value for all through IT, pp. 197–213. Springer
24. Vuori, V., Okkonen, J.: Knowledge sharing motivational factors of using an intra-organizational social media platform. J. Knowl. Manag. **16**(4), 592–603 (2012)
25. Yuan, D., Lin, Z., Zhuo, R.: What drives consumer knowledge sharing in online travel communities? personal attributes or e-service factors? Comput. Hum. Behav. **63**, 68–74 (2016)
26. Fowler, F.J.: Survey Research Methods. Sage publications, Thousand Oaks (2013)
27. Zikmund, W., Babin, B., Carr, J., Griffin, M.: Business Research Methods. Cengage Learning, Mason (2012)

Lost in Time: Temporal Analytics for Long-Term Video Surveillance

Huai-Qian Khor(✉) and John See

Faculty of Computing and Informatics, Center for Visual Computing, Multimedia University, 63100 Cyberjaya, Malaysia
hqkhor95@gmail.com, johnsee@mmu.edu.my

Abstract. Video surveillance is a well researched area of study with substantial work done in the aspects of object detection, tracking and behavior analysis. With the abundance of video data captured over a long period of time, we can understand patterns in human behavior and scene dynamics through data-driven *temporal analytics*. In this work, we propose two schemes to perform descriptive and predictive analytics on long-term video surveillance data. We generate heatmap and footmap visualizations to describe spatially pooled trajectory patterns with respect to time and location. We also present two approaches for anomaly prediction at the day-level granularity: a trajectory-based statistical approach, and a time-series based approach. Experimentation with one year data from a single camera demonstrates the ability to uncover interesting insights about the scene and to predict anomalies reasonably well.

1 Introduction

In the domain of video surveillance, there has been a significant amount of research done in the past few decades relating to a variety of sub-tasks such as object detection and tracking [1], and behavior analysis [2]. The abundance of video data in the "Big Data" era has resulted in far more data collected than analysed or processed [3]. Across a long period of time, *temporal analytics* can offer interesting data-driven insights into a variety of contemporary problems such as retail location analysis, and understanding of commuting behaviors and crowd patterns.

The Long-term Observation of Scenes with Tracks (LOST) dataset [4] is the only known dataset established for the purpose of studying scene behavior and changes at a longer time scale. Its data consists of videos captured from a number of outdoor streaming cameras located in different parts of the world, over a period of 1–3 years. Each video contains the same half an hour period captured each day. The dataset comes with rich metadata (i.e. blob, trajectory) that can be readily used for further analysis of long term trends in the scene.

The analysis of long term trends has been much studied in a wide variety of fields such as climatology [5] and epidemiology [6]. Within the engineering and computing domains, long term trend analysis has also been investigated in

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 347–357, 2018.
https://doi.org/10.1007/978-981-10-8276-4_33

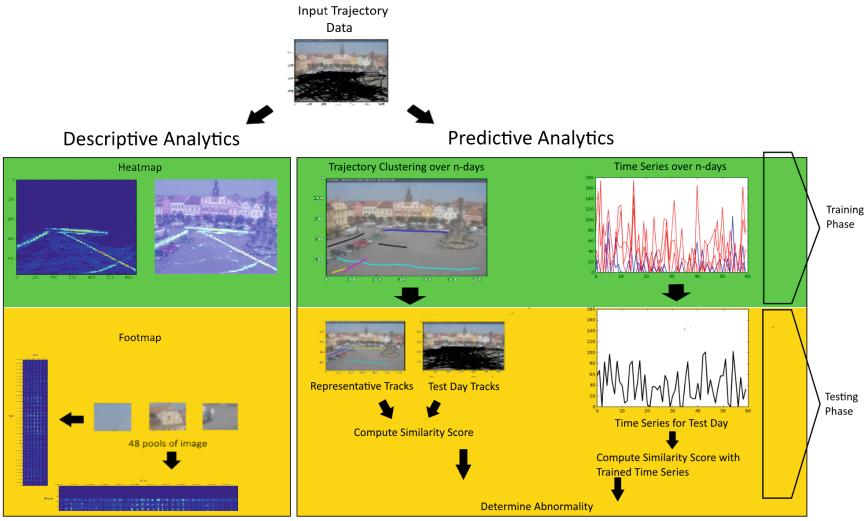


Fig. 1. Proposed temporal analytics framework for long-term video surveillance

the area of time series forecasting [7] and social media analytics [8]. While video surveillance research has progressed tremendously in many aspects, most of the data considered are short term in nature. Moreover, data analytics applied to long-term surveillance data could potentially derive a deeper understanding into changes that occur at a longer time scale (months to years). A few recent works have begun exploring the usefulness of long-term video data for anomaly mining and prediction [9–11].

In this paper, we describe the notion of *temporal analytics* and how it is carried out to extract valuable insights from long-term video surveillance data. The main contribution of this work is to propose feasible techniques for descriptive and predictive analytics on video surveillance data that spans a long period of time. Firstly, we performed a descriptive extraction of trajectory patterns from a monitored scene to generate *heatmap* and *footmap*, which can capture time- and location-based trends pooled from the accumulated trajectories. Secondly, we proposed two approaches for predicting anomalies at the *day-level granularity*: a trajectory-based statistical approach that calculates statistical difference with distance between trajectories and a time series-based classification approach that computes statistical difference in terms of number of trajectories per day. The statistical approach is motivated by the work in [12] which used trajectory-based information to predict abnormal trajectories, while time-series based approach allows daily trajectory information to be represented temporally. We report the insights obtained from these proposed schemes by experimenting with one-year data from a single camera of the LOST dataset.

2 Related Work

Long-term video surveillance research is still in a nascent stage. Recently, there has been an increase of interest in investigating how common behavioral patterns can be mined and how anomalies can be predicted across longer time spans.

In the original work that proposed the LOST dataset, Abrams et al. [4] demonstrated a few example cases of how trajectories of moving objects (or "tracks") can be clustered and aggregated into statistics that effectively captures the long-term trends in track variation over the scene. They also showed how these statistics also correlated to external signals such as day of the week and weather condition.

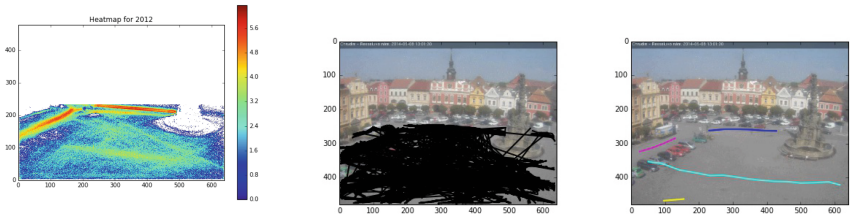
Following that, a work by See and Tan [10] proposed a time-scale based framework for mining anomalous track patterns on selected cameras from the LOST dataset. Their method first clusters unlabeled tracks using two temporal levels to find common modes of behavior, represented by track exemplars. Then, a probabilistic anomaly prediction algorithm was devised to evaluate the abnormality of new tracks. Due to the lack of ground truth labels, prediction of abnormal tracks was performed using synthetically generated anomalous tracks. A subsequent work [12] performed object classification on objects extracted from LOST videos for more than 23,000 frames under a variety of weather conditions.

Zen et al. [9] proposed a pixel-wise approach to determine the density of traffic captured for a whole month in New York City. The segmented foreground regions of moving objects were extracted to compute traffic density. Typical patterns and anomalous events were discovered by an anomaly score, which defines the distance between the traffic densities taken at a specific day of the week and time of the day. The vulnerability of this method is that the presence of noise as foreground pixels can produce inaccurate traffic patterns.

3 Data Preparation

In this work, we use the object trajectory data from camera number '001' (Ressel Square, Chrudim, Czech Republic) of the LOST dataset [4], spanning 228 days between 01-01-2012 to 31-12-2012. In this period of 1 year, a portion of videos had a lower frame rate than most of the other videos while some had erroneous trajectory data. These videos were omitted as they were found to be unsuitable for our analysis. The trajectory data contains essential information of the moving objects in the scene, i.e. track ID and object centroid coordinates and dimensions.

One pressing issue is the lack of annotations in the original dataset [4], which is essential for validating the predictive analytics task. Hence, we sought the help of three annotators to manually annotate the selected videos with a binary anomaly label (i.e.'1' corresponds to an anomalous day while '0' corresponds to a normal day). The annotators are asked to provide labels independently (no knowledge of labels given by the others), and the final annotated label is decided on the basis of consensus where two or more annotators agree to the same annotation. The criteria for a day to be considered as an anomaly are:



(a) (b) (c)

Fig. 2. (a) Heatmap constructed from a period of one year (January–December 2012) for camera 001, (b) All tracks accumulated over a span of three similar days, (c) Representative tracks of clusters, each shown with a distinct color.

(1) The occurrence of ad-hoc events at the plaza, (2) The occurrence of vehicles driving through the plaza which is only meant for pedestrians only.

4 Framework and Methods

In this section, we present our proposed temporal analytics framework, and the methods associated to the two schemes – descriptive and predictive analytics. Figure 1 shows a graphical illustration of the proposed framework, outlining how trajectory data can be visualized and used for the purpose of day-granularity anomaly prediction. The purpose of selecting day-level granularity is to experiment with a coarser granularity that spans a longer period of time.

4.1 Trajectory Data

The initial data contains trajectories (or tracks, in short), which are each represented by a set of points in 2D space. Each track T is denoted as

$$t_i = \{x_i, y_i\} \qquad \forall i \in 1...|T| \tag{1}$$

where x and y are the track coordinates at frame i.

4.2 Descriptive Analytics

The aim of performing *descriptive analytics* [13] on long-term surveillance data is to derive useful knowledge from historical data that can potentially be used for further analysis. In our work, descriptive analytics can be performed to extract higher level information such as the busiest period in the year or busiest day in a usual week, or paths that most objects will pass through. The outcome of this module is two-fold – we utilize *heatmaps* to discover the paths that are most commonly taken by moving objects. Besides, we also introduce a new visualization called *footmap* that is able to summarize the intensity of activity in the monitored scene based on time and location.

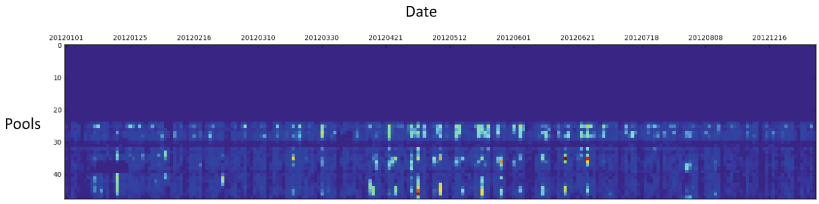


Fig. 3. Horizontal Footmap (HF)

Heatmaps [14] are constructed by accumulating the track coordinates $t(x, y)$ "traveled" by the moving objects,

$$\mathcal{H} = \sum_{k}^{|K|} \sum_{i}^{|T_k|} \delta_{t_i} \quad \forall t = t_i \tag{2}$$

for K number of tracks per day, over a period of N days.

Since the traveled locations are largely centered on a small number of prominent paths, we used a logarithmic scaled heatmap to provide a better balance in the color intensity distribution in the scene. Figure 2a shows the generated heatmap of moving objects represented by an array of different colors. Blueish regions represent lesser activities while the reddish regions represent higher level of activities. In this 1-year heatmap of camera '001' taken from the LOST dataset, we observe that the two main thoroughfares are distinctly marked in red, while two pedestrian pathways at the plaza are in noticeable green streaks.

To extract information on the intensity of activity based on time and location, we create a new type of visualization called a *footmap* for the monitored scene. Footmap can be used to capture time- and location-based trends pooled from patches of accumulated trajectories. *Footmap* is constructed by first accumulating all trajectory points t within each 80×80 non-overlapping square patch, also called a *pool*, by sum operation. To ensure all patches in a scene are of same size, we select the patch size based on two criteria: (1) it must be a common divisor of both the width and height of the scene; (2) it is of a reasonable size to allow the difference in the intensity of activity to be clearly depicted without being too homogeneous (patches too small) or too coarse (patches too large).

We name the footmap as a Horizontal Footmap (HF) (in Fig. 3). The HF is generated by spatially pooling all pixels in each patch before rearranging the output values vertically, taking from the original scene in left-to-right row-wise fashion. This causes the patches from the top parts of the scene to be located near the top of the footmap. This is repeated for all days in the period of analysis; we use a period of 1 year from January to December 2012.

The intuition behind the use of footmap is the visualization of activities from the perspective of both duration and location. Using a typical jet colormap (red for high intensity, blue for low intensity), we observe that the footmap demonstrates the intensity of activities across different times of the year (i.e. date axis), and also across different locations in scene (i.e. pool numbers).

The spring/summer season, which is near the middle zone, is noticeably busier particularly around April to June. In terms of location, the road area is visibly busy on most days of the year (middle rows of the HF) while there are obviously no activities detected in the top portion of the scene that corresponds to the buildings and sky (top half rows of the HF).

4.3 Predictive Analytics

We propose two strategies for *predictive analytics* on long-term surveillance data: the first being trajectory based while the second is time-series based.

Trajectory Based Prediction. In this method, we first detect for anomalies at a finer *trajectory-level* granularity. This is then used to make an inference at the *day-level* granularity, whether a particular day is likely to be abnormal, relative to what usually happens on that same day of the week (Mondays, Tuesdays, ..., Sundays). Hence, this assumes that each day exhibits a consistent pattern of activities throughout the year.

First step involves clustering the trajectories T to obtain common modes of movements in the scene. We utilize a trajectory-specific clustering algorithm, TraClus [15] to perform clustering on all tracks in the scene. The clustering is done on a time-scale of $(\omega, \epsilon) = (28, 7)$ as defined in [10]; a time-scale defines a temporal window that spans a specific number of days (ω) at a specific stride (ϵ) or periodicity. This can be interpreted as taking 3 previous similar days (e.g. Mondays) as training days for cluster generation while the 4th similar day is used as the test day for anomaly prediction. Figure 2b shows all tracks from the three training days, which are then grouped by TraClus algorithm into four clusters, each represented by its representative track (centroid of cluster) in Fig. 2c.

Motivated by previous works [10,16] that apply a statistical approach for probabilistic anomaly prediction, we utilize a similar concept to that, extending it further to day granularity anomalies. We formulate the distance metric,

$$D(P, Q) = \frac{1}{|P|} \sum_{t_p \in P} \min_{t_q \in Q} |t_p - t_q|^2 \tag{3}$$

as a random variable that measures the distance between new track and clusters.

Given the computed distance between a test track T' and the j-th cluster containing the representative track, we approximate the likelihood probability,

$$P(T'|X_j) = e^{-n_j D(T', X_j)} \tag{4}$$

The parameter η_j can be computed based on maximal likelihood evaluation, which is the reciprocal of mean distances learnt from the distribution of K training tracks:

$$\eta_j = \frac{K}{\sum_{k=1}^{K} D(T_k, X_j)} \tag{5}$$

For ease of computation, we compute the thresholds $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_J\}$ for all J clusters during the training process. Each threshold γ_j is calculated by taking the minimum likelihood of all training tracks $T_{j,k}$ belonging to the j-th cluster,

$$\gamma_j = \min_k P(T_{j,k}|X_j) \tag{6}$$

where a test track T' is classified as an anomalous track if the likelihood of the test track given cluster X_j,

$$j^* = \arg\max_j P(T'|X_j) \tag{7}$$

is less than its respective threshold γ_j^* and the distance between the test track and the nearest cluster $D(T'|X_{j^*}) > \delta$, an empirically defined distance threshold; we use $\delta = 1000$.

To predict a day-level anomaly on a test day A, we compute the ratio of number of anomalous tracks over the total number of tracks on that day, $\psi'_T = N_{ano}/N_{total}$. If ψ_T is more than a empirically defined anomaly threshold λ, then the test day will be predicted as anomalous.

$$A = \begin{cases} 1 & \text{if } \psi_T < \lambda \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Time Series Based Prediction. The second strategy performs anomaly prediction using time series trajectory data. Data can be summarized for each day, by counting the unique number of active trajectories (i.e. moving objects) within a defined interval θ seconds. For instance, a 30-min video with interval $\theta = 15\,\text{s}$ would produce a time series of 120 values. We denote the time-series count data as,

$$C(s) = \sum_{s=0}^{S} [O(s)] \tag{9}$$

where $S = D * 60/\theta$ and $O(s)$ represent the unique trajectories occurring at time interval $[\theta s, \theta(s+1)]$.

To measure the similarity between two time series, we opt for dynamic time warping (DTW) which seeks to find the optimal non-linear alignment. To speed things up, we use the Keogh lower bound (LB) variant of DTW (famously known as 'LB_Keogh') [17] which computes in linear time. With this quick method to determine the similarity between two time series, we predict an anomaly at the day-level using the classic k-Nearest Neighbor algorithm with LB_Keogh as distance measure. For every time series in the test set, a search is performed through all points in the training set to match with the most similar time series in the training set. We applied a 50:50 training-test split on a total of 212 days that contain valid time series data; the first half data is used to predict the second half data.

(a) (b)

Fig. 4. (a) The X-marks denote suitable places for marketplace stalls or advertisements (b) Two patches that have the highest amount of activity throughout the year, i.e. the 27th and 28th pool of the HF (c.f. Fig. 3).

5 Experimental Results and Discussion

In this section, we discuss the outcome of the two schemes performed on long-term video trajectory data.

5.1 Descriptive Analysis

In a year's worth of data, we uncover certain patterns through our descriptive analytics of heatmap and footmap.

The first analysis that we present is the outcome of heatmap. By overlaying the heatmap on top of the scene as in Fig. 4a, we can see that a reasonable amount of moving objects, possibly pedestrians and crowds, occurred at the plaza area. Hence, future developments such as marketplace stalls or advertisement billboards can be strategically placed at potential locations, adjacent to the paths that are mostly used by pedestrians (shown with X marks in Fig. 4a).

The second analysis is done based on the footmap in Fig. 3. From the Horizontal Footmap, we observe the difference in color intensities around April–June. This is indicative of an increase in activities due to more movements, hence retail activities can be suggested to increase during this busy period.

5.2 Anomaly Prediction

One of the main issues that we faced in measuring the performance of predicting anomalies is the imbalanced number of typical (negative) and anomalous (positive) days. As such, the standard accuracy metric is less suitable, and may not reflect the actual performance. Hence, we obtain the full confusion matrix, which allows us to determine the Precision, Recall and F1-score measures.

For trajectory-based prediction, we report the best F1-score of 0.46 based on the threshold $\lambda = 0.01$. Meanwhile, the time series-based method appears to perform much better at predicting anomalies, achieving the best F1-score of 0.67 with DTW window size of 2 ($k = 1$ therefore 1-NN). This demonstrates the robustness of using the time-series data for predicting anomalies.

Table 1. Confusion matrix for trajectory based prediction

Predicted \ Desired	Anomalous	Typical
Anomalous	50	110
Typical	7	46

Table 2. Confusion matrix for time series based prediction

Predicted \ Desired	Anomalous	Typical
Anomalous	13	23
Typical	10	60

Table 3. Performance results of various metrics for anomaly prediction

Metric	Trajectory	Time-series
Precision	0.31	0.67
Recall	0.88	0.69
F1-Score	**0.46**	**0.67**

Tables 1 and 2 are confusion matrices which display the number of predicted anomalies for trajectory-based and time-series-based methods respectively. The metrics used in Table 3 are described as follows: Precision is the number of correct results divided by the total number of returned results, Recall is the number of correct results divided by the number of results that should have been returned, F1-score is the harmonic mean of precision and recall which often used for class-imbalanced data.

5.3 Discussion

To the best of our knowledge, there are no prior works related to anomaly prediction on a day-level granularity for this dataset due to the lack of ground truth labels. Hence, it is not possible to compare directly in terms of standard performance measures. Prior works [10,18] that focus on trajectory-level prediction showed some promising results but no ground truth labels were available for validation. Interestingly, using this coarser granularity, we show that the visualization of traffic patterns can yield a distinction between high and normal traffic flow from both the temporal (Fig. 3) and spatial (Fig. 2a) perspectives.

6 Conclusion

In this work, we present a framework for performing temporal analytics for long-term video surveillance; it consists of a descriptive extraction of trajectory patterns to generate useful visualizations, and two predictive schemes for identifying

anomalies at the day-level granularity. This is a preliminary attempt at proposing descriptive and predictive analytics on long-spanning temporal information from surveillance videos. There is still plenty of room for improvements on the techniques proposed, and how object trajectories can be better represented with additional directional information. We hypothesize that temporal analytics on long-term video data will have far-reaching benefits for various domains such as urban planning, market strategy for businesses, and public security.

References

1. Zhang, S., Wang, C., Chan, S.C., Wei, X., Ho, C.H.: New object detection, tracking, and recognition approaches for video surveillance over camera network. IEEE Sens. **15**, 2679–2691 (2015)
2. Morris, B.T., Trivedi, M.M.: Understanding vehicular traffic behavior from video: a survey of unsupervised approaches. J. Electron. Imaging **22**, 041113–041113 (2013)
3. Porikli, F., Brémond, F., Dockstader, S.L., Ferryman, J., Hoogs, A., Lovell, B.C., Pankanti, S., Rinner, B., Tu, P., Venetianer, P.L.: Video surveillance: past, present, and now the future [DSP forum]. IEEE Signal Proc. Mag. **30**, 190–198 (2013)
4. Abrams, A., Tucek, J., Little, J., Jacobs, N., Pless, R.: Lost: Longterm observations of scenes (with tracks). In: WACV, pp. 297–304. IEEE (2012)
5. Collins, M., Knutti, R., Arblaster, J., Dufresne, J.L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W., Johns, T., et al.: Long-term climate change: projections, commitments and irreversibility (2013)
6. Velagaleti, R.S., Pencina, M.J., Murabito, J.M., Wang, T.J.: Long-term trends in the incidence of heart failure after myocardial infarction. Circulation **118**, 2057–2062 (2008)
7. Simon, G., Lendasse, A., Cottrell, M., Fort, J.C., Verleysen, M.: Time series forecasting: obtaining long term trends with self-organizing maps. Pattern Recognit. Lett. **26**, 1795–1808 (2005)
8. Vallis, O., Hochenbaum, J., Kejariwal, A.: A novel technique for long-term anomaly detection in the cloud. In: HotCloud (2014)
9. Zen, G., Krumm, J., Sebe, N., Horvitz, E., Kapoor, A.: Nobody likes mondays: foreground detection and behavioral patterns analysis in complex urban scenes. In: ARTEMIS 2013, pp. 297–304 (2013)
10. See, J., Tan, S.: Lost world: looking for anomalous tracks in long-term surveillance videos. In: Proceedings of IVCNZ, pp. 224–229 (2014)
11. Hu, S., Gurram, P., Chan, A.L.: Detection of anomalous track patterns for long term surveillance. In: SPIE Defense+ Security, pp. 946404–946404. International Society for Optics and Photonics (2015)
12. Saemi, M.M., See, J., Tan, S.: Lost and found: identifying objects in long-term surveillance videos. In: IEEE ICSIPA, pp. 99–104 (2015)
13. Kaisler, S., Armour, F., Espinosa, J.A., Money, W.: Big data: issues and challenges moving forward. In: Proceedings of HICSS, pp. 995–1004 (2013)
14. Fisher, D.: Hotmap: looking at geographic attention. IEEE Trans. Vis. Comput. Graph. **13**, 1184–1191 (2007)
15. Lee, J.G., Han, J., Whang, K.Y.: Trajectory clustering: a partition-and-group framework. In: Proceedings of ACM SIGMOD ICMD, pp. 593–604 (2007)

16. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE PAMI **28**, 1450–1464 (2006)
17. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**, 358–386 (2005)
18. Zhou, S., Shen, W., Zeng, D., Zhang, Z.: Unusual event detection in crowded scenes by trajectory analysis. In: ICASSP (2015)

Synergy in Facial Recognition Extraction Methods and Recognition Algorithms

Rayner Pailus Henry(✉) and Rayner Alfred(✉)

Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
rayner.pailus@gmail.com, ralfred@ums.edu.my

Abstract. This paper aims to survey on the existing research works done on facial recognition and acknowledge their differences. Understanding facial recognition processes such as facial normalization, facial detection, facial extraction and facial recognition methods and algorithms are part of the essence of this paper. This paper outlines the purposes of existing techniques as well as its challenges. This paper also looks at the idea whether combining several techniques is feasible in order to produce a better synergy result. Methods are evaluated based on their classification rate percentages as well as the numbers of dimensionality reductions. Based on the literature reviews conducted, the facial recognition algorithm is made up of two steps. The first step is when an individual model is modeled in the database based on the color appearance and geometrical information provided by the available images whereby each model characterizes an individual like a bar code or a unique serial number and discriminates it from the other people in the database. The second step is to carry out the identification using a classifier, related with the standard Gaussian distribution, to decide whether a face image belongs to one person in the database or not. This paper has performed a comparative analysis of previously conducted experiments and based on the findings obtained, a schema of the framework for face recognition is proposed.

Keywords: Facial normalization · Facial detection · Facial extraction
Facial recognition

1 Introduction

Studies on facial recognition had started for the past 30 years. It is to be known as programmed software which capable of recognizing and verifying a facial image either from a digital image or several digital images or from a video source. Despite of receiving several challenges in its real world application, especially in the field of commercial, banking, social and law enforcement realms these past few years, it has been actively researched involved with several well known processes such as image normalization process, pattern recognition process, neural networks and computer vision processes. Studies are crucial in facial recognition areas to overcome its key problems such as the illumination problem (lighting and background), face orientation and expression and the pose problem.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 358–369, 2018.
https://doi.org/10.1007/978-981-10-8276-4_34

Biometric method is replacing the traditional authentication methods of an individual that includes PINs numbers, plastic cards, keys smart cards, passwords, tokens which can be exposed or hacked, stolen, misplaced, duplicated. Other biometric methods had been implemented such as fingerprints biometric. However, it cannot be rendered or used if the epidermis tissue is damaged due to bruises or crack due to injuries or skin disease infections or worn out due to aging factors. As for iris and retina biometric, it requires expensive equipment and it is an intrusive process whereby this recognition process has to extract rays that penetrates one's cornea to capture the pattern of one's retina. This process will reduce the layer of one's cornea thickness and is believed to harm user's cornea and retina in a long run. If a person's eyes are disorders (56 types) or infected with eye diseases (12 common ones) that commonly are contagious, iris and retina biometric may cause an epidemic if it is used by publics. Roizenblatt has conducted several researches and looked into the outcome of cataract surgery and discovered that the surgical intrusion for cataract may result ones eye difficult to be identifiable by the iris recognition system [1]. Thus, unlike other biometric, face recognition that performs contactless process is a non- health risk biometric recognition. Facial recognition has numerous application areas that cover various important aspects in various sectors of industries that include surveillance security, individual's identity verification, criminal investigation system, smart card biometrics application, information security and access control.

In this paper, the evaluation of facial extraction and facial recognition for both subspace method and dimensionality reduction are discussed. The concept and algorithms of linear and nonlinear subspace methods for global against local structure are also discussed below. Nonlinear methods lose their advantages especially when there is an inconsistency subspace in the data sets. We will use results obtained from several previously conducted experiments using different types of data sets to be used to clarify our examinations. Some representation tasks outcome seem to favor nonlinear methods however between linear and nonlinear their performances frequently vary from their preprocessing methods, training analysis schemes and classifiers selections. For that reason, our goal is to evaluate on the properties of which combination of algorithm are to be used in order to best produce advanced techniques in realistic face recognition systems. This paper presents an investigation and a comparison among several experiments that their results were statistical tested. At the end, a nonlinear analysis and subsequent discussions on the complexity of real-world face data are presented to further explaining the findings.

The remaining Sections of this paper are structured as follows. Section 3 briefly reviews a range of nonlinear and linear's subspace learning and dimensionality reduction methods. Several experiments related to the application on results of face recognition on various benchmark data sets and two-dimensional representation of real world data sets are reported in Sect. 4. In Sect. 5 we will look at the selected algorithms and methods that we believe may help in improving the result of facial recognition where we will focus on the methodology and implementation of this suggested methods. Finally, the conclusion of the research is presented in Sect. 6 by suggesting future works that can be conducted based on the results obtained in this work.

2 Related Works

The work of Darwin [2] and Galton [3] are to be known as the earliest studies on face recognition. Darwin' studies have reached until facial expressions analysis, where he believes that featuring various different emotional states is essential in facial recognition. Galton continued the facial recognition studies by contributing studies on facial profiles. In the late 1960's Galton had made his first endeavor in developing semi-automated facial recognition systems. In early 1970's, Galton had improved facial recognition studies by incorporating geometrical information where landmarks and points were placed on facial photo image to locate major facial features. Features such as mouth corners, ears, eyes, noses identified as land marks and by using their relative distances and angles. Galton used these series of computed numbers as reference points and they are to be collected either as arrays registered to each facial image. Further to that, Goldstein Harmon [4] has initiated a method or a system of 21 point that marked included the lips thickness and the hair color. Since the measurements of these markers were made by hand, automating them makes it very hard. Yang [5] came up with a more consistent approach where no manual marking is done. Later Yuille et al. [6] improvised Fishler's approach on facial recognition by measuring the facial features with templates of single facial features and mapped them onto a global template. It turns out that, they, Fishler and Yuille, were known to be the first stages of facial recognition focused on automatic detection of individual facial features. Despite of geometrical feature-based methods being insensitive towards illumination, measurement techniques and geometric feature-based recognition of a face alone are inadequate for face recognition accuracy.

Geometric feature-based recognition has slowly been left out due to several disadvantages in the technique and mainly because of the increasing interest in holistic color research which able to provide consistency in result. By aligning a set of different faces with holistic color-based techniques, enable to gather relationship between pixels concentration or intensities, where by locating the nearest neighbor classifier to organize new facial image first aligned to the set of already aligned images. Eigenfaces technique, a well known statistical learning technique which has indeed enhance the learning techniques of facial recognition. Unlike, geometrical direct comparison, Eigenface which converts images into pixel size and comparing them by focus on the concentrated areas of the various facial images. The dimension of the input intensities were first compact (reduced) with the Principal Component Analysis (PCA), known as the first principal to apply Eigenface technique. After PCA, the next found or evolved technique is known as Fisherfaces. [5], also known as Fisher Linear Discriminant Analysis (FLDA) that incorporates Eigenfaces techniques with better segregation of the individual faces vectors. Fisherfaces will first reduce the dimension of the input intensity vectors with PCA and then apply FLDA for better segregation of faces vectors in order to get an optimal result.

3 Methodology

Facial Recognition systems usually consist of four steps [7], as shown in Fig. 1 below; (1) Face Detection (localization), (2) Face Preprocessing (face alignment/normalization, light correction and etc.), (3) Feature Extraction and (4) Feature Matching. These steps are described in the following sections.

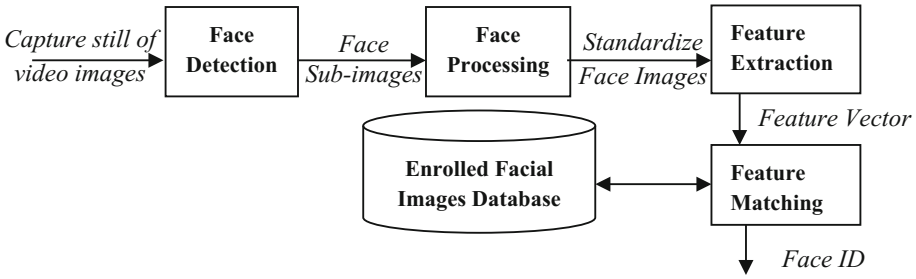


Fig. 1. Facial recognition systems

The main objective of face detection is to localize a facial image. It is an advantage to track the face in between multiple frames from video or when able to track from multiple images produced by 3 cameras, the more images able to reduce computational time and conserve the facial structure and texture of a face between frames. Face detection uses methods include: Active Appearance Models (AAM), Neural Networks, and Shape templates.

The main objective of the face preprocessing is to normalize the raw/coarse face detection so that able to extract the important essence for the process of feature extraction. This is achieved by using face preprocessing encompass – scaling, rotation, translation which are the process of alignment, light normalization and correlation. On the other hand, the main goal of feature extraction is to remove the mean redundant value, a set of similar selective geometrical or/and photometrical features of the face. Methods that are popular for feature extraction include: Principal Component Analysis (PCA), Fisher Linear Discriminant Analysis (FLDA) or Linear Discriminant Analysis (LDA) and Locality Preserving Projections (LPP).

Feature matching is the actual recognition process but it cannot be achieved without the previous preparation steps. The feature vector produced or extracted from the feature extraction is to be matched to classes of facial images vector which are already registered in a database. There are various matching algorithms which are applicable from Nearest Neighbor to advanced schemes like Neural Networks. Table 1 below are the list of Facial Extraction Algorithm Methods that were developed through time for Appearance based and Model based.

This paper focuses on the face recognition on Appearance based and not on Model based because apparently modelling processing had to deal with a very long processes and requires more raw images of an individual or a video of a person in order to initiate any modelling process mainly because modelling based requires a clear land marking

of a facial but in a real life scenario most unsupervised images are not perfect enough for 3D source or raw materials. Section 3 discusses in details all the appearance based algorithms. Pose variation and illumination are the two main problems face by face recognition researchers. Raw images are taken from source such as augmented reality home entertainment systems, video security systems or video surveillance where input data may came from uncontrolled environment. The uncontrolled environment constraint involves several obstacles for face recognition. Pose variation and illumination are the two main problems face by face recognition researchers. Raw images are taken from source such as video surveillance or augmented reality home entertainment systems where input data may came from uncontrolled environment. The uncontrolled environment constraint involves several obstacles for face recognition.

4 Linear vs Non-linear Dimensional

4.1 Linear Algorithms

PCA (Principal Component Analysis) is an analysis of data using linear technique in a global structure method (captures the variance of the input data in a linear format) that was invented by Karl Pearson back in 1901. It is designed to model linear variation in high-dimensional and reduced dimensional. Often used to reveal the concentrated structure from a group of data which in a way best depicts the main distribution and directions

Table 1. Facial extraction algorithm methods

Method	Preserving	Technique	Analysis
Appearance based	Global structure	Linear	PCA (Principal Component Analysis)
			ICA (Independent Component Analysis)
			LDA (Linear Discriminant Analysis) or FLDA (Fisher Linear Discriminant Analysis) or FLD (Fisher Linear Discriminant)
		Non-linear	KFDA (Kernel Fisher Discriminant Analysis)
			KPCA (Kernel Principal Component Analysis)
	Local structure	Linear	LPP (Locality Preserving Projections)
		Non-linear	ISOMAP (Isomap)
			Laplacian Eigenmap
			Elastic Bunch Graph
Model based		2D	Active Appearance Model
			3D Morphable Model
		3D	

of a group of data. PCA linear dimensional reduction reduces the data dimensionality by converting them to eigenvectors and select related data along the direction of data distributed variability. The output vector in the transformed embedding subspace will no longer connected nor represent a specific link from the original high-dimensional space however, it rather encapsulate data similarities in low-dimensional space in a unsupervised manner. It seems that PCA is not sensitive towards embedded data or their classification, therefore ICA, LDA are created [8].

Independent Component Analysis (ICA) [5] was designed to solve blind source separation, non-gaussian distribution and directional issues. Unlike PCA, ICA does not perform feature extraction, but using the feature space obtained by PCA. When given a face image as a source, PCA will extract the mean image, perform an eigenface process and will produce a brighter version of the face image. On the other hand, ICA will find particular feature in face such as nose, eyebrows, mouth, hair or other parts of the face. This way ICA is able to compute quicker, independently unsupervised learning and able to get rid of other distractions. ICA algorithms are known to have difficulties when the sources are nearly Gaussian. Two-component ICA problems with identical source distributions are studied in order to address these issues. These distributions were chosen at random among a set of mixtures of Gaussians which are at various distances from Gaussianity. This set includes both supergaussian (positive kurtosis) and subgaussian distributions (negative kurtosis) [8]. Thus, ICA method alone is not appropriate in facial recognition.

LDA Linear discriminant analysis is a pattern recognition statistic analysis that uses PCA's eigenvector, linear technique and preserving global structure of features which later separates two or more group of classes (discrimination or segregation among vectors). Unlike PCA (unsupervised), LDA is a supervised method, which implies that all training-data samples must be associated (manually) with a class. In this way LDA increases the between-class variance as well as reduces the within-class variance. LDA will first search for the midpoint axes where this point is the point that separates different classes the most further among classes but requiring data points of the same class to be nearer to each other. If PCA encodes data to an orthogonal linear space, we find LDA to encode segregating information in a linearly separable form which is not necessarily according to orthogonal format. LDA and PCA both try to reduce dimension however, LDA creates two axes that maximize the separations or gaps between classes of the data. Like PCA, LDA also creates a face vector space, normalizes the face vector, calculates covariance matrix, conduct eigenfaces from co-variance matrix. In algorithms based evaluation wise, LDA method is superior to those method on PCA, nevertheless, according to recent works using several types of dataset, it shown that, when data set that processed to training data set is small, PCA can outperform LDA. Unlike LDA, PCA is less sensitive towards different set of training data sets.

Locality Preserving Projections (LPP) is another linear technique, but an orthogonal linear, that uses locality structure for face recognition data vector [9] by combining embedded data as a way to preserve local information. It uses face subspace to best detect the important part of face manifold structure. LPP is a new linear projection occurs when there is a high dimensional data lies embedded in the ambient space on a low dimensional manifold. LPP is the result when finding Eigen functions in optimal linear

approximations of the Laplace Beltrami operator on the manifold. Thus, with this, despite of being a linear, LPP also being used by many of the data representation properties of nonlinear analysis methods, for example Laplacian Eigenmaps or Locally Linear Embedding (LLE).

4.2 Non-linear Algorithms

These nonlinear methods such as Isomap [9], LLE [9], and Laplacian Eigenmap [9] produce impressive results on some benchmark artificial data sets but never the less, they lose in performance on novel test data points maps that they are only defined on the training data points but very unclear to evaluate their maps on novel/original test data points. As a result, these nonlinear manifold learning techniques [9], may or might not be appropriate or suitable for the application of some computer vision tasks, which in this case, face recognition.

Kernal PCA [10] is a nonlinear analysis of data, which is an extension of PCA, where by a data set is projected by using a hypothetic non-linear function in a high dimensional feature space. PCA cannot reduce the dimensionality from two to one because the data points may not be located along a straight line, so when these data are located around a one-dimensional non-linear curve, PCA cannot be applied, but only by Kernel PCA that this non-linear manifold can be discover along the data which are in fact nearly one-dimensional. By mapping these data in a higher dimensional space, we finally discover that they are lying on a lower dimensional subspace. So in KPCA dimensionality is increased in order to be able to decrease it.

Kernel Fisher Discriminant Analysis (KFDA) projects data into a high-dimensional feature space, where a Linear Discriminant Analysis (LDA) or also known as Fisher Discriminant Analysis (FDA) concept is performed on the data [10]. There are some similarity approach towards LPP despite the data structure is locally preserved in the subspace, LPP allows nonlinear classifications like KFDA. KFDA is an alternative KPCA algorithm that employs a kernel function which intrinsically maps the original data distributions to a space. KFDA [11] moves is classified nonlinear in the original space but linear in the kernel one which redefines the within and between-class scatter matrices in the kernel space to obtain feature extraction algorithms.

The Isometric Mapping (Isomap) algorithm [12] modifies classical Multidimensional Scaling (MDS) by the use of a neighborhood mapping to manage nonlinearities data. Compared to Isomap, the locally linear embedding (LLE) [11] exploits another idea to preserve the structure in the original feature space: the neighbourhood preservation. Based on the same assumption that in a small scale of neighbourhood the local distance metric is Euclidean in dimensional reduction, LLE preserves the local relationship, also called the local geometry, of each sample with its neighbors while ignores the global geometry in large scale.

Laplacian Eigenmap is another nonlinear just like Isomap and LLE, in 2002, Belkin et al. [11] had proposed for another way of thinking in manifold learning, which he called the Laplacian eigenmap. In their claim, to have an optimal embedding should keep neighbor points in the d-dimensional space and still close in the p-dimensional

space. This idea can be formulated as minimizing the summation of the Laplacian-Beltrami operator over the entire manifold. The Laplacian-Beltrami operator computes the divergence of mapping function from a point on the manifold. Laplacian eigenmap is based on the same idea of LLE, preserving only the local geometry, while uses another method to model the objective function. The experimental results in showed that Laplacian eigenmap could achieve nonlinear dimensionality reduction in both visualization and data representation tasks.

5 Comparative Analysis of Previously Conducted Experiments

In this section, a comparative analysis is made for three experiments conducted previously in order to gauge into the significance of all these experiments.

5.1 CASE 1: Multiple Individual Discriminative Models (MIDM)

Multiple Individual Discriminative Models has been introduced in which not only the texture intensities are taken into account but also the geometrical information [13]. The aim of each of the individual models, obtained by the projection of the training set into the c one-dimensional spaces, is to characterize a given person. This means that every person in the data set is represented by one model. In creating the individual models, the geometry of the face (e.g., landmarks alignment) and the texture information (e.g., normalized texture) are captured. PCA is used to produce Eigenfaces by removing redundancy from both sets of features before combining them. After combining them, the PCA will be used again to extract the redundancy or mean from the combined features. Finally, FLDA is used to build the individual model.

Experiments have been conducted in order to compare the outcome of using MIDM method with respect to the Fisherface's methods in terms of correct classification rates. In order to evaluate the importance of the geometrical information, the Fisherface technique was modified replacing the texture data with the shape data and also combining the shape with the texture. These two modified techniques will be referred to as Fishershape and Fishercombined. The Euclidean Nearest-Neighbor algorithm was used as classifier algorithm in the Fisher methods. Based on the results obtained shown in Table 2, the proposed method classified the images as the person associated to the model that yields the highest probability. From Table 2, it is observed that the proposed method has a slightly better performance than the Fisher methods, in which using the texture data one obtains a higher accuracy than when the shape is used. This implies that the information contained in the texture is more significant than that included in the shape. However, the highest correct classification rate in both techniques is attained when both shape and texture are considered.

Table 2. Average correct classification rates

Method	Input features	Correct identification rate
PCA + FLDA	Shape	86.4% (95)
PCA + FLDA	Texture	99.6% (3)
PCA + FLDA	**Texture and Shape**	**99.9% (1)**
Fishershape FLDA	Shape	85.9% (99)
Fisherface FLDA	Texture	96.9% (22)
Fishercombined FLDA	Texture and Shape	97.1% (20)

Note: Number of misclassified images reported in parentheses.

5.2 CASE 2: Linear and Nonlinear Dimensionality Reduction

There were several previously works conducted on evaluating both linear and nonlinear projections on face recognition evaluation [14]. Six dimensional reduction methods were evaluated that includes (1) PCA, (2) KPCA1 (polynomial), (3) KPCA2 (Radial Basis), (4) LLE, (5) Isomap and (6) CCA, using four different classifiers that include (1) NN, (2) soft k-NN, (3) LDA and (4) SVM on ORL database (composed of 400 images of 40 persons, where each person has 10 images) and Yale database (composed of 5760 single light source images of 10 persons under 576 viewing conditions with 9 poses and 64 illumination conditions) Based on the experiments conducted results, overall, it showed that the more training samples available, the higher the classification rate is. From the above experiments, LLE had slightly outperforms the others by less than 1% in most implementations with both NN and soft k-NN classifiers. For a combination performance of 1 or 2 method(s) and 1 or 2 classifier(s), we found that PCA has the best performance with LDA and SVM classifiers, and improved further after implementing both NN and soft k-NN classifiers on the ORL database. As for the Yale database with PCA coupled with LDA classifier had shown the highest accuracy rate in all implementations. Therefore, based on this experiment we can see that nonlinear methods do not always significantly perform better than linear method such as PCA especially in reducing dimensionality of face data [14].

There has been previous works conducted that showed that nonlinear techniques are more prevailing or more influential than linear PCA for capturing nonlinear structure on a high-dimensional data. Before we proceed, we have to understand that most experiments with these nonlinear projections were conducted using artificial data sets that are not based on real captured data especially for Yale database. Further analysis have been done, to see the individual method performance of PCA (linear) with LLE and CCA (2 non linear) using the NN as a classifier. This comparison is to compare between linear and nonlinear in increasing and reducing dimensionality by varying from 5 to 70 dimensions. From the results that were obtained, several conclusions were made. First conclusion is that the performance of three methods increases with the increased the number of dimensionality reduction, where lower nonlinearities as compared to high nonlinearities lead to better performances. PCA showed slightly better than the two others nonlinear due to smaller training images set in ORL database (10 images per person). Second, by comparing lower dimensionality reduction and classification rate in ORL

database, it showed stable performances among all three methods however nonlinear methods had similar or slightly better performances which were only less than 1% improvement in classification rate than PCA. Third, LLE had lower classification rate significantly than that of PCA especially in the reduced dimension. Data sets structure cannot be preserve in LLE. Fourth, in Yale database, nonlinear techniques had outperformed PCA in the reduced dimensions below 30 because the huge data set per individual (576 sample images per person) which allows nonlinearity to perform better in these dimensions [14].

Though nonlinear techniques have capabilities in capturing data structure of nonlinear, they may not often lead to significant improvement in face recognition performance in mainly because real face data may distribute fairly linearly and the nonlinear capabilities of those nonlinear methods may not be effective in projecting these high-dimensional faces data [14].

5.3 CASE 3: Multilinear Principal Component Analysis (MPCA) and Locality Preserving Projections (LPP)

LPP algorithm advantages are not only that it is an unsupervised projection that performs a linear transformation and preserves the local information of the face image space (set neighborhood structure) but also covers an adjacency graph models expressing local nearness data along the manifold structure, However, LPP has a limitation that it represents an image by a vector in high dimensional space. In order to address this issue, an approach has been proposed that Multilinear Principal Component Analysis (MPCA) subspace and LPP combination algorithm is to be used to preserve the local structure's information. This combined approach will improve both the global and local structure of the face image space in order to obtain a more optimal and effective subspace for face representation and recognition. Firstly, LPP will compress and preserve the principal information in a matrix form, here; it removes more inherent redundancy and acquires a much lower dimensional face representation by which the recognition speed is greatly enhanced. Secondly, once it is in a low dimensional representative with MPCA it achieves a more competitive accurate recognition rate than the Laplacianface.

Based on the analysis, it can be concluded that the performance of LPP supersedes the PCA and LDA methods in which, based on the latest experimental results conducted, it has shown that the combination of LPP with MPCA has improved the in face recognition rate accuracy [15].

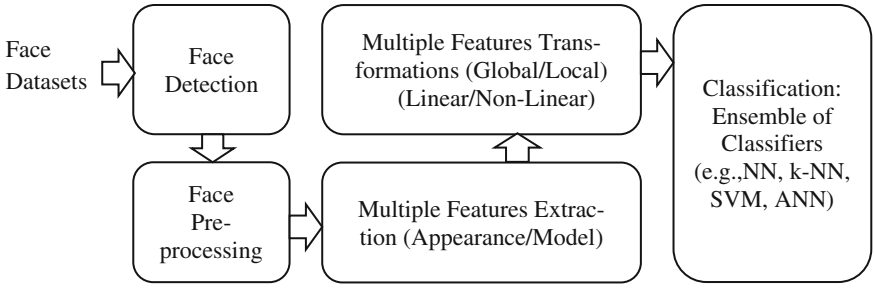


Fig. 2. Schema of the proposed face recognition ensemble

6 Framework for Effective Synergy in Facial Recognition

Based on the findings, it can be concluded that

(1) Multiple sets of extracted features can be combined to enrich the facial representation according to the Appearance and Models based as listed in Table 1.
(2) The PCA is much easier to implement and less computationally demanding.
(3) Multiple feature dimensionality reduction methods (e.g., feature transformation) can be combined to preserve global and local structure.
(4) Ensemble of classifiers can be designed to optimize the performance.

Figure 2 illustrates the proposed framework to be addressed for effective synergy in facial recognition. As illustrated in Fig. 2, the proposed approach can be broken down into the following steps:

Face detection: first the precise position of the face image is detected and the resulting face is cropped and aligned according to eye position.
Preprocessing: several enhancing methods will be tested in this work in order to make the feature extraction more robust to changes in illumination, noise, etc.
Feature extraction: this step is performed separately on each image resulting from the previous preprocessing method in order to obtain different sets of features based on appearance and models.
Feature transformation: before classification the dimensionality of each descriptor is reduced (e.g., via Principal Component Analysis (PCA)).
Classification: a set of general-purpose classifiers is trained on each reduced set of features. The final decision is then determined according to consensus results obtained from multiple classifiers.

7 Conclusion

This paper has performed a comparative analysis of previously conducted experiments. Based on the findings from the comparative analysis of previously conducted experiments, it can be concluded that the performance of the facial recognition can be improved by combining multiple sets of extracted features in order to enrich the facial

representation according to the Appearance and Models based. In addition to that, the global and local properties of the image can be preserved with more effectively by combining multiple feature dimensionality reduction methods. Finally, ensemble classifier can be proposed to accommodate multiple perspectives of the dataset and to ensure better facial recognition performance can be achieved. Based on these finding, a schema of the proposed framework for face recognition is proposed.

References

1. Roizenblatt, R., Schor, P., Dante, F., Roizenblatt, J., Belfort, R.: Iris recognition as a biometric method after cataract surgery. Biomed. Eng. Online **3**, 2 (2004)
2. Darwin, C.: The Expression of the Emotions in Man and Animals. John Murray, London (1872)
3. Galton, F.: Personal identification and description. Nature **38**, 173–188 (1888)
4. Goldstein, A.J., Harmon, L.D.: Identification of human faces. Proc. IEEE **59**, 748–760 (1971)
5. Yang, M.H.: Kernel Eigenfaces vs. Kernel Fisherfaces: face recognition using kernel methods. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (2002)
6. Yuille, A., Hallinan, P., Cohen, D.: Feature extraction from faces using deformable templates. Int. J. Comput. Vis. **8**, 99–111 (1992)
7. Cox, I.J., Ghosn, J., Yianilos, P.N.: Feature-based face recognition using mixture-distance. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 209–216, June 1996
8. Turk, M., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (1991)
9. de Carrera, P.F.: Face Recognition Algorithms (2010)
10. Goldberg, Y., Zakai, A., Kushnir, D., Ritov, Y.: Manifold learning: the price of normalization. J. Mach. Learn. Res. **9**, 1909–1939 (2008)
11. Belkin, M., Niyogi, P., Sindhwani, V.: Mainfold regulation: a global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
12. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neurosci. **3**(1), 71–86 (1991)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
14. Huang, W., Yin, H.: Linear and nonlinear dimensionality reduction for face recognition. School of Electrical and Electronic Engineering, The University of Manchester, UK (2009)
15. Shermina, J.: Application of locality preserving projections in face recognition. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **1**(3), 82–85 (2010)

Detection and Defense Algorithms of Different Types of DDoS Attacks Using Machine Learning

Mohd Azahari Mohd Yusof[1(✉)], Fakariah Hani Mohd Ali[2], and Mohamad Yusof Darus[2]

[1] Kolej Universiti Poly-Tech MARA Kuala Lumpur, Kuala Lumpur, Malaysia
azaha_ri@yahoo.com
[2] Universiti Teknologi MARA Shah Alam, Shah Alam, Malaysia
{fakariah,yusof}@tmsk.uitm.edu.my

Abstract. Recently, many organizations require security tools to maintain their network or IoT environment from DDoS attacks. Most security tools today, do not have enough power to detect whether the incoming packet is a normal packet or DDoS packet. The purpose of the DDoS attack is to undermine the web server of an organization that may run a business. Therefore, this research is conducted to design a technique called Packet Threshold Algorithm (PTA) coupled with SVM in order to detect four types of DDoS attacks such as TCP SYN flood, UDP flood, Ping of Death and Smurf. The results of this research on the use of this technique is claimed enable the action of minimizing false positive rates and increases the detection accuracy in comparison to the other three current techniques. The TPA-SVM technique has the capability of detecting incoming packets as normal packets or DDoS attacks. The DDoS attack type of detection is based on the packet threshold.

Keywords: DDoS · Internet of Thing (IoT) · Packet Threshold Algorithm (PTA) Support Vector Machine (SVM)

1 Introduction

Nowadays, networks are very crucial to everyone as they offer many advantages. One of the main advantages is the sharing of resources. A network is a connection between two or more computers, so that users can exchange information with each other. The combination of computer networks around the world has formed an indispensable technology known as the Internet. The Internet brings many advantages, but it depends on the purpose of using the Internet. Internet technology provides various conveniences for users to enjoy entertainment such as Online Games to connect with others who are far away. It also provides social media platform such as Facebook, Twitter and Instagram for socializing users even though they do not know each other. Most importantly, the Internet is accessible 24 h a day because the Internet is an economical communications platform. However, the network and IoT environment face various threats launched by attackers such as the Distributed Denial of Service (DDoS) attack.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 370–379, 2018.
https://doi.org/10.1007/978-981-10-8276-4_35

The DDoS attack makes the networks or systems unavailable for use of network users even if they are legitimate users and had full authority to access them. Commonly, attackers will use many computers called botnet to launch DDoS quickly to one or more targets. DDoS attacks can be divided into three categories which are volume-based attack, protocol attack and application layer attack [1].

Volume-based attack includes UDP flood, ICMP flood, and other spoofed-packet flood. The UDP flood occurs when the attacker sends a large UDP packet to the target server to slow down the response process to a legitimate user [2]. ICMP flood occurs when an attacker sends abnormal IP packets to the target server to make the server inaccessible by other users [3]. Protocol attack includes TCP SYN flood, Ping of Death and Smurf attack. TCP SYN flood occurs when an attacker sends repeated SYN packets to the target server using spoofed IP address and to make the server crashes [4]. Ping of Death occurs when an attacker sends ping packets greater than 65535 bytes to make the target server inaccessible by users [5]. Smurf attack occurs when an attacker sends a large number of ICMP packets to crash, freeze or destabilize the target server [6]. The application layer attack includes Slowloris and Zero-day attack. Slowloris is a DDoS software that enables a single computer to take down a web server. The Zero-day attack occurs when an attacker exploits the vulnerability of software or hardware by releasing malware before a developer has an opportunity to create a patch to fix the vulnerability [7]. No matter what kind of DDoS attacks launched by the attacker, it could paralyze the target server even though the method of attack is simple.

2 Related Work

DDoS attacks can be a serious problem for businesses, system administrators and network users as it can interfere with various systems across the Internet using infected zombies [8]. Zombies are created to strike targets with various types of packets such as TCP, UDP and ICMP. This will result in inaccessible of business information, losing business opportunities and contracts, corporate credit ratings and insurance premium increases as reported by Malaysian Insider (2015). However, some solutions have been proposed to combat some of these DDoS attacks.

The study conducted by [9] has proposed a multi-queue algorithm for gateway and router to detect DDoS attacks. The algorithm was formed from the combination of two congestion control methods, namely drop tail congestion control algorithm and random early detection algorithm. The algorithm can increase network throughput even the network is under DDoS attack. However, tail drop congestion control algorithm is cannot properly distinguish between normal packets and DDoS packets and does not have the technical policy. Apart from that, there is a lack of synchronization techniques and what makes it worse is its open mechanism. Random early detection algorithm also could not distinguish between normal packets and DDoS packets and the queue is not filled by a single flow.

The study conducted by [10] has proposed an algorithm called cumulative sum algorithm, where there are two states, namely Not Under Attack (NA) and Under Attack (A). *ipac* and *ddos* are two top-level functions that are included in the algorithm. The

ddos function will determine whether the network system is A or NA state by analyzing the incoming packets. The *ipac* functions as a way to check the IP address of incoming traffic either new IP address or not. However, the algorithm still has the problem of false positive rates because when the system is under attack, the new IP will also be detected as DDoS packets. Apart from that, a very slow time interval used by *ipac* and *ddos* can cause the received packet to not be processed correctly.

The study conducted by [11] have proposed DDoS attack prevention strategy called the dynamic security level changing strategy algorithm for a server node. The strength of the algorithm is to protect neighboring nodes that are under attack. Apart from that, the algorithm can specify the types of DDoS attacks. However, the algorithm still faces the problem of false positive rates when a normal or clean traffic is incorrectly identified as an attack and it will affect the detection accuracy.

[12, 13] have proposed SVM to detect DDoS attacks with the pattern generated from the DARPA dataset. Furthermore, the researchers did a comparison between SVM with other techniques such as naive bayes, bagging, radial basis function network, J48 decision tree and random forest and found SVM performs better in terms of false positive rates and detection accuracy. However, SVM cannot protect victims who are attacked if the attacker uses the actual IP address. Therefore, detection of incoming packet requires full treatment even DDoS attack coming from a spoofed IP address has stopped. Moreover, SVM takes a long time to be trained to learn real users in a network environment.

The study conducted by [14] has developed an algorithm called worldwide SYN flooding attack detection algorithm, where it is used to detect DDoS attack. They have conducted their research by introducing eight attack scenarios with 14 types of the SYN flooding attacks. However, the algorithm still has weaknesses due to normal packet is detected as DDoS attacks. For example, only 80 incidents were detected from 307 incidents during the algorithm implemented.

The study conducted by [15] proposed that IAFV and correlation algorithm to detect normal packets and DDoS attacks. Both algorithms are also used to determine the state of false negatives and false positives when it is carried out for the purposes of detecting normal packets and DDoS attacks. However, IAFV algorithm is still not sufficient to detect DDoS attacks compared to correlation algorithm as IAFV algorithm still has higher false positive rates. Correlation algorithm also has the weakness of reducing the problem of false positive rates even though it is better than IAFV algorithm.

The study conducted by [16] proposed that modified k-means algorithm could detect DDoS attack in DARPA 98 dataset. However, the resulting experiments showed that false positive rates are still at high rates and the technique still needs some additional features to improve detection accuracy.

The study conducted by [17] proposed a lightweight detection algorithm and coupled with the hop-count filter to observe and detect DDoS attacks. This algorithm is set to threshold, which if incoming packets exceeds the threshold, it indicates that it is a DDoS attack and an alarm signal will be triggered. They used Flooding DDoS Attack 2007 dataset from CAIDA for the data traffic. However, this algorithm still has shortage if the throughput of DDoS packets is not consistent for specific timeslots, it will not be able to detect the synchronization of attack flows.

The study conducted by [18] has proposed the logistic regression to detect the application layer DDoS attack. Researchers have constructed different characteristics to differentiate attackers and normal users to model user behavior. They have compared the proposed techniques with existing techniques and they are hidden semi markov model, random walk graph, and hierarchical clustering. They found out that their proposed techniques showed better results than the other three existing techniques. However, the technique still projects high false positive rates because the incoming traffic cannot be tracked properly whether it is an attacker or a normal user.

The study conducted by [19] proposed a model system to detect and specify DDoS packets using artificial neural network. They have collected data used in their research through online sources, which contains 4986 network traffic coming from four datasets. However, this technique failed to accurately classify the types of DDoS attacks.

From some current techniques that were implemented, it can be concluded that the problem of false positive rates and detection accuracy is still ongoing and as researchers need to produce a new method of detection and defense against DDoS attacks.

3 Proposed Technique

To address the problems identified by current techniques, the research will continue based on the basis of four phases as shown in Fig. 1.

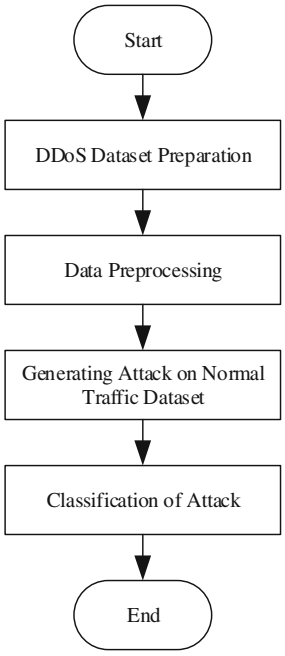


Fig. 1. Research methodology

3.1 DDoS Dataset Preparation

A dataset is a collection of indispensable data to get great classification results. This research uses a new dataset that was collected using Wireshark tool as shown in Fig. 2. Moreover, this research uses machine learning techniques and it is great for detecting and classifying network packet whether it is normal packets or DDoS packets.

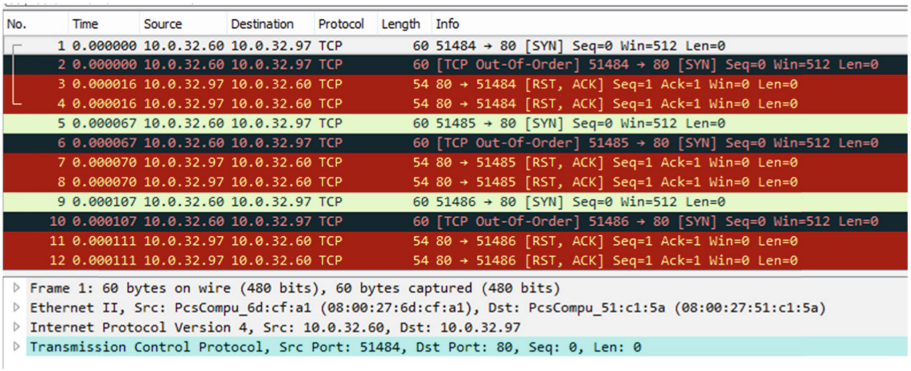


Fig. 2. Captured packet

3.2 Data Preprocessing

Data preprocessing is executed to transform raw data into an understandable format, which the previous data had multiple duplicate values, redundant data and missing values. After the data preprocessing is done as sample shown in Table 1, this study has obtained quality results.

Table 1. Sample dataset

Src_Add	Des_Add	Pkt_Type	Pkt_Size	…	Pkt_Class
10.0.32.97	10.0.32.60	TCP	54	…	Normal
10.0.32.60	10.0.32.97	ICMP	60	…	Ping-of-Death
10.0.32.97	10.0.32.60	ICMP	42	…	Smurf
10.0.32.60	10.0.32.97	UDP	60	…	UDP-Flood
10.0.32.60	10.0.32.97	TCP	60	…	TCP-SYN-Flood

3.3 Generating Attacks on Normal Traffic Dataset

There are several requirements that are used for generating attacks on normal packet dataset as shown in Fig. 3.

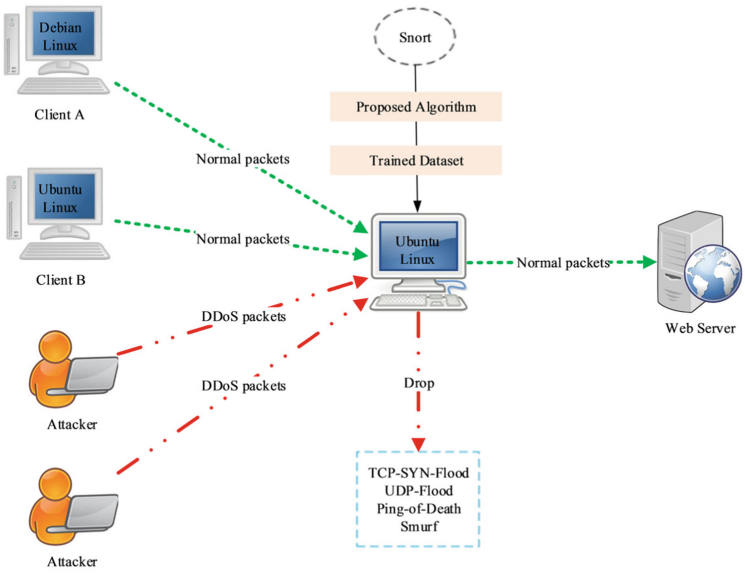


Fig. 3. Experimental setup

Experiments start with generate normal and DDoS packet using Hping3 tool. Then, Snort will capture the incoming packet and specify the packet is normal or DDoS packet based on proposed algorithm and trained dataset. DDoS packet detected by Snort will drop and only normal packet is allowed into the network.

3.4 Classification of Attack

Detection and defense technique for this research uses Packet Threshold Algorithm (PTA) as shown in Fig. 4 coupled with SVM to classify and mitigate network and IoT environment as well against DDoS attacks. The advantage of this proposed algorithm is that it can detect incoming packets as normal or DDoS packets. If the packet is DDoS packet, then it will learn the type of DDoS attack based on the specified packet threshold. If TCP SYN with threshold 60 SYN per second is detected, then the type of DDoS attack is TCP SYN flood. If UDP packet with threshold 60 UDP per second is detected, then the type of DDoS attack is UDP flood. If a packet size is greater than 75 bytes of ICMP per second, then the type of DDoS attack is Ping of Death. If ICMP replies or abnormal operation exist between client and server, then the type of DDoS attack is Smurf. If the packet is not within the packet threshold range, it is a normal packet that is allowed into the network.

Check incoming traffic
If ($P_{Threshold}$ = $N_{Threshold}$)
Traffic is $N_{Traffic}$
Else
Traffic is $A_{Traffic}$
Study $A_{Traffic}$
If $P_{Threshold}$ >= 60_{SYN}/second
$A_{Traffic}$ is TCP SYN flood
If $P_{Threshold}$ >= 60_{UDP}/second
$A_{Traffic}$ is UDP flood
If $P_{Threshold}$ >= 75_{ICMP}/second
$A_{Traffic}$ is Ping of Death
If $ICMP_{Reply}$
$A_{Traffic}$ is Smurf
Else
Traffic is $N_{Traffic}$

Fig. 4. Packet Threshold Algorithm

This research uses SVM as a classifier, which is used to train the dataset, so that it can detect DDoS attacks according to appropriate accuracy. Moreover, SVM can provide an accurate classification because it can reduce false positive rates [12] and it is one of the most popular supervised learning algorithms for certain applications like intrusion detection system, spam filtering and pattern recognition [20].

4 Experimental Results

This section describes the experimental results obtained from the use of PTA-SVM, which shows the number of packets detected as shown in Fig. 5. The number of packets that have been captured is 228425 packets, which consist of 109403 normal packets, 49644 TCP SYN flood packets, 7677 UDP flood packets, 41136 Ping of Death packets and 20565 Smurf packets.

PTA-SVM is compared with Lightweight, Modified K-Means and Logistic Regression technique as shown in Fig. 6. The comparison between three current techniques, it shows the Modified K-Means is better than Logistic Regression with 98.9% detection accuracy and Logistic Regression is better than Lightweight with 98.6% detection accuracy. However, the PTA-SVM technique in this study is better than the three current techniques with 99.1% detection accuracy and 1.11 false positive rates.

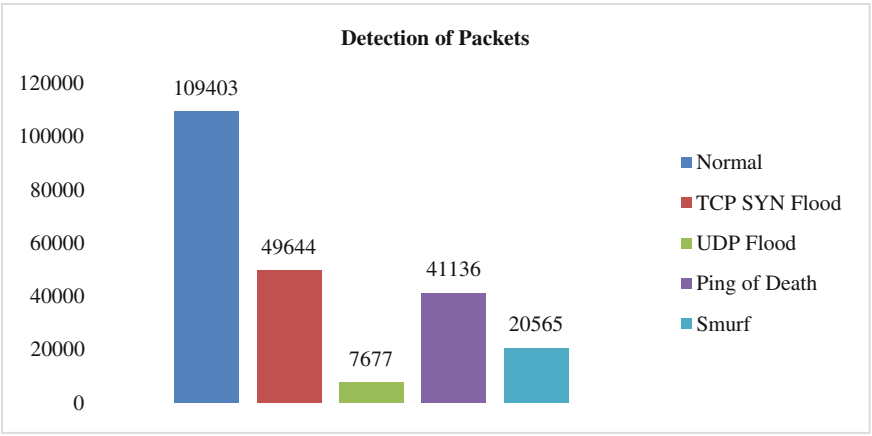


Fig. 5. Detection of packets

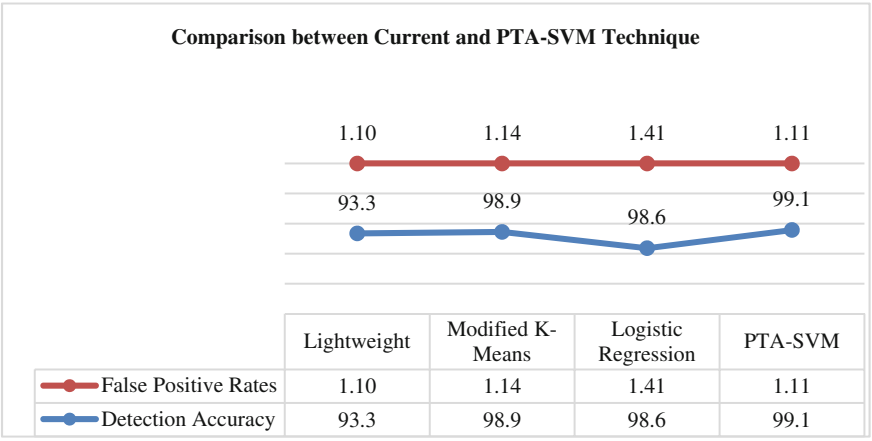


Fig. 6. Comparison between current and PTA-SVM technique

5 Conclusion

The network and IoT environment are very vulnerable to the ever-increasing DDoS attack. Therefore, this study was conducted to propose a technique called PTA-SVM to detect four types of DDoS attacks, which are TCP SYN flood, UDP flood, Ping of Death and Smurf. In this technique, it attempts to detect incoming packets as normal packets or DDoS attacks and then it can detect the type of DDoS attack. The results of this research indicate that the PTA-SVM technique is able to ensure that DDoS attacks are better than the three selected techniques in terms of detection accuracy 99.1% with false positive rates 1.11.

Acknowledgment. This research was supported by the Research Management Institute, Universiti Teknologi MARA and registered under the LESTARI #600-IRMI/DANA 5/3/LESTARI (0107/2016).

References

1. Imperva: DDoS Protection Center, Incapsula for Cloud Providers, 12 January 2017. www.incapsula.com/. Accessed 5 July 2017
2. Kolahi, S.S., Treseangrat, K., Sarrafpour, B.: Analysis of UDP DDoS flood cyber attack and defense mechanisms on web server with Linux Ubuntu 13. In: International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah (2015)
3. Gupta, N., Jain, A., Saini, P., Gupta, V.: DDoS attack algorithm using ICMP flood. In: International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi (2016)
4. Geetha, K., Sreenath, N.: SYN flooding attack - identification and analysis. In: International Conference on Information Communication and Embedded Systems (ICICES), Chennai (2014)
5. Buvaneswari, M., Subha, T.: IHONEYCOL: a distributed collaborative approach for mitigation of DDoS attack. In: International Conference on Information Communication and Embedded Systems (ICICES), Chennai (2013)
6. Guerid, H., Serhrouchni, A., Achemlal, M., Mittig, K.: A novel traceback approach for direct and reflected ICMP attacks. In: Conference on Network and Information Systems Security, La Rochelle (2011)
7. FireEye Inc.: What is a zero-day exploit? 2 January 2017. https://www.fireeye.com/. Accessed 6 July 2017
8. Rawal, B., Ramcharan, H., Tsetse, A.: Emergence of DDoS resistant augmented split architecture. In: International Conference on High Capacity Optical Networks and Enabling Technologies (HONET-CNS), Magosa (2013)
9. Nkemneme, F., Wei, R.: A multi-queue algorithm for DDoS attacks. In: International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hsinchu (2014)
10. Ahmed, E., Mohay, G., Tickle, A., Bhatia, S.: Use of IP addresses for high rate flooding attack detection, pp. 1–13. IEEE (2014)
11. Lim, S.-H., Kim, J.-H.: Dynamic security level changing strategy using attack predictions, pp. 1–4. IEEE (2014)
12. Kokila, R.T., Selvi, S.T., Govindarajan, K.: DDoS detection and analysis in SDN-based environment using Support Vector Machine classifier. In: International Conference on Advanced Computing (ICoAC), Chennai (2014)
13. Devi, B.S.K., Preetha, G., Shalinie, S.M.: DDoS detection using host-network based metrics and mitigation in experimental testbed, pp. 423–427. IEEE (2012)
14. Miao, L., Ding, W., Gong, J.: A real-time method for detecting internet-wide SYN flooding attacks. In: IEEE International Workshop on Local and Metropolitan Area Networks (LANMAN), Beijing (2015)
15. Saboor, A., Aslam, B.: Analyses of flow based techniques to detect distributed denial of service attacks. In: International Bhurban Conference on Applied Sciences & Technology (IBCAST), Islamabad (2015)
16. Pramana, M.I.W., Purwanto, Y., Suratman, F.Y.: DDoS detection using modified k-means clustering with chain initialization over landmark window, Bandung (2015)

17. Li, C., Yang, J., Wang, Z., Li, F., Yang, Y.: A lightweight DDoS flooding attack detection algorithm based on synchronous long flows. In: IEEE Global Communications Conference (GLOBECOM), San Diego (2015)
18. Yadav, S., Selvakumar, S.: Detection of application layer DDoS attack by modeling user behavior using logistic regression, Noida (2015)
19. Peraković, D., Periša, M., Cvitić, I., Husnjak, S.: Artificial neuron network implementation in detection and classification of DDoS traffic. In: Telecommunications Forum (TELFOR), Belgrade (2016)
20. Kato, K., Klyuev, V.: An intelligent DDoS attack detection system using packet analysis and Support Vector Machine. Int. J. Intell. Comput. Res. (IJICR) **3**(5), 464–471 (2014)

Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation

Edy Budiman[1(✉)], Haviluddin[1(✉)], Nataniel Dengan[1], Awang Harsa Kridalaksana[1], Masna Wati[2], and Purnawansyah[2]

[1] Faculty of Computer Science and Information Technology, Mulawarman University, Samarinda, Indonesia
edy.budiman@fkti.unmul.ac.id, haviluddin@gmail.com, ndengen@gmail.com, awangkid@gmail.com

[2] Faculty of Computer Science, Universitas Muslim, Makassar, Indonesia
masnawati.ssi@gmail.com, purnawansyah@gmail.com

Abstract. Student academic evaluation is part of academic information system (AIS) performance, in order to control student learning progress is necessary. Furthermore, the evaluation showing whether the student will pass or fail would benefit the student/instructor and act as a guide for future recommendations/evaluations on performance. An in depth study on the student academic evaluation technique by using Decision Tree C4.5 has been conducted. Specific parameters including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and progress GPA (pGPA) and Total GPA (tGPA) from semester 1–4 with three times graduation criteria (i.e., *fast*, *on*, and *delay*) times have been defined and tested. The scope of the paper has been set for undergraduate programs. The experimental results show that accuracy algorithm (AC) of 78.57% with true positive rate (TP) of 76.72% by using quality training data of 90% have best performance accuracy value.

Keywords: Tree C4.5 · Confusion matrix · Student academic evaluation

1 Introduction

Learning process evaluation is a process to determine an academic performance level of students which comprehensive and continuous in accordance with educational regulations. Where, the student's achievement of subjects mastery are determined by quizzes, examinations, practicums, and other tasks that covering cognitive, affective, and psychometrics capacity [1–3]. Furthermore, in general, student academic assessment are based on progress report including progress GPA (pGPA) and Total GPA (tGPA). Where, pGPA and tGPA are calculated from course subjects values. Therefore, it is of great interest to identify the students to understand which factors have a larger influence on this. Hence, a data mining model is an appropriate tool for covering these tasks, i.e., classification [4], prediction [5, 6], cluster [7] etc. [8].

Furthermore, an application of data mining in the Educational context is referred as Educational Data Mining (EDM) that defined by the International Educational Data

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 380–389, 2018.
https://doi.org/10.1007/978-981-10-8276-4_36

Mining Society [9]. In other words, EDM is talked about fields of education and information or computer sciences [2, 10]. Numerous methods in data mining are widely applied in order to perform student academic evaluation tasks, including statistical and smart computing methods. [11] have implemented two classification techniques, namely Naïve Bayes and Decision Tree Classifier to model academic attrition (loss of academic status) at the Universidad Nacional de Colombia. This studied were used academic datasets 2007-II and 2012-II from two programs, Agricultural (AE) and Computer and Systems (CE) Engineering. The results showed that NBC and Decision Tree models can be used as models in the prediction of the loss of academic status. [12] have conducted research with C4.5 and ID3 algorithms of student dropout, predicting and characterizing students at the University Simón Bolívar. This experiment was used WEKA as a tools for data processing. The results of this study confirmed that these algorithms can be used as an alternatives model. [13] have conducted study Naive Bayes, the 1-NN and the WINNOW algorithms in order to predict a student's performance. The results showed that this algorithm was the most appropriate to be used for the construction a software support tool.

The aim of this study is to investigate Tree C4.5 algorithm in order to student academic learning evaluate performance. Therefore, all students might improve and increase the learning process. It is expected that this model analysis can be used in order to support academic decisions. This paper is consists of four sections. Section 1 is the motivation to do the writing of the article. Next, the methodology and techniques is discussed in Sect. 2. Section 3 presents the experimental results and discussion, and finally Sect. 4 describes the research summaries and conclusion.

2 Methodology

2.1 Tree C4.5 Algorithm

Decision tree is a data structure consisting of nodes (i.e., root, branch, leaf) and edge. Tree C4.5 algorithm is a part of decision tree algorithm that supervised learning method [13, 14]. Tree C4.5 developed by Quinlan in the 1996s, which is derived from the algorithm *Iterative Dichotomiser* (ID3), efficient, powerful and popular [4]. In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.

In principle, Tree C4.5 algorithm consists of four steps in order to generate decision tree. First, choose attribute as a root. Second, generate branch every value. Third, put dataset in branch, and. Four, repeat the second process until every class have the same value. Formula of Entropy is shown below where S is entropy, and p is class proportion in the output.

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i * \text{Log}_2 p_i \tag{1}$$

Furthermore, the attribute with the highest gain value is used as the root attribute. Equation 2 shows the formula of the gain where, S is a set of case; A is an attribute of case; $|S_i|$ is a number of cases to i; and $|S|$ is number of cases in the set.

$$\text{Gain}(S, A) = \text{Entropy } (S) \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy } (S) \tag{2}$$

The pseudocode of Tree C4.5 is shown as follow:

```
Input: an attribute-valued dataset D
Tree = [14]
if D is "pure" or other stopping condition met then
    stop
end if
for all attribute a ∈ D do
    Compute information-theoretic condition if we split on a
end for
a_best  = Best attribute according to above computed condition
Tree = Create a decision node that tests a_best in the root
D_v = Induce sub-datasets from D based on a_best
for all D_v do
   Tree_v = C4.5 (D_v)
   Attach Tree_v to the corresponding branch of Tree
end for
return Tree
```

2.2 Datasets

In this study, the student dataset includes biographical, academic portfolios, course duration, and student participation in the organization's activities has been used. The data were collected from academic information system (AIS) in 2014–2017 (279 samples data). Before training, all datasets will be normalization by using cleaning, integration and transformation, Fig. 1. First, cleaning process; total data collected of 459, then 180 data have been cleaned because some attribute value uncompleted. Second, integration and transformation process; total attribute value of 15, then 11 attribute have been applied in order to reduce and integrated unconditional attributes.

Furthermore, the performance of Tree C4.5 algorithm is measured by using the confusion matrix (CM) in which the true positive rate (TP) has been applied. Then, Rapid Miner Studio 7.3 software for the process of calculation and modeling has been used (Table 1).

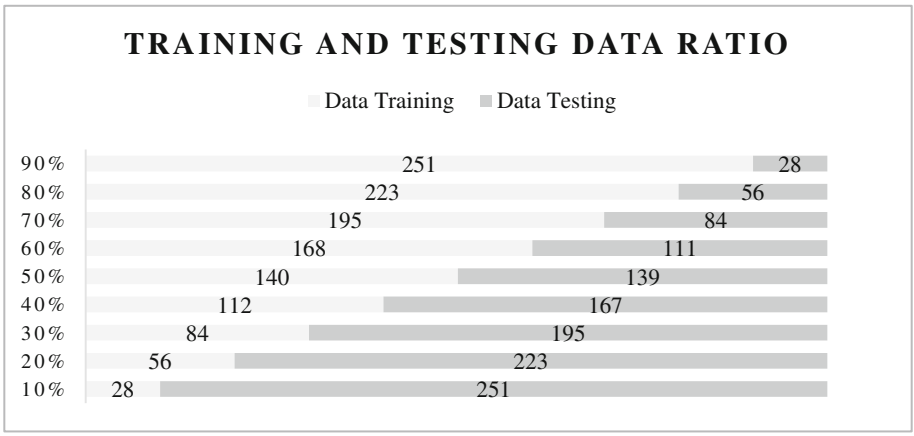


Fig. 1. Distribution of training data

Table 1. Data attribute after integration and transformation

No.	Attributes	Scale	Argument
1	Sex	Nominal	Male, Female or (M, F)
2	Age	Ordinal	Student age
3	Place of birth	Nominal	Town, Village
4	School status	Nominal	State, Private
5	School program	Nominal	Science, Non-science
6	GPA semester 1	Ordinal	1 (GPA $\leq$ 1.5)
7	GPA semester 2	Ordinal	2 (1.5 < GPA $\leq$ 2.5)
8	GPA semester 3	Ordinal	3 (2.5 < GPA $\leq$ 3.5)
9	GPA semester 4	Ordinal	4 (3.5 < GPA $\leq$ 4.0)
10	Organization	Nominal	Activist, Non-activist
11	Graduation time	Ordinal	Delay-time (>4,6 years) On-time (4–4,6 years) Fast-time (<4 years)

2.3 Performance of Evaluation

In this study, confusion matrix (CM) and with true positive rate (TP) for evaluation of Tree C4.5 model have been implemented. Where, CM is a matrix of prediction that will be compared with the original class of input, Tables 2 and 3. In other words, the matrix contains the actual value information and predictions on the classification [15]. Then, the equation of the accuracy (AC) measurement is shown as follows, where, *AC* is accuracy percentage proportion of predictions correct number; *a* is the exact number of predictions for the "*Fast-Time*" graduation; *b* is the exact number of predictions for the "*On-Time*" graduation; *c* is the exact number of predictions for the "*Delay-Time*" graduations; and *N* is total training data.

$$AC = \frac{a+b+c}{N} \tag{3}$$

Where, a is correct number of predictions, that negative instance; b is wrong number of predictions, that negative instance; c is wrong number of predictions, that positive instance; and d is correct number of predictions, that negative instance.

Table 2. Confusion matrix 2 class

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Source: [15]

Table 3. Confusion matrix of Tree C4.5 algorithm

Confusion matrix	Time		
	Fast	On	Delay
Fast-time	115	27	8
On-time	19	11	46
Delay-time	6	10	9

In this study, total course subject has been used as a student academic evaluation in Year 1, 2, and 3. In other words, student will be through the next level by this evaluation. The student evaluation term can be seen in Table 4.

Table 4. Student evaluation term

Evaluation	Degree	
Year I	Total course subject	24
	Total GPA	2,00
Year II	Total course subject	48
	Total GPA	2,00
Year III	Total course subject	72
	Total GPA	2,00

Meanwhile, true positive rate (TP) is also implemented for measured training data of Tree C4.5 model. The formula of TP as follows.

$$TP = \frac{\sum_{i=1}^{3} \frac{a_i}{n_i}}{3} \tag{4}$$

Where, *TP* is a percentage of predictions correct number; a_i is the exact number of predictions for the "*fast, on, delay*" graduation time; n_i is total of training data for the "*fast, on, delay*" graduation time. In this study, Receiver Operating Characteristic (ROC) was not chosen for evaluation the model because ROC analysis is particularly useful for threshold selection of CM and TP. Furthermore, in this study, analysis stages using Tree C4.5 algorithm is shown in Fig. 1.

3 Experimental and Results

This section describes the test of student academic evaluation variables using Tree C4.5 models. Based on predetermined rules, nine training and testing classes' dataset have been established. In this experiment, the dataset among others students' academic

Table 5. Training and testing dataset

Confusion matrix	Training data		
	Fast-time	On-time	Delay-time
10%	101	21	4
	20	14	41
	19	13	18
20%	78	13	8
	32	18	24
	15	11	24
30%	72	10	3
	22	15	6
	15	12	40
40%	63	8	2
	26	17	12
	5	7	28
50%	47	5	3
	14	9	19
	17	12	13
60%	43	8	6
	13	9	2
	6	4	20
70%	35	8	5
	9	6	2
	3	2	14
80%	24	5	3
	3	5	0
	4	1	11
90%	14	3	1
	1	2	0
	1	0	6

performance evaluation variables including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and *pGPA* and *tGPA* from semester 1–4. Furthermore, 10% to 90% of CM as a quality training data has been explored. Meanwhile, in order to get the best accuracy, CM as a performance of Tree C4.5 algorithm by using three times criteria (i.e., *fast*, *on*, and *delay* times) has been utilized, Table 5.

Based on the experiment conducted, the CM of Tree C4.5 algorithm shows that 78.57% AC with 76.72% TP of 90% quality training data have best accuracy value. It means that the best accuracy of Tree C4.5 algorithm is obtained when using 90% of training data ratio as shown in Table 6 and Fig. 2.

Table 6. Confusion matrix and true positive rate of Tree C4.5 algorithm

Algorithm evaluation		Algorithm Accuracy (AC)	True Positive rate (TP)
Training data ratio	10%	52.99%	44.94%
	20%	53.81%	50.37%
	30%	65.13%	59.76%
	40%	64.29%	62.40%
	50%	49.64%	45.95%
	60%	64.86%	59.87%
	70%	65.48%	60.63%
	80%	71.43%	68.75%
	90%	**78.58%**	**76.72%**

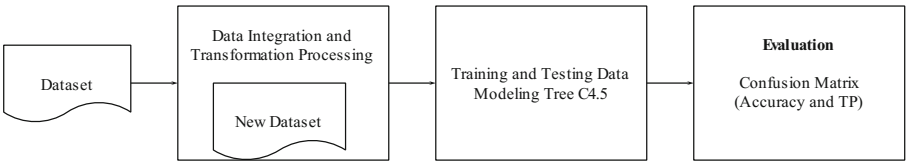


Fig. 2. Analysis stages of Tree C4.5 algorithm

The best performance of entropy and gain by using training data of 251 has been established. Where, GPA semester 4 as an initial node (root) has been settled. Detailed of entropy and gain values with 90% training data can be seen in Table 7. Based on Table 7, the highest gain value for the initial node of the manual calculation on the GPA semester 4 variables with 1.019 has been found. In other words, the initial node has corresponded with modeling (Fig. 3).

Table 7. Entropy and gain values with 90% training data

Root	Total graduation	Fast-time	On-time	Delay-time	Entropy	Gain
	251	140	48	63	1,43	
Sex						0,94
M	182	97	38	47	0,49	
F	69	43	10	16	0,47	
Age						0,96
16	1	0	1	0	0	
17	36	22	4	10	0,43	
18	156	83	35	38	0,49	
19	46	26	7	13	0,46	
20	5	3	1	1	0,49	
…	…	…	…	…	…	
23	2	2	0	0	0	
Place of birth						0,94
Town	107	59	22	26	0,49	
Village	144	81	26	37	0,48	
School status						0,95
State	193	101	43	49	0,49	
Private	58	39	5	14	0,41	
School program						0,94
Science	173	111	29	33	0,48	
Non-Science	78	29	19	30	0,48	
Organization						0,95
Activist	103	47	19	37		
Non-Activist	148	93	29	26		
GPA Sem. 1						0,96
1	0	0	0	0	0	
2	0	0	0	0	0	
3	170	73	42	55	0,50	
4	81	67	6	8	0,38	
GPA Sem. 2						1,01
1	0	0	0	0	0	
2	0	0	0	0	0	
3	156	56	39	61	0,49	
4	95	84	9	2	0,28	
GPA Sem. 3						0,99
1	0	0	0	0	0	
2	0	0	0	0	0	
3	173	74	38	61	0,49	
4	78	66	10	2	0,31	

(*continued*)

Table 7. (*continued*)

Root	Total graduation	Fast-time	On-time	Delay-time	Entropy	Gain
	251	140	48	63	1,43	
GPA Sem. 4						1,02
1	0	0	0	0	0	
2	0	0	0	0	0	
3	135	43	32	60	0,49	
4	116	97	16	3	0,31	

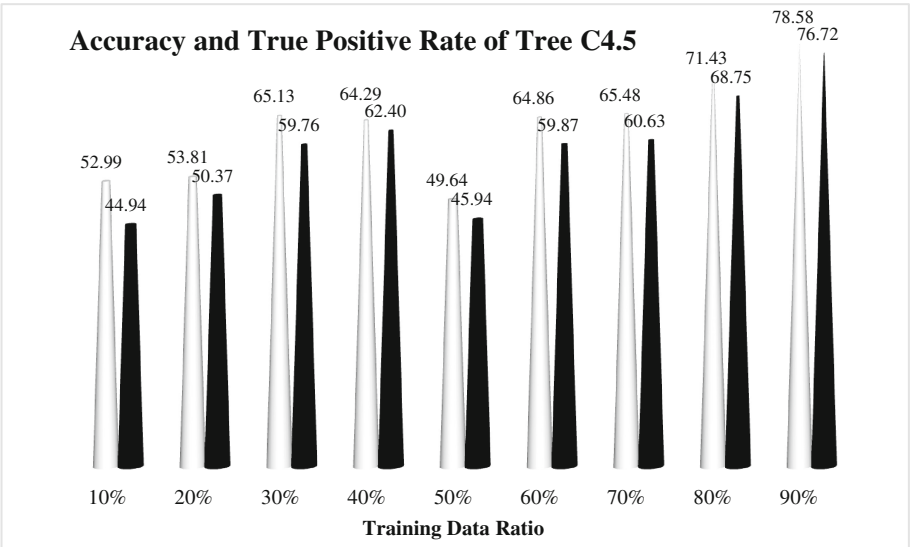


Fig. 3. Graphic of confusion matrix and true positive rate of Tree C4.5 algorithm

4 Conclusion

This paper has presented the Tree C4.5 algorithm in order to evaluate students' academic performance. Based on experiment, particular variables includes student activeness in the organization (activist and non-activist), place of birth, and age have been influence in student academic performance. This study indicated that Tree C4.5 algorithm have an accuracy better on evaluate students' academic performance. In other words, Tree C4.5 algorithm could be applied as an alternative model in student academic evaluation. Therefore, one of the planned future works is to implement Naïve Bayes Classifier (NBC), K-Means Cluster and Support Vector Machine (SVM) algorithms in order to get the better accuracy performance.

References

1. Ktona, A., Xhaja, D., Ninka, I.: Extracting relationships between students' academic performance and their area of interest using data mining techniques. In: 2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks. IEEE (2014)
2. Xu, B., et al.: Clustering educational digital library usage data: a comparison of latent class analysis and K-means algorithms. J. Educ. Data Min. **5**(2), 38–68 (2013)
3. Mahboob, T., Irfan, S., Karamat, A.: A machine learning approach for student assessment in e-learning using Quinlan's C4.5, Naïve Bayes and random forest algorithms. IEEE (2016)
4. Lakshmi, B.N., Indumathi, T.S., Ravi, N.: A study on C.5 decision tree classification algorithm for risk predictions during pregnancy. In: International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST-2015), Procedia Technology (2016)
5. Haviluddin, et al.: Modelling of network traffic usage using self-organizing maps techniques. In: 2016 2nd International Conference on Science in Information Technology (ICSITech). IEEE (2016)
6. Haviluddin, et al.: A performance comparison of statistical and machine learning techniques in learning time series data. Adv. Sci. Lett. **21**(10), 3037–3041 (2015)
7. Purnawansyah, Haviluddin: K-means clustering implementation in network traffic activities. In: 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia. IEEE (2016)
8. Pandey, M., Taruna, S.: Towards the integration of multiple classifier pertaining to the student's performance prediction. Perspect. Sci. **2016**(8), 364–366 (2016)
9. Yunianta, A., et al.: Data mapping process to handle semantic data problem on student grading system. Int. J. Adv. Intell. Inform. (IJAIN) **2**(3), 157–166 (2017)
10. Dangi, A., Srivastava, S.: Educational data classification using selective Naïve Bayes for quota categorization. In: 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE). IEEE (2014)
11. Guarín, C.E.L., Guzmán, E.L., González, F.A.: A model to predict low academic performance at a specific enrollment using data mining. IEEE Rev. Iberoam. Tecnol. Aprendiz. **10**(3), 119–125 (2015)
12. Amaya, Y., Barrientos, E., Heredia, D.: Student dropout predictive model using data mining techniques. IEEE Lat. Am. Trans. **13**(9), 3127–3134 (2015)
13. Kotsiantis, S., Patriarcheas, K., Xenos, M.: A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowl. Based Syst. **23**, 529–535 (2010)
14. Jiawei, H., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Ltd., San Francisco (2001)
15. Gorunescu, F.: Data Mining. Intelligent Systems Reference Library, vol. 12. Springer, Craiova (2011)

Computing Complex Roots of Systems of Nonlinear Equations Using Spiral Optimization Algorithm with Clustering

Kuntjoro Adji Sidarto and Adhe Kania(✉)

Department of Mathematics, Institut Teknologi Bandung, Bandung, Indonesia
{sidarto,adhe.kania}@math.itb.ac.id

Abstract. Finding complex roots of a system of nonlinear equations is not an easy numerical computation problem. A method of locating and finding all real and complex roots of systems of nonlinear equations in a single run is proposed here. The method that was first proposed for finding all real roots of systems of nonlinear equations is now slightly modified and adapted so that it can be used also for finding complex roots of the corresponding system. The root finding problem is transformed to optimization problem and then a spiral optimization algorithm of Tamura and Yasuda is used to solve the optimization problem. In order to locate the position of the roots, we proposed a certain clustering technique. Several test problems have been examined. This combination of technique enables ones to locate and find all real and complex roots within a bounded domain in all test cases.

Keywords: Systems of nonlinear equations
Real and complex roots finding problem
Spiral optimization algorithm · Clustering

1 Introduction

Many problems in the real world can be modeled in the systems of nonlinear equations form. Deterministic methods such as Newton and quasi-Newton methods are commonly used for solving the problem because of their speed of convergence once a sufficiently accurate initial approximation to the root is known. However, the convergence of these methods cannot be ensured if the accurate initial approximation to the root is not provided. Metaheuristic methods that are initially proposed for solving optimization problem can also be used for solving systems of nonlinear equations by first converting the problem as an optimization problem and then find the root as a point that solve the optimization problem. However, these techniques are only able to find a single root of a system of nonlinear equations at single run of the algorithm.

The problem of finding all real roots of systems of nonlinear equations based on metaheuristic optimization algorithms have been proposed on several recent articles, such as [2–5]. The problem of finding all real and complex roots also have been proposed on several recent articles, such as [6, 7].

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 390–398, 2018.
https://doi.org/10.1007/978-981-10-8276-4_37

This paper focuses on locating all real and complex roots of systems of nonlinear equations using a combination of certain clustering technique with Spiral Optimization Algorithm.

2 Problem Description

A standard form of system of nonlinear equations is

$f_1(x_1, x_2, \ldots, x_n) = 0, f_2(x_1, x_2, \ldots, x_n) = 0, \ldots, f_n(x_1, x_2, \ldots, x_n) = 0,$ with $(x_1, x_2, \ldots, x_n) \in D = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \subset \mathbb{R}^n$. $f_i : D \to \mathbb{R}, i = 1, 2, \ldots, n$ is continuous functions with one or more nonlinear functions. The vector form of this system can be written as $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ where $\mathbf{f} = (f_1, f_2, \ldots, f_n)^T$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. A vector $\mathbf{x}^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)^T \in D$ where $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$ is the solution of the system.

The system of nonlinear equations can be solved by optimization methods [1]. A vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is the root of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ precisely when the function: $F(\mathbf{x}) = F(x_1, x_2, \ldots, x_n) = \frac{1}{1 + \sum_{i=1}^{n} |f_i(x_1, x_2, \ldots, x_n)|} = \frac{1}{1 + \sum_{i=1}^{n} |f_i(\mathbf{x})|}$ has the maximal value 1. Hence, finding all $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = 1$ corresponds to locating all the roots of the system $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. It suggests that global optimization methods can be used to find the solution of system of nonlinear equations.

3 Spiral Optimization Algorithm

Spiral Optimization Algorithm (SOA) [8] is inspired by spiral phenomena. In 2-D spiral model, a vector $\mathbf{x}$ in $\mathbb{R}^2$ can be rotated at the origin to $\mathbf{x}'$ with formula $\mathbf{x}' = rR^{(2)}\mathbf{x}$ where $0 < r < 1$ and $R^{(2)} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ is a rotation matrix with rotation angle $0 < \theta < 2\pi$. The mapping $\mathbf{x}(k+1) = r\,R^{(2)}\,\mathbf{x}(k), k = 0, 1, 2, \ldots$ will produce a sequence of vectors $\mathbf{x}(0), \mathbf{x}(1), \mathbf{x}(2), \ldots$ which converge to the origin along a trajectory of spiral form. We can use an arbitrary point $\mathbf{x}^*$ as the center and rotated $\mathbf{x}$ by spiral model with the formula $\mathbf{x}' = S_2(r, \theta)\mathbf{x} - (S_2(r, \theta) - I_2)\mathbf{x}^*$ where $S_2(r, \theta) = rR^{(2)}$.

The extension to n-Dimensional spiral model may be obtained with $R^{(n)}$ being an $n \times n$ matrix defined as $R^{(n)} = \prod_{i=1}^{n-1} \left(\prod_{j=1}^{i} R_{n-i,n+1-j}^{(n)} \right)$ where $R_{i,j}^{(n)}$ is $n \times n$ matrix with entries $r_{ii} = r_{jj} = \cos\theta, r_{ji} = \sin\theta, r_{ij} = -\sin\theta$ and $r_{st} = \delta_{st}$ for all other entries of $R_{i,j}^{(n)}$ (where $\delta_{st} = 1$ if $s = t$ and $\delta_{st} = 0$ if $s \neq t$). Thus $R^{(n)}$ is a composition of plane rotations matrix $R_{i,j}^{(n)}$. Using this composition rotation matrix, the n-Dimensional spiral model is formulated similar with the formula for 2-D spiral model.

For the maximization problem: $\underset{\mathbf{x} \in \mathbb{R}^n}{maximize}\ F(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, the Spiral Optimization Algorithm (SOA) can obtain the optimal value by updated the new

value of $\mathbf{x}$ in each iteration as $\mathbf{x}_i(k+1) = S_n(r,\theta)\mathbf{x}_i(k) - (S_n(r,\theta) - I_n)\mathbf{x}^*$, where $x^* = x_{i_g}(k+1), i_g = \arg\max_i F(x_i(k+1)),\ i = 1,2,\ldots,m.$

4 A Clustering Technique for Roots Finding Problem

As described in Sect. 2, finding the root of system of non linear equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ can be obtained by finding the maximum point of the function $F(\mathbf{x}) = \left(1 + \sum_{i=1}^{n} |f_i(\mathbf{x})|\right)^{-1}$. We can use SOA to find the maximum point of $F(\mathbf{x})$. In general, the nonlinear system $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ may have many roots in the specified domain, while a single run of SOA is only capable of obtaining a single maximum point of $F(\mathbf{x})$. To obtain the other roots, a certain clustering technique that was recently proposed [1] combined with SOA is capable of finding as many global maximum points of $F(\mathbf{x})$ as possible which correspond to many roots of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, in a single run.

The algorithm to get many roots of system of nonlinear equation need input $m_{cl}, r_{cl}, \theta_{cl}, k_{cl}$ as input parameters for SOA at diversification phase, $\gamma\ (0<\gamma<1)$ as a 'cut-off' parameter for function value $F(\mathbf{x})$, $\varepsilon\ (0<\varepsilon<1)$ as parameter for roots acceptance, $\delta(0<\delta<1)$ as parameter to distinguish between one candidate root and another one in case they are very close each other, and $m, r, \theta, k_{\max}$ as parameters for SOA at intensification phase. The process of algorithm can be described as follows.

1. Generate Sobol sequence of points $\mathbf{x}_i(0) \in \mathbb{R}^n, i = 1,2,\ldots,m_{cl}$ in the feasible region D, where $D = [a_1,b_1] \times [a_2,b_2] \times \ldots \times [a_n,b_n] \subset \mathbb{R}^n$ and set $k = 0$.
2. Set $\mathbf{x}'$ as $\mathbf{x}' = \mathbf{x}_{ig}(0),\ i_g = \arg\max_i F(\mathbf{x}_i(0))\ i = 1,2,\ldots,m_{cl}$.
3. Store $\mathbf{x}'$ as center of the first cluster with radius $\frac{1}{2}\left(\min_l |b_l - a_l|\right),\ l = 1,2,\ldots,n$
4. For $i = 1,2,\ldots,m_{cl}$ do

If $F(\mathbf{x}_i) > \gamma$ and $\mathbf{x}_i$ is not the center of existing cluster, then $\mathbf{x}_i$ is possible to become a center of new cluster, and do the following function cluster with input $\mathbf{x}_i$:

Function Cluster (input: $\mathbf{y}$)

(a) Find a cluster with the center nearest to $\mathbf{y}$. Let C be that cluster with center $\mathbf{x}_C$.
(b) Set $\mathbf{x}_t$ as mid-point between $\mathbf{y}$ and $\mathbf{x}_C$.
(c) Compare $F(\mathbf{y})$, $F(\mathbf{x}_C)$, and $F(\mathbf{x}_t)$:

- If $F(\mathbf{x}_t) < F(\mathbf{y})$ and $F(\mathbf{x}_t) < F(\mathbf{x}_C)$: Set a new cluster with center at $\mathbf{y}$ and radius equal to the distance between $\mathbf{y}$ and $\mathbf{x}_t$.
- Else, if $F(\mathbf{x}_t) > F(\mathbf{y})$ and $F(\mathbf{x}_t) > F(\mathbf{x}_C)$: Set a new cluster with $\mathbf{y}$ as its centre and radius equal to the distance between $\mathbf{y}$ and $\mathbf{x}_t$. Redo Function Cluster with $\mathbf{x}_t$ as its input.
- Else, if $F(\mathbf{y}) > F(\mathbf{x}_C)$, set $\mathbf{y}$ as the centre of C.

(d) Change the radius of C equals to the distance between $\mathbf{y}$ and $\mathbf{x}_t$.

5. Set $\mathbf{x}_p = \mathbf{x}_{i_g}$ where $i_g = \arg\max_i F(\mathbf{x}_i(k)), i = 1,2,\ldots,m_{cl}$
6. Update $\mathbf{x}_i$: $\mathbf{x}_i(k+1) = S_n(r_{cl},\theta_{cl})\mathbf{x}_i(k) - (S_n(r_{cl},\theta_{cl}) - I_n)\mathbf{x}_p,\ i = 1,2,\ldots,m_{cl}$

Do steps 4 to 6 k_{cl} times, and after that, we have a number of clusters. Each cluster has its center and radius. To each cluster, perform SOA to obtain a candidate of root in each cluster. Use $m, r, \theta,$ *and* $k_{\max}$ as SOA input in this phase. Keep only candidate roots which satisfy condition $1 - F(\mathbf{x}) < \varepsilon$.

Suppose after the previous step, there are n_g candidate roots. From these candidates, select only those in which the distance between the candidates is more than δ. In case the distance is less than δ, select only one root that have the larger function value.

Note that we have used Sobol sequence of points instead of pseudo-random points in step 1 above, since generated initial population of points randomly may not uniformly distribute in the search feasible region of the problem [9]. This paper applies algorithm developed by Joe and Kuo [10] to construct Sobol sequence of points, which does not involve generation of pseudo-random points, as initial population of points. This implies that the solution generated in this paper always produce the same result for the same input. Hence, we can focus on getting the best input to get the result without running it many times to one input.

5 Finding the Real Roots

In order to verify the technique, two test cases from benchmark problems have been examined to get the real roots of system of nonlinear equations. In this study, all the numerical experiments were performed on a Notebook equipped with processor Intel Core™ i5 with 4 GB ram and 1.6 GHz CPU running Ubuntu Linux 12.04. The code was written in C++ and compiled using g++.

Test Problem 1. The system of equations is defined as follow [5]:

$$f_1(x_1, x_2) = x_1 - \cos(4\pi x_2) = 0, f_2(x_1, x_2) = x_1^2 + x_2^2 = 1,$$
$$\text{with } D = \{(x_1, x_2) : -1 \le x_1 \le 1, -1 \le x_2 \le 1\}$$

A single run of the clustering technique was performed with parameters: $m_{cl} = 250$, $k_{cl} = 10$, $r_{cl} = 0.99$, $\theta_{cl} = \pi/16$, $\gamma = 0.6$, $\varepsilon = 10^{-7}$, $\delta = 10^{-3}$, $m = 250$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 250$ and obtain the 15 results that appear as the intersection of the graph $f_1(x_1, x_2) = 0$ and $f_2(x_1, x_2) = 0$. Their positions can be seen in Fig. 1. Here we obtained simultaneously all the 15 distinct roots in a single run which took 1.61 s.

Test Problem 2. The Weierstrass function defined as follow [7]:

$$g(x) = \sum_{n=1}^{N \to \infty} \lambda^{(s-2)k} \sin\left(\lambda^k x\right), \text{ where } 1 < s < 2 \text{ and } \lambda > 1$$

This function is known as a function which is continuous everywhere but differentiable nowhere. Here, we consider the truncated form of the above function for $N = 20$ with $s = 1.1$ and $\lambda = 1.5$. To find the real roots of $g(x) = \sum_{k=1}^{20} 1.5^{-0.9k} \sin\left(1.5^k x\right) = 0$, $0 \le x \le 5$, a single run of the clustering technique was performed with parameters:

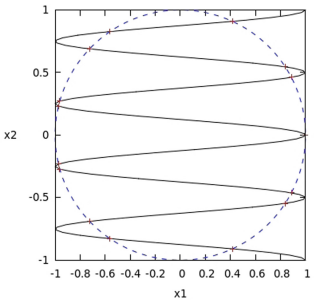


Fig. 1. The graph of $f_1(x_1, x_2) = 0$ and $f_2(x_1, x_2) = 0$ for problem 1

$m_{cl} = 200$, $k_{cl} = 50$, $r_{cl} = 0.99$, $\theta_{cl} = \pi/8$, $\gamma = 0.9$, $\varepsilon = 10^{-7}$, $\delta = 10^{-4}$, $m = 150$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 150$ and the results are presented in Table 1. The root positions can be seen in Fig. 2. Here we obtained simultaneously all the 9 distinct roots in a single run which took 9.89 s, in which the 3rd to 7th roots are very close each other. Hence the proposed algorithm is capable of finding the closely located roots.

Table 1. Results for problem 2.

Solution	x	$g(x)$	Solution	x	$g(x)$
1	0	0	6	3.73962	9.30683e−08
2	1.88871	−9.77192e−08	7	3.74071	9.10978e−08
3	3.73173	−9.0547e−08	8	4.54986	−9.4056e−08
4	3.73499	−9.69124e−08	9	5.01996	9.56515e−08
5	3.73819	3.67608e−08			

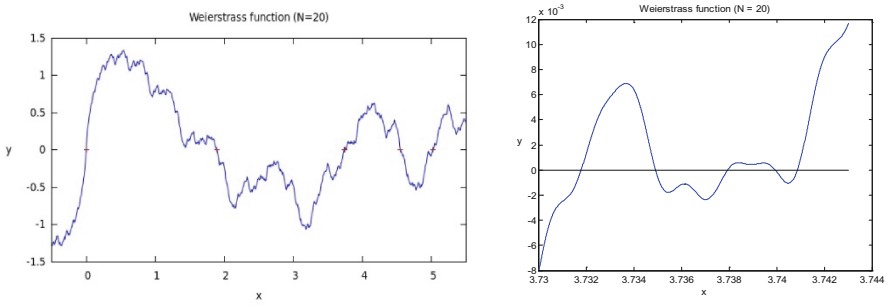


Fig. 2. The graph of $y = g(x)$ in the intervals [−0.5, 5.5] (left) and [3.73, 3.744] (right)

6 Finding the Complex Roots

A complex number z can be written as $z = u + vi$, where u is its real part and v is its imaginary part. Hence, the complex function can be written as $f(z) = f(u + vi)$. For some simple functions, their real and imaginary part can be written easily. For instance, consider the following system:

$$\begin{bmatrix} f_1(z_1, z_2) \\ f_2(z_1, z_2) \end{bmatrix} = \begin{bmatrix} z_1^2 + z_2^2 + z_1 + z_2 - 8 \\ z_1 z_2 + z_1 + z_2 - 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1}$$

Notice that:

$$\begin{aligned} f_1(z_1, z_2) &= z_1^2 + z_2^2 + z_1 + z_2 - 8 = (u_1 + v_1 i)^2 + (u_2 + v_2 i)^2 + (u_1 + v_1 i) + (u_2 + v_2 i) - 8 \\ &= \left(u_1^2 - v_1^2 + u_2^2 - v_2^2 + u_1 + u_2 - 8\right) + (2u_1 v_1 + 2u_2 v_2 + v_1 + v_2) i \end{aligned} \tag{2}$$

$$\begin{aligned} f_2(z_1, z_2) &= z_1 z_2 + z_1 + z_2 - 5 \;=\; (u_1 + v_1 i)(u_2 + v_2 i) + (u_1 + v_1 i) + (u_2 + v_2 i) - 5 \\ &= (u_1 u_2 + u_1 + u_2 - v_1 v_2 - 5) + (u_1 v_2 + u_2 v_1 + v_1 + v_2) i \end{aligned} \tag{3}$$

Hence, finding the complex roots of system (1) is equivalent to find the real roots of system below:

$$\mathbf{g}(\mathbf{x}) \;=\; \begin{bmatrix} g_1(x_1, x_2, x_3, x_4) \\ g_2(x_1, x_2, x_3, x_4) \\ g_3(x_1, x_2, x_3, x_4) \\ g_4(x_1, x_2, x_3, x_4) \end{bmatrix} = \begin{bmatrix} x_1^2 - x_2^2 + x_3^2 - x_4^2 + x_1 + x_3 - 8 \\ 2x_1 x_2 + 2x_3 x_4 + x_2 + x_4 \\ x_1 x_3 + x_1 + x_3 - x_2 x_4 - 5 \\ x_1 x_4 + x_3 x_2 + x_2 + x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4}$$

where x_1 represents the real part of z_1, x_2 represents the imaginary part of z_1, x_3 represents the real part of z_2, and x_4 represents the imaginary part of z_2.

To search for the complex roots for the system (1) in the region $D = \{x_i : -10 \le x_i \le 10, i = 1, \ldots, 4\}$, we use parameters: $m_{cl} = 1000$, $k_{cl} = 60$, $r_{cl} = 0.99$, $\theta_{cl} = \pi/32$, $\gamma = 0.1$, $\varepsilon = 10^{-7}$, $\delta = 0.1$, $m = 300$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 300$, and obtained the results as solution of system (4) shown in Table 2.

Table 2. The real roots for system (4) with four variables.

Sol	x_1	x_2	x_3	x_4	$g_1(\mathbf{x})$	$g_2(\mathbf{x})$	$g_3(\mathbf{x})$	$g_4(\mathbf{x})$
1	2	−5.3339e−09	1	1.23297e−08	−4.5118e−08	1.032e−08	1.7347e−08	2.6321e−08
2	1	1.17625e−08	2	−6.96105e−09	−8.7564e−09	4.8227e−10	5.2617e−08	2.1365e−08
3	−3	−1.41421	−3	1.41421	−1.2475e−08	−4.9283e−08	−2.3663e−08	1.2819e−08
4	−3	1.41421	−3	−1.41421	−3.7987e−08	−1.3877e−08	7.8744e−09	1.4025e−08

The results in Table 2 show that the system (1) has essentially two real roots and two complex roots which are mutually conjugate: $\mathbf{z}_1 = (2+0i, 1+0i)$, $\mathbf{z}_2 = (1+0i, 2+0i)$, $\mathbf{z}_3 = (-3-1.41421i, -3+1.41421i)$, $\mathbf{z}_4 = (-3+1.41421i, -3-1.41421i)$.

However, not all complex functions can be easily written explicitly its real part and its imaginary part as above. But we can still find the complex roots of system of nonlinear equation by using similar algorithm as for finding the real roots. By using *complex.h* library that has been provided by C++ standard platform, it is possible to find the complex roots with the clustering technique described in Sect. 4. Each variables that were previously initiated as *double* which is a real number, to be changed to *complex<double>*.

7 Numerical Experiments

In order to verify the proposed technique, several test cases from various benchmark problems have been examined.

Test Problem 1. The system of equations considered in [6] is defined as follows, together with our chosen domain:

$$f_1(x_1, x_2) = x_1^4 + 4x_2^4 - 6 = 0\,, f_2(x_1, x_2) = x_1^2 x_2 - 1.6787 = 0,$$
$$\text{with } D = \{(x_1, x_2) : -2 \le x_1 \le 2, -2 \le x_2 \le 2\}$$

A single run of the clustering technique was performed with parameters: $m_{cl} = 200$, $k_{cl} = 50$, $r_{cl} = 0.99$, $\theta_{cl} = \pi/8$, $\gamma = 0.9$, $\varepsilon = 10^{-7}$, $\delta = 10^{-4}$, $m = 150$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 150$ and the results are presented in Table 3. Here we obtained simultaneously all the 12 distinct roots in a single run which took 58.82 s. All results are similar with those reported in [6].

Table 3. Results for problem 1.

Sol	x_1	x_2	$g_1(\mathbf{x})$	$g_2(\mathbf{x})$
1	1.43098 − 5.08152e−05i	0.819816 + 6.53159e−05i	−2.6359e−05 − 1.9781e−05i	4.4030e−05 + 1.4521e−05i
2	−1.4311 + 2.82656e−05i	0.819663 + 3.8692e−05i	3.7359e−05 + 9.5324e−06i	1.3268e−05 + 1.2931e−05i
3	1.39602 − 4.59531e−05i	0.861362 + 5.32735e−05i	−4.7551e−06 + 4.4651e−05i	−2.3060e−05 − 6.6922e−06i
4	−1.39615 − 8.3548e−05i	0.861226 − 9.40931e−05i	2.5552e−05 − 5.2203e−05i	2.5856e−05 + 1.7507e−05i
5	0.840215 + 0.840215i	1.00202e−07 − 1.18897i	−1.0749e-05 + 7.1284e−06i	2.779e−05 − 1.7253e−06i
6	0.840205 − 0.840206i	−1.62053e−06 + 1.18896i	−1.6171e-05 + 4.6572e−05i	−1.4268e−05 + 1.0281e−06i
7	−0.840203 + 0.84022i	3.68565e−06 + 1.18896i	−6.6519e−06 − 2.0486e−05i	1.2003e−05 − 3.8311e−05i
8	−0.840214 − 0.840218i	1.17022e−07 − 1.18897i	−9.5319e−06 − 1.4893e−05i	3.2477e−05 + 7.7607e−06i
9	−8.67857e−05 − 1.43107i	−0.819693 − 0.000114517i	−6.6203e−05 − 8.2757e−06i	1.1955e−06 + 3.0921e−05i
10	0.000159901 + 1.43101i	−0.819779 − 0.000211417i	−7.1804e−07 − 1.073e−05i	3.8877e−05 + 5.7773e−05i
11	1.8196e−05 − 1.39627i	−0.861096 + 1.63072e−05i	−5.6278e−06 + 3.1534e−05i	5.8816e-05 + 1.1963e−05i
12	6.66309e−07 + 1.39611i	−0.861273 + 2.27387e−06i	4.8289e−05 − 3.0496e−05i	1.6184e−05 − 6.0344e−06i

Test Problem 2. The system of equations considered in [6] is defined as follows, together with our chosen domain:

$$f_1(x_1,x_2) = e^{x_1} - e^{x_2} + 2 = 0\,, f_2(x_1,x_2) = x_1^3 - x_2^3 - 1 = 0,$$
$$\text{with } D = \{(x_1,x_2) : -5 \le x_1 \le 5, -5 \le x_2 \le 5\}$$

A single run of the clustering technique was performed with parameters: $m_{cl} = 2000$, $k_{cl} = 20$, $r_{cl} = 0.95$, $\theta_{cl} = \pi/4$, $\gamma = 0.01$, $\varepsilon = 10^{-5}$, $\delta = 0.01$, $m = 500$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 500$ and the results are presented in Table 4. Here we obtained simultaneously all the 6 distinct roots in a single run which took 380.09 s.

Table 4. Results for problem 2.

Sol	x_1	x_2	$g_1(\mathbf{x})$	$g_2(\mathbf{x})$
1	−0.662665 + 0.956196i	0.848236 + 0.181314i	1.96347e−06 − 4.71156e−07i	4.69593e−07 + 9.11494e−07i
2	−0.662666 − 0.956197i	0.848238 − 0.181314i	−2.88996e−06 + 2.18101e−06i	7.28555e−07 + 5.90633e−07i
3	0.680596 − 3.16038i	−3.10316 + 0.972481i	6.59986e−07 + 2.35274e−06i	−9.76271e−07 + 4.93106e−06i
4	0.680596 + 3.16038i	−3.10316 − 0.97248i	2.41556e−07 + 2.33136e−07i	9.89233e−07 − 9.29327e−06i
5	1.98471 − 3.05201i	1.66544 + 3.26502i	−1.31746e−06 + 1.53558e−06i	−3.36184e−06 + 2.4365e−06i
6	1.98471 + 3.05201i	1.66544 − 3.26502i	2.25969e−06 − 1.25927e−07i	2.29536e−06 + 6.32589e−06i

Test Problem 3. The system of equations considered in [11] is defined as follows, together with our chosen domain:

$$f_1(x_1,x_2) = e^{x_1 - x_2} - \sin(x_1 + x_2) = 0, f_2(x_1,x_2) = x_1^2 x_2^2 - \cos(x_1 + x_2) = 0$$
$$\text{with } D = \{(x_1,x_2) : -10 \le x_1 \le 10, -10 \le x_2 \le 10\}$$

A single run of the clustering technique was performed with parameters: $m_{cl} = 20000$, $k_{cl} = 30$, $r_{cl} = 0.95$, $\theta_{cl} = \pi/4$, $\gamma = 0.01$, $\varepsilon = 10^{-5}$, $\delta = 0.01$, $m = 500$, $r = 0.95$, $\theta = \pi/4$, $k_{\max} = 500$ and obtained simultaneously 27 distinct roots in which 6 are real roots and 21 are complex roots in a single run which took 1413.37 s.

Test Problem 4. The system of equations considered in [12] and also in [13] is defined as follows, together with our chosen domain:

$$f_1(\mathbf{x}) = x_1 + \frac{x_2^2\, x_4\, x_6}{4} + 0.75 = 0, f_2(\mathbf{x}) = x_2 + 0.405\, e^{1 + x_1\, x_2} - 1.405 = 0,$$
$$f_3(\mathbf{x}) = x_3 - \frac{x_4\, x_6}{2} + 1.5 = 0, f_5(\mathbf{x}) = x_5 - \frac{x_2\, x_6}{2} + 1.5 = 0, f_6(\mathbf{x}) = x_6 - x_1\, x_5 = 0$$
$$\text{with } \mathbf{x} = (x_1, x_2, \ldots, x_6)^T \in \mathbb{R}^6 \text{ and } D = \{\mathbf{x} : -3 \le x_i \le 3, i = 1, 2, \ldots, 6\}$$

A single run of the clustering technique was performed with parameters: $m_{cl} = 10000$, $k_{cl} = 50$, $r_{cl} = 0.99$, $\theta_{cl} = \pi/64$, $\gamma = 0.1$, $\varepsilon = 10^{-3}$, $\delta = 0.5$, $m = 1000$, $r = 0.99$, $\theta = \pi/4$, $k_{\max} = 1000$, and obtained simultaneously 12 distinct roots in which 2 are real roots and 10 are complex roots in a single run which took 3423.33 s. The two references only reported one real root, that is: $\mathbf{x} = (-1, 1, -1, 1, -1, 1)^T$.

8 Conclusions

Combination of the proposed Clustering technique with SOA have been shown able effectively to locate and find all the real and complex roots of systems of nonlinear equations in various test cases considered, each in a single run, without a priori knowledge of the number of the roots. The use of Sobol sequence of points instead of pseudo-random points to generate initial populations of points in the feasible region may increase the potential capacity of SOA during diversification in the early phase to locate the potential positions of the candidate roots of the system of equations.

Acknowledgement. This work was partially supported by the P3MI Industrial and Financial Mathematics Research Group, Institut Teknologi Bandung, 2017.

References

1. Sidarto, K.A., Kania, A.: Finding all solutions of systems of nonlinear equations using spiral dynamics inspired optimization with clustering. J. Adv. Comput. Intell. Intell. Inf. **19**(5), 697–707 (2015)
2. Tsoulos, I., Stavrakoudis, A.: On locating all roots of systems of nonlinear equations inside bounded domain using global optimization methods. Nonlinear Anal. Real World Appl. **11**, 2465–2471 (2010)
3. Sacco, W., Henderson, N.: Finding all solutions of nonlinear systems using a hybrid meta-heuristic method with Fuzzy Clustering Means. Appl. Soft Comput. **11**, 5424–5432 (2011)
4. Grosan, C., Abraham, A.: A new approach for solving nonlinear equations systems. IEEE Trans. Syst. Man Cybern. **38**(3), 698–714 (2008)
5. Song, W., Wang, Y., Li, H.-X., Cai, Z.: Locating multiple optimal solutions of nonlinear equation systems based on multiobjective optimization. IEEE Trans. Evol. Comput. **19**(3), 414–431 (2015)
6. Bahrami, M., Oftadeh, R.: An effective iterative method for computing real and complex roots of systems of nonlinear equations. J. Appl. Math. Comput. **215**, 1813–1820 (2009)
7. Pourjafari, E., Mojallali, H.: Solving nonlinear equations systems with a new approach based on invasive weed optimization algorithm and clustering. Swarm Evol. Comput. **4**, 33–43 (2012)
8. Tamura, K., Yasuda, K.: Spiral dynamics inspired optimization. J. Adv. Comput. Intell. Intell. Inform. **15**(8), 1116–1122 (2011)
9. Seydel, R.: Tools for Computational Finance. Springer, Heidelberg (2002)
10. Joe, S., Kuo, S.: Constructing sobol sequences with better two dimensional projections. SIAM J. Sci. Comput. **30**, 2635–2654 (2008)
11. Chen, K., Giblin, P., Irving, A.: Mathematical Explorations with MATLAB. Cambridge University Press, Cambridge (1999)
12. Luo, Y.-Z., Tang, G., Zhou, L.-N.: Hybrid approach for solving systems of nonlinear equations using chaos optimization and quasi-Newton method. Appl. Soft Comput. **8**, 1068–1073 (2008)
13. Krzyworzcka, S.: Extension of the Lanczos and CGS methods to systems of nonlinear equations. J. Comput. Appl. Math. **69**(1), 181–190 (1996)

A Survey on Context-Aware Information Retrieval Research

Shaiful Bakhtiar bin Rodzman(✉), Normaly Kamal Ismail, and Nurazzah Abd Rahman

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor Darul Ehsan, Malaysia
shybug_2@yahoo.com, {normaly,nurazzah}@tmsk.uitm.edu.my

Abstract. Most of the retrieved documents from the *Information Retrieval* (IR) System are irrelevant to the user because the IR cannot determine the user's context. One of the main issues is that the relevancy of the retrieved documents is based on personal assessment that depends on the task to be done and its context. This paper provides the review of prior researches (2003–2016) and concludes the review by providing the summary of the research's current trends, future direction and opportunity and defining the research gap. First, the findings show that in prior studies, there is no identification of contextual aspect has been done in optimizing the ranking function of the Malay IR. Second, in optimizing the ranking function, the integration process of context representation and document ranking must be done. This approach also has not been done yet in the development of Malay Document Retrieval. If it still stays in the current status, the Malay Document Retrieval system cannot be improved compared to the traditional languages of Context Aware IR System (English).

Keywords: Context-Aware Information Retrieval · Malay Document Retrieval IR research

1 Introduction

As a summation, Information Retrieval (IR) focuses on the searching of the documents for information that satisfies a user's need. In traditional IR, the results of the searching process will be presented to the user in a form of a ranked list that contains the most relevant documents [38]. However, most of the documents that are retrieved in IR are irrelevant to the user because the search engines cannot determine the user's context [15]. Context can be employed from the dimension of user's prior knowledge or user's interest and it's understandable, that relevance can change with time, location, and size of the document and also what in the user's mind [6]. Ideally, the relevance of documents and the problem of ranking of the retrieved documents should be based on the user's context and preferences [15]. To apply context in IR for personal assessment, researchers have examined new technique that known as Contact Awareness, due to the existing personalization techniques that are typically applied out of context [15]. During the last decade, research on Context-Awareness technique has actively evolved, which takes the

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 399–409, 2018.
https://doi.org/10.1007/978-981-10-8276-4_38

advantage of recent developments in related fields like IR, Mobile Application, GeoInformatic, Computational Linguistics, Artificial Intelligence, and Soft Computing.

This paper provides the review of prior research (2003–2016) and conclude the review by providing the summary of the research's current trends, future direction, opportunity and defining the research gap. The rest of the paper is organized as follows: Sect. 2, presents the overview of Context Aware IR research. Section 3, explores the concept of Topic Model and its potential application with IR. Section 4, explores the Evolution of Malay IR and Sect. 5 provide the research gap and Sect. 6, draws a conclusion for this paper.

2 Framework for Context-Aware Information Retrieval

One of the primary questions on Context-Aware IR (CAIR) research is to determine which types of context that should be considered in the retrieval process? Knowing more about what features are important in a context and what they are used for, than can be helped in design more beneficial and successful IR systems. Referring to [15], the author has visualised the context in the Fig. 1 below into five specific dimensions, such as Device, User Context, Task/Problem, Document Context and Spatio-Temporal Context.

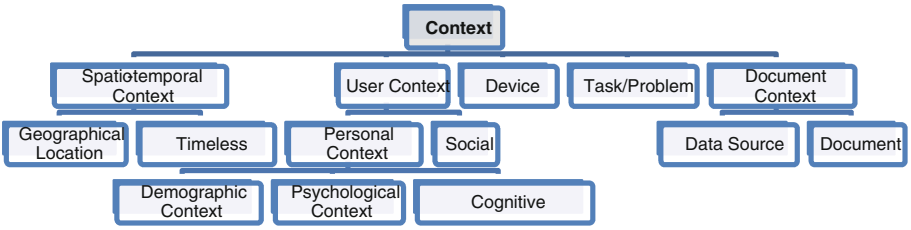


Fig. 1. The multi-faceted concepts of contextual IR, Source: [16]

The prior studies also shown various features were taken as the context in defining the contextual model and utilize it to improve the IR such as in the timeline shown in Fig. 2 and this section will elaborate more on this particular topic.

The first types of context model, we can see in development of Ontology Model on [3], in 2003 that utilised the Document Context. The author proposed the IR Framework with Ontology Model in English Semantic Web to improve the ranking of the IR. In evaluation part, the User Oriented Assessment Criterion have been done, due to the author has ranked the results based on a context specified by the user, and the evaluation criterion would be very subjective according to the user's interests. From this work, we can see the first example of integration of Context Awareness and Document Ranking to improve IR. Later in 2007, on [36] the author proposed Ontology Knowledge Model to produce the Fuzzy Representation, then will be used in Clustering and Classification process for improving the performance of personalized IR [36]. The author has proposed a method for automatic extraction of Context Awareness factor such as persistent semantic user preferences, live and ad-hoc user interests, which are integrated with the

Document Ranking in order to improve the accuracy and reliability of personalization and the ranking of the IR. In the evaluation process, English date set and evaluation metrics such as Precision, Recall and MAP have been used. 5 years later in 2012, Keßler on [5], proposed the Semantics-Based IR, that involves the development of Semantic Web Rule Language (SWRL) Ontology Model that will be utilized in the process of Cognitively Plausible Dissimilarity Measure for IR Results (DIR) to improve the retrieval and ranking IR. The overall works have been evaluated by using a human participant test.

***Context-Aware IR* (CAIR) Research**				
Types of Context Model	2001-2005	2006-2010	2011-2015	2016-2020
Ontology Model and Document Context	Boanerges et al [3]	Ph. Mylonas et al [36], Carsten [5]		
Boolean Model and User Context	Dongpyo et al [8]			
x-Relation Model and Document Context		Lijun et al [19]		
Learning To Rank Model and Document Context		Biao et al [2]		Daan [7]
Thesaurus Model and Document Context			Nurfadhlina et al [37]	
User Profile Model and User Context			Kehinde [15]	Kehinde et al [14]
Graph Based Model and Document Context				Daan [7]
Neural Networks Model and Situational / Document Context				Hamed [10]

Fig. 2. Timeline of the research on *Context-Aware IR* (CAIR)

Other than that, we can see the application of *Mathematical Model* in CAIR on [8] in 2005; the author has proposed a *Framework of Personalized IR* with a *Mathematical Model*, a weighted *Boolean* that exploits the user's context within the well-defined components when the user is trying to retrieve information from objects in an environment. The author does not mention any ranking improvement, but we understand the author is trying to improve IR in general. The proposed method was evaluated with the *Precision* and *Recall* metrics and the English data set.

Another different work of types of context model then can be seen on [19] in 2009, the author proposed a *Context (multi-attribute graph) Model* or *x-Relation Model* (Using Graph: *Document Context* and multi attribute nodes) to cater the uncertainty problem and the tuples in the relational database and integrate with the *Document Ranking* to improve the ranking of the personal IR. The author presents the score function with two

components (the IR-styled score and the structural cost). The author evaluated the approach using real data set and *Discounted Cumulative Gain* metrics.

Xiang et al. on [2] in 2009, adopt a *learning-to-rank* approach and integrate the ranking principles into a state-of-the-art ranking model by encoding the *data source of the document context (context information)* as features of the model. In the evaluation, empirical test has been done using a large of English log of search engine data set and also involves human judgments and implicit user click data. The experimental results show the author *Context-aware* ranking approach improves the ranking of IR system which ignores *context information*. Furthermore, the method also outperforms a baseline method which considers contextual information in the ranking.

Agbele on [14, 15] has proposed the *Document Ranking Optimization (DROPT)* algorithm to optimize the ranking of IR system according to the *user context (user search context)* in a variety of environments with respect to *Document Context* such as *index keywords* and *the query vectors*. The author also explained about the process of identification of context, modelling it and the integration of this context model with *Document Ranking* towards clustering technique, in order to optimize the ranking and produce the adaptive IR system. The evaluation part has been done using *Precision* and *Recall* metrics and the English data set. The results demonstrate on how the attributes from the *user context* (*user search context)* can be applied in (*User Profile Model*), then can be used to improve the IR effectiveness.

Odijk in [7], presented *Retrieval-based Context Model* and *Graph-based Context Model* based on the *document context, such* as *Features of English Wikipedia Article* to improve the ranking of the IR system. The author proposes an entity linking approach for generating links from streaming text, consisting of two steps: (1) link retrieval and (2) *link reranking*. The author uses *learning to rerank* for improving the ranking, upon a strong link retrieval approach. Both context models, are evaluated with *Precision, Recall, Discounted Cumulative Gain (DCG)* metrics and was proven highly accurate and fast.

Zamani in [10], develops the *Neural Networks Model* in *Deep Neural Network Architecture* and utilised *the Document Context* of 10 language corpus and the *Situational Context*. The semantic matching and *Deep Network* involved in the process of integration of context representation and document ranking. The author evaluated the models using click data collected from the personal search engines and using the standard evaluation metrics for ranking such as *MRR* and *Precision@K*. The interesting aspect of this research is that it used the documents from the 10 language corpus. Unfortunately, this article, did not give any specific of what the 10 languages' name really is.

3 Topic Model for Context Aware Information Retrieval

Research works have focused on exploiting the sources of evidence to build the user profile that involves process of learning user's context by implicitly inferring the information from the user's behavior and from external or local context sources. One model of the User Profile Model is the Topic Model and it's proven to bring the significant

improvement such as in [12]. The author of [13] proposes Query-Specific Domain Model or Topic Domain Model and proposes a framework based on the language modelling approach (term relation technique), to integrate multiple contextual factors. In the evaluation part, several TREC collections of English data set and Average Precision, Recall and Precision@K evaluation metrics have been used. The results shown that multiple types of context can be used to produce significant improvements in retrieval effectiveness. In defining the context representation as well as the Topic Model of the text document, the researcher, always used the thesaurus as showed in [27], that used Sociopolitical Thesaurus for organizing the conceptual indexing and in IR, automatic text categorization and text summarization as well as in the works of [11] that proposed a new method to improve the text classification using a Thesaurus.

The example of *CAIR* that have been done in Malaysia can be seen in Sharef and Madzin work on [34], however, it does not focus on *Malay Text Corpus* but instead use the *ImageCLEF 2010* as the data set and *English* terms in their terminologies. The authors have proposed the semantic-based retrieval technique, by utilizing *Context-aware* query expansion and search ranking method and also utilized *Medical Thesaurus Based Model* as the *Contextual Representation*. The evaluation has been done using *Precision, Recall* and *Mean Average Precision* (MAP) metrics and achieved high *recall*. The author's finding, moreover, proven that a thesaurus can be utilized as the reliable knowledge base at hand for modelling context.

4 Evolution of Malay Information Retrieval

Based on the works that's available on the internet, we can summarised the timeline of the evolution of Malay IR that begin from 1995 until now and derived what research have been done so far and discover that any works have been done in aspects of Contextual IR.

In the period of the year 1995–2000, there are a couple of works have been done such as *Multilingual Document Retrieval System* by [1] in 1995, Stemming *Algorithm* by [9] in 1996 (that also indicate the beginning of the works of stemming technique for the *Malay IR)*, *Japanese-to-Malay Translation System* by [16] in 1999, *Syntax Analysis* by [4] in 2000 and the evaluation of retrieval effectiveness using spelling correction and *String-Similarity Matching* methods by [44] in 2000. The works of Malay IR, that continues in period of year 2001–2005 also show the improvements in stemming technique with *Automatic Learning of Stemming Rules* by [20] in 2003 and introduction of clustering technique in Malay IR using *Singular Value Decomposition* by [30] in 2005.

The period of year 2006–2010 showed the beginning of a Booming Era in the works of *Malay IR*, when many papers and journals have been produced to improve in a variety of areas. The era, started by the work on the *Malay WordNet System* by [21] in 2006, *Malay-Language Stemmer* by [22] in 2006, *Monolingual* and *Cross-Language IR* by [26] in 2006, Semantic *Similarity Measures* for *Sentence*s by [40] in 2007 and *Terms Visualization for Quran Documents* by [32] in 2007. The works followed by *Stemming Algorithms* and *Retrieval* in Malay and Arabic Documents by [43] in 2007, Feature Selection and Classification Techniques by [41] in 2008, *Classification of Proverbs*

using *Naïve Bayesian Algorithm* by [39] in 2008, *Categorization* using *Latent Semantic Indexing* by [31] in 2008, Term Weighting Schemes by [29] in 2008, Query translation for *Malay-English Cross-Language* IR system by [35] in 2010 (Fig. 3).

The variety of works, then continue to be produced in period of year 2011–2015, such as *Simple Rules Stemmer* by [42] in 2012, *Stemming Algorithm with Background Knowledge* by [19] in 2012, *Semantic Method* for *Query Translation* by [24] in 2013, *Sentiment Mining* using *Artificial Immune System* by [23] in 2013, Topic Analysis by [12] in 2014, and lastly *Named Entity Recognition Based on Rule-Based Approach* by [37] in 2014. The new era in 2016–2020, prior to that, there are many works also have been produced such as *Semantic-based Ontology* for *Qur'an Reader* by [28] in (2016). In *Malay IR* research, Rahman et al. in [33] in 2005, the author used *Latent Semantic Indexing* algorithm to improve the performance of retrieving relevant documents. Although, one of these research objectives is to retrieve documents in similar context, the author, just focusing on improving indexing of the Malay Retrieval with the parallel method and provide no explanation on the integration of the context representation and document ranking.

***Malay IR* Research**					
Research	1995-2000	2001-2005	2006-2010	2011-2015	2016-2020
Cross-Language	Belal et al [1]		Muhammad[26], Nurjannaton et al [35]		
Malay Stemmer	Fatimah et al (1996) [9]	Lily et al (2003) [20]	Mangalam et al [22]	Leow et al [18]	Nurazzah et al [33]
Malay Translation	Kentaro et al [16]		Nurjannaton et al [35]	Mohd Amin et al [24]	
Syntax / Lexical / Topic Analysis /	Bobby (2000) [4]			Haslizatul et al [12]	
Similarity	Zainab et al [44]		Shahrul et al [40]		
Clustering / Classification		Nordianah et al [30]	Shyamala [41], S. A. Noah et al [40], Nordianah et al. [31]		
Malay Wordnet / Ontology			Lim et al [21]		Nor Diana et al [28]
Indexing / Term (Visualisation / Weighting / Extraction)			Normaly et al [32], Nordianah et al [30]	Mazidah et al [23], Rayner et al [37]	

Fig. 3. Timeline of the *Malay IR* research

What we can conclude based on the review, there is no works have been utilised the contextual aspect in development or improvement of *Malay IR* research, and there is no research of *Context-Aware* application in *Malay IR* have been done so far.

5 Research Gap

As a conclusion of the review, we have identified several aspects and criteria of *Context-Aware* IR that are important to be utilized in the our future work, such as:

1. Most of the overall *Context Aware* IR is developing with the process of identification of context, and modelling it as a context representation. The *Context Representation* then will be involved in the process of integration with the document retrieval or *Document Ranking* to improve the IR. This kind of processes are obviously can be seen in the works of [3, 5, 14, 15], which all of these works apply the process of integration between context representation and document ranking to improve document retrieval.
2. The clearest findings that we retrieve in the review is that, there has been until this point no available *Malay* IR research that identifies the contextual aspect in bringing the improvement of the ranking. Then it also goes without saying that there has been no available *Malay* IR research that applied the process of integration between *Context Representation* and *Document Ranking* to improve document retrieval.
3. From the review too, the application of context modelling and integration was then generalized in one big framework such as in [3, 8, 13].
4. The identification of the features of the document in specific document as the contextual aspect and utilize in optimize the ranking result can also be seen in the work of [2, 7].
5. Most of the *CAIR* research used the evaluation metrics such as *Precision, Recall, Mean Average Precision* (MAP), *Discounted Cumulative Gain (DCG)* and also Human Judgement (*User Oriented Assessment Criterion*) to evaluate their finding.

6 Conclusion

As a conclusion, for our future work, we want to develop the application of context modelling that applied one of the *User Profile Model* in *Context Awareness* that is described as *Topic Model* and involve the *Characteristic Modelling* of the documents (features) and integrate it with the *Document Ranking*. Then, we will put the implementation in a framework that is called as a *Context-Aware Malay Document Retrieval Frameworks*.

This approach has been motivated from the review, that show to us, in development of the Context-Aware Information Retrieval System, generally the application of context modelling and integration has been generalized in one big framework such as in [3, 8, 13]. We also can see from the review, most of the overall *Context Aware* IR is developing with the process of identification of context, and modelling it as a *Context Representation*. The *Context Representation* then will be involved in the process of integration with the document retrieval or *Document Ranking* to improve the IR. This kind of processes are obviously can be seen in the works of [3, 5, 14, 15], which all of these works apply the process of integration between *Context Representation* and Document Ranking to improve document retrieval. This review has been done on the purpose of continuing

the research and developing the application for the proposed frameworks as shown in Figs. 4 and 5 below:

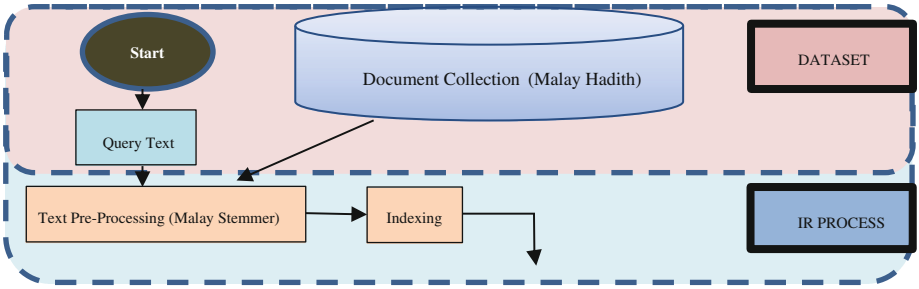


Fig. 4. Conceptual framework of context-aware Malay Document Retrieval frameworks.

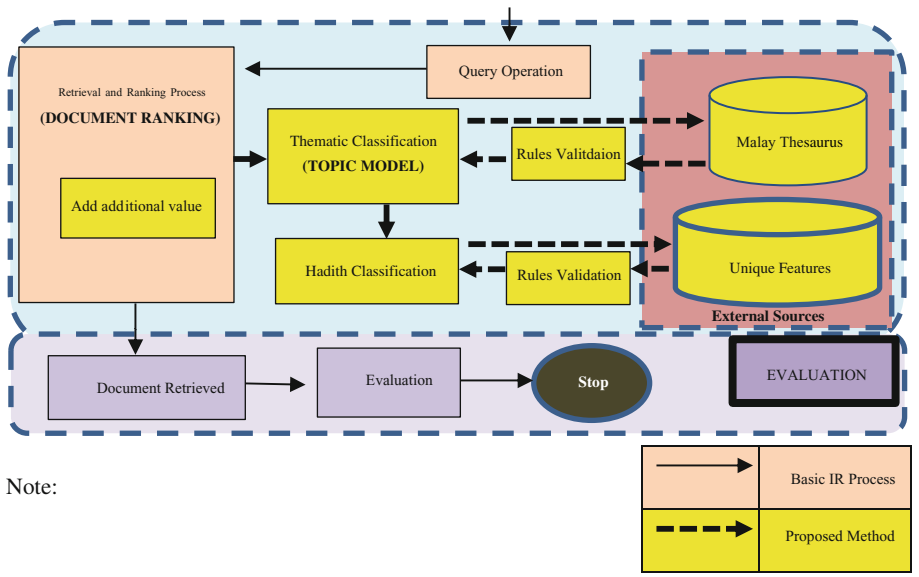


Fig. 5. Conceptual framework of context-aware Malay Document Retrieval frameworks (Continued)

The conceptual framework, features the classification of the Malay hadith using the unique features of the *Document Context* and related term in Malay Thesaurus as the Topic Model (*Context Representation)* and also act as the classifier rules. The classifier in this approach, can be represented as the process of integration between *Context Representation* and *Document Ranking*, by classifying the text document and give the specific score in this particular process, then this score will be used as an additional value in ranking function and will be improved the final ranking score of the IR system.

Acknowledgement. All authors are grateful for Universiti Teknologi MARA, Shah Alam, Selangor for financial support.

References

1. Ata, B.M.A., Sembok, T.M.T., Yusof, M.: SISDOM: multilingual document retrieval system. Asian Librar. Bradford **4**(3), 37–46 (1995). http://ezaccess.library.uitm.edu.my/login?url=http://search.proquest.com/
2. Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., Li, H.: Context-aware ranking in web search. In: SIGIR 2010, Geneva (2009)
3. Aleman-Meza, B., Halaschek, C., Arpinar, I.B., Sheth, A.: Context-aware semantic association ranking. Technical report 03-010, University of Georgia (2003)
4. Nazief, B.: Development of computational linguistics research: a challenge for Indonesia. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL 2000, pp. 1–2 (2000)
5. Keßler, C.: Context-aware semantic-based information retrieval. Inaugural-Dissertation for Doctor of Phylosophy. Wilhelms-Universität Münster (2010)
6. Akinribido, C.T., Afolabi, B.S., Akhigbe, B.I., Udo, I.J.: A fuzzy ontology based information retrieval system for relevant feedback. Int. J. Comput. Sci. **8**(1), 382–389 (2011)
7. Odijk, D.: Context & semantics in news & web search. SIKS Dissertation Series. No. 2016-20. Dutch Research School for Information and Knowledge Systems, Eindhoven (2016)
8. Hong, D., Park, Y.-K., Lee, J., Shin, V., Wo, W.: Personalized information retrieval framework. In: Proceedings of the First International Workshop on Personalized Context Modeling and Management for UbiComp Applications (2005)
9. Ahmad, F., Yusof, M., Sembok, T.M.T.: Experiment with a stemming algorithm for Malay words. J. Am. Soc. Inf. Sci. **47**(12), 909–918 (1996)
10. Zamani, H., Bendersky, M., Wang, X., Zhang, M.: Situational context for ranking in personal search. In: International World Wide Web Conference Committee (IW3C2), pp. 1531–1540. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva (2017)
11. Parvin, H., Dahbashi, A., Minaei, B.: Improving Persian text classification and clustering using Persian thesaurus. In: Distributed Computing and Artificial Intelligence. Advances in Intelligent and Soft Computing, vol. 151, pp. 493–500. Springer, Heidelberg (2012)
12. Hanum, H.M., Bakar, Z.A., Rahman, N.A., Rosli, M.M., Musa, N.: Using topic analysis for querying halal information on Malay documents. Procedia Soc. Behav. Sci. **121**, 214–222 (2014)
13. Bai, J., Nie, J.-Y., Bouchard, H., Cao, G.: Using query contexts in information retrieval. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 15–22. ACM, New York (2007)
14. Agbele, K.K., Ayetiran, E.F., Aruleba, K.D., Ekong, D.O.: Algorithm for information retrieval optimization. In: IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, Vancouver (2016)
15. Agbele, K.K.: Context-awareness for adaptive information retrieval systems. A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science. Faculty of Science Department of Computer Science, University of the Western Cape, Cape Town (2014)

16. Ogura, K., Bond, F., Ooyama, Y.: A prototype Japanese-to-Malay translation system. In: Machine Translation Summit VII. Asia-Pacific Association for Machine Translation, Tokyo (1999)
17. Derczynski, L.R.A., Yang, B., Jensen, C.S.: Towards context-aware search and analysis on social media data. In: Proceedings of the 16th International Conference on Extending Database Technology, EDBT 2013 (2013)
18. Leong, L.C., Basri, S., Alfred, R.: Enhancing Malay stemming algorithm with background knowledge. In: Proceedings of the 12th Pacific Rim International Conference on Trends in Artificial Intelligence, pp. 137–142. ACM, New York (2012)
19. Chang, L., Yu, J.X., Qin, L.: Context-sensitive document ranking. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1533–1536. ACM, New York (2009)
20. Suryana, L., Bressan, S.: Automatic learning of stemming rules for the Indonesian language. In: Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, pp. 62–68. Colips Publications, Singapore (2003)
21. Tze, L.L., Hussein, N.: Fast prototyping of a Malay WordNet system. In: Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing (LAICS-NLP) Summer School Workshop, pp. 13–16 (2006)
22. Sankupelayya, M., Valliappan, S.: Malay-language stemmer. Sunway Acad. J. **3**, 147–153 (2006)
23. Puteh, M., Isa, N., Puteh, S., Redzuan, N.A.: Sentiment mining of Malay newspaper (SAMNews) using artificial immune system. In: Proceedings of the World Congress on Engineering, vol. 3, pp. 1498–1503. Newswood Limited, International Association of Engineers, IAENG, Hong Kong (2013)
24. Yunus, M.A.M., Zainuddin, R., Abdullah, N.: Semantic method for query translation. Int. Arab J. Inf. Technol. **10**(3), 253–259 (2013)
25. Abdullah, M.T., Ahmad, F., Mahmod, R., Sembok, T.M.T.: Rules frequency order stemmer for Malay language. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **9**(2), 433–438 (2009)
26. Abdullah, M.T.: Monolingual and cross-language information retrieval approaches for Malay and English language documents. Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy. Universiti Putra Malaysia, Serdang (2006)
27. Loukachevitch, N.V., Dobrov, B.V.: Construction of thematic representations of texts based on domain-specific thesaurus. AAAI Technical report SS-02-09 (2002)
28. Ahmad, N.D., Bennett, B., Atwell, E.: Semantic-based ontology for Malay Qur'an reader. In: 4th International Conference on Islamic Applications in Computer Science and Technologies (2016)
29. Ab Samat, N., Murad, M.A.A., Abdullah, M.T., Atan, R.: Term weighting schemes experiment based on SVD for Malay text retrieval. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **8**(10), 357–361 (2008)
30. Ab Samat, N., Murad, M.A.A., Abdullah, M.T., Atan, R.: Malay documents clustering algorithm based on singular value decomposition. J. Theor. Appl. Inf. Technol. **8**(2), 180–186 (2005)
31. Ab Samat, N., Murad, M.A.A., Atan, R., Abdullah, M.T.: Categorization of Malay documents using latent semantic indexing. In: Proceedings of Knowledge Management International Conference, pp. 87–91. ACM, New York (2008)
32. Ismail, N.K., Rahman, N.A., Bakar, Z.A., Sembok, T.M.T.: Terms visualization for Malay translated Quran documents. In: Proceedings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung, Indonesia, pp. 17–19. IEEE Xplore (2007)

33. Rahman, N.A., Mabni, Z., Omar, N., Hanum, H.F.M., Rahim, N.N.A.T.M.: A parallel latent semantic indexing (LSI) algorithm for Malay Hadith translated document retrieval. J. Commun. Comput. Inf. Sci. **545**, 154–163 (2015)
34. Sharef, N.M., Madzin, H.: Semantic-based medical records retrieval via medical-context away query expansion and ranking. J. Theor. Appl. Inf. Technol. **58**(3), 697–706 (2013)
35. Rais, N.H., Abdullah, M.T., Kadir, R.A.: Query translation architecture for Malay-English cross-language information retrieval system. In: International Symposium in Information Technology (ITSim). IEEE Xplore (2010)
36. Mylonas, P., Vallet, D., Castells, P., Fernandez, M., Avrithis, Y.: Personalized information retrieval based on context and ontological knowledge. Knowl. Eng. Rev. **23**(1), 73–100 (2007)
37. Alfred, R., Leong, L.C., On, C.K., Anthony, P.: Malay named entity recognition based on rule-based approach. Int. J. Mach. Learn. Comput. (IJMLC) **4**(3), 300–306 (2014)
38. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval the Concepts and Technology Behind Search, 2nd edn. Pearson Education Limited, Harlow (2011)
39. Noah, S.A., Ismail, F.: Automatic classifications of Malay proverbs using Naïve Bayesian algorithm. Inf. Technol. J. **7**, 1016–1022 (2008)
40. Noah, S.A., Amruddin, A.Y., Omar, N.: Semantic similarity measures for Malay sentences. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 117–126. Springer, Heidelberg (2007)
41. Doraisamy, S., Golzari, S., Norowi, N.M., Sulaiman, M.N.B., Udzir, N.I.: A study on feature selection and classification techniques for automatic genre classification of traditional Malay music. In: 9th International Conference on Music Information Retrieval, ISMIR 2008, pp. 331–336. Drexel University, Philadelphia (2008)
42. Fadzli, S.A., Norsalehen, A.K., Syarilla, I.A., Hasni, H., Dhalila, M.S.S.: Simple rules Malay stemmer. In: The International Conference on Informatics and Applications (ICIA 2012), pp. 28–35. The Society of Digital Information and Wireless Communication (SDIWC) (2012)
43. Sembok, T.M.T.: Word stemming algorithms and retrieval effectiveness in Malay and Arabic documents retrieval systems. Int. J. Comput. Electr. Autom. Control Inf. Eng. **1**(10), 3197–3199 (2007)
44. Bakar, Z.A., Sembok, T.M.T., Yusoff, M.: An evaluation of retrieval effectiveness using spelling correction and string-similarity matching methods on Malay texts. J. Am. Soc. Inf. Sci. Technol. **51**, 691–706 (2000)

Improved Cascade Control Tuning for Temperature Control System

I. M. Chew[1(✉)], F. Wong[2], A. Bono[2], J. Nandong[1], and K. I. Wong[1]

[1] Curtin University Malaysia, Miri, Sarawak, Malaysia
{chewim,jobrun.n,wong.kiing.ing}@curtin.edu.my
[2] Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia
farrah@ums.edu.my, awangbono@gmail.com

Abstract. Single loop feedback control is commonly used in process control. The main drawback of single loop feedback control is its less effectiveness in rejecting the external disturbances. In order to improve speed of disturbance rejection and stability of closed-loop system, cascade control was studied and analyzed. To design the cascade control, first-order plus deadtime (FOPDT) models of both inner and outer loop were developed and applied for both sequential and simultaneous tuning methods. For sequential tuning, an IMC-based tuning was used whereas for simultaneous tuning method, Multiscale and Enhanced Cascade Control tunings were chosen. Relative performance of various controller settings for single and cascade control were compared. Moreover, recommendation for optimized tuning used "Step Response Checker" from System Design Toolbox was also elaborated and tested. Performance results were evaluated through Minimum Integral Error measurement. The effectiveness of the tuning methods was compared and evaluated using a lab-scale air flow rig.

Keywords: Single loop control · Cascade loop control · Optimized tuning
Relative performance · Improved temperature control

1 Introduction

1.1 A Single Loop Control System

Single loop feedback control compensates the process upsets or to provide high consistency of process control that ensures the operational safety and product quality [1, 2]. Basically, the process variable, PV is measured and then compared with the applied setpoint, SP so as to produce manipulated value, MV for re-positioning the control of final control element. Eventually, the parameter has been controlled according to the applied set point, SP.

The block diagram of a feedback control system is shown in Fig. 1.

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 410–419, 2018.
https://doi.org/10.1007/978-981-10-8276-4_39

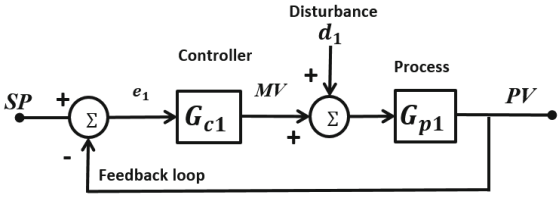


Fig. 1. A typical block diagram of a closed loop system

It is noted that a single loop control system responded poorly to the imposed disturbances. As depicted by Fig. 1, disturbance d_1 enters the process, it has to propagate through the process and error signal, e_1 is produced to initiate the control. Meaning, the controller starts to react after the process is affected by the upset. Besides, the controller also has its limitation to control a large process by itself due to possession of large dead time or time constant.

If the system has one process variable to interact with another parameter, the overall control action can be improved by adding measurement and control the other parameter that will impact the actual (primary) controlled parameter. In a nutshell, we can develop another feedback control loop to deal with imposed disturbance before it begins to affect the controlled parameter. It has greatly overcome problems due to slow control's reaction to any unpredicted disturbances as well as required new setpoint.

1.2 Cascade Loop Control System

Nowadays, cascade control system is widely implemented in various chemical process plants, boilers and utilities systems. Introduced by Frank and Worley, a typical cascade control system consists of two closed loops in such a way that inner loop is embedded into an outer control. Basically, cascade control is classified in series and parallel form [3]. Lee et al. [9] had extended the studies of cascade control to integrating and unstable processes. Interestingly, Nandong and Zhang [4, 8] studied cascade control in multi-scale control scheme.

In this paper, the structure of cascade control system is developed and visualized through Matlab simulation tool. Typical block diagram of a cascade control loop is shown in Fig. 2. Two controllers are used but only primary variable, PV_1 eventually is to be regulated at setpoint, SP_1 and despite the imposed disturbance, d_2 [3].

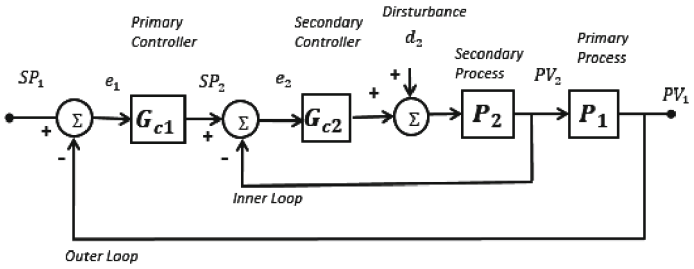


Fig. 2. Block diagram of cascade control system

The resultant impact to the outer loop is much reduced compared to the standard feedback control. It is desired that, outer controller G_{c1} should drive process variable to the setpoint as to abide to the Minimum Integral Absolute Error criterion [6]. Thereby, cascade control loop often requires a more vibrating dynamic behaviour of inner loop rather than outer loop. Otherwise, the cascade loop will not offer significant improvement in controlling controlled parameters. For the faster inner loop, its time constant, τ_p and dead time, τ_d should be less than the values of outer loop. In general, P or PI controller mode is selected for inner loop whereas PI or PID controller mode is selected for outer loop.

There are two tunings method to be discussed, known as Sequential Tuning Method and Simultaneous Tuning Method.

Sequential Tuning Method. This is a state-of-art tuning method used by experienced personnel with minimum guidance from the available tuning methodologies. The steps for sequential tuning method are shown as below [5–7].

(a) Set both controllers to the Manual Mode.
(b) Perform the bump test to the inner loop and develop the First Order Plus Dead Time, FOPDT model by using open loop test method.
(c) Determine proportional gain, K_c and integral time, T_i of secondary controller. Then, set the secondary controller to Auto Mode.
(d) Perform the bump test to the outer loop and develop FODPT model for outer loop by using open loop test method.
(e) Determine proportional gain, K_c, integral time, T_i and derivative time, T_d for primary controller.
(f) Tune the primary controller. Then, set the primary controller to Auto Mode.

Simultaneous Tuning Method. There is an alternative tuning methodologies to determine correlation values for both primary and secondary controllers concurrently through a single bump test. In this research paper, Multi-scale Control [8] and Enhanced Cascade control [9] are analysed and compared with single loop temperature control system. The tuning is initialized by setting both inner and outer loop into Manual Mode. A bump test is conducted to the controller output, see Fig. 2. Corresponding dynamic behaviours of

both inner and outer loop are defined for developing FOPDT models. Correlation values for both controllers G_{c1} and G_{c2} are calculated using formulas in both literatures [8, 9].

Correlation Tuning Using IMC-Based Model. The IMC-based tuning method is chosen because it has good capability to drive process variable to new set point in a consistent fast speed track without overshooting. This characteristic is still obvious even the controller is set with overdamped characteristic that gives slower and consistency reaction towards d_1. The details of formulation are available from the literature [11]. The used formulas for both single loop and cascade control are shown in Table 1.

Table 1. Formula of PID settings

Heading level	Proportional gain, Kc	Integral time constant, τ_p	Derivative time, τ_d
PI controller	$\frac{1}{K_p}\frac{\tau_p}{(\theta_p+\tau_c)}$	τ_p	–
PID controller	$\frac{1}{K_p}\left(\frac{\tau_p+0.5\theta_p}{\tau_c+0.5\theta_p}\right)$	$\tau_p+0.5\theta_p$	$\frac{\tau_p\theta_p}{2\tau_p+\theta_p}$
Aggressive tuning: τ_c = larger of 0.1 τ_p or 0.8 θ_p			
Moderate tuning: τ_c = larger of 1.0 τ_p or 8.0 θ_p			
Conservative tuning: τ_c = larger of 10 τ_p or 80 θ_p			

Performance Measurement with IAE, ISE, ITAE, ITSE. The purpose of controller's setting is enabling control action to drive controlled variable converge to its setpoint. The smaller value of integral error means high degree of controllability and robustness in response to the new setpoint and disturbances. Overall performance is measured through Integral Absolute Error, *IAE*, Integral Time-weighted Absolute Error, *ITAE*, Integral Square Error, *ISE*, and Integral Time-weighted Square Error, *ITSE*. The respective measurement is developed in Simulink environment as illustrated in Fig. 5.

Response Optimization with Step Response Checker. Step Response Checker Function is used for repetitively evaluating and comparing response of generated iterations so to generate the best PID tuning parameters that meets step response design requirements [12, 13]. Applying this block function do not need to deal with control knowledge as well as tuning skills for obtaining the PID setting values. The block function can be extracted from Simulink Design Optimization of Simulink Library Browser. After it is connected to the model, requirements such as rise time, settling time, overshoot, initial value and final value are inserted to the sink block parameter window. The optimized operation is started by clicking optimized button. As the optimized tunings are obtained, plot displays and PID tuning values are present. The Step response checker is applied to the model as shown in Fig. 3.

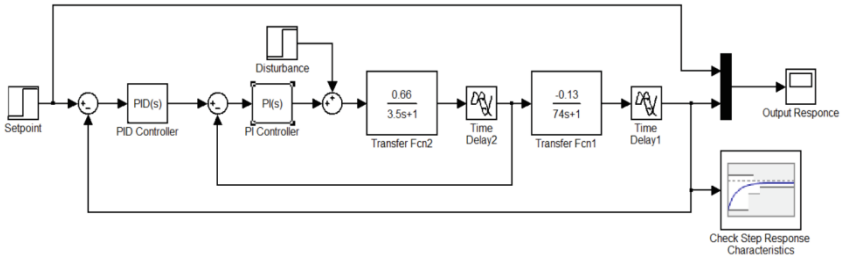
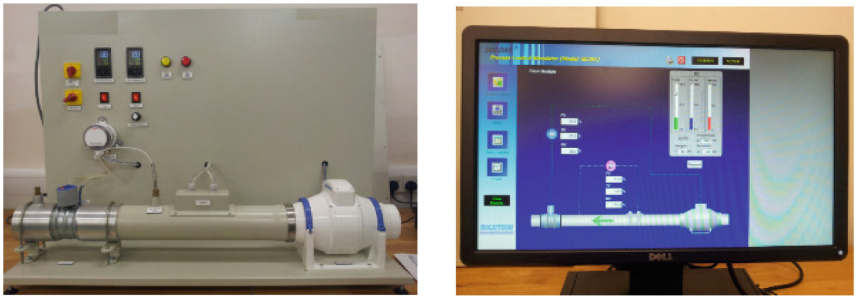


Fig. 3. Optimized tuning of PID controller with Step Response Checker

2 Experimental Case Studies

The three basic controls can be performed by Process Control Simulator, SE-201 includes Flow control, Temperature control and Cascade control [10]. The structure of SE-201 is depicted in Fig. 4(a) and (b).



(a) (b)

Fig. 4. Process Control Simulator SE-201

The flow control loop is nested into temperature control loop where both temperature and air flow are controlled simultaneously by manipulating the speed of the fan. Any disturbance to the air flow will be compensated for by the flow controller in the inner control loop, and therefore minimizes the noticeable disturbance to the temperature control loop.

3 Analysis and Results

3.1 First Order Plus Dead Time (FOPDT) for Single Loop and Cascade Loop Control System

Dynamic behaviour of open loop for temperature control is obtained through formulating FOPDT model as shown in (1).

$$P_2 = \frac{0.68e^{-15s}}{150s + 1} \tag{1}$$

For cascade control loop, open loop test method is applied to define FOPDT models. As the cascade control system has utilises flow control loop to improve temperature control, two FOPDT models are determined for both sequential tuning and simultaneous tuning.

Table 2 illustrates the developed FOPDT models for sequential tuning and simultaneous tuning method. It is noted that both tuning method is having similar inner loop model while the outer loop model is obtained differently. Simultaneous tuning method only requires a single bump test, and evaluates the reaction for both flow and temperature at the same time.

Table 2. FOPDT for sequential and simultaneous tuning methods

Cascade control system	Sequential tuning	Simultaneous tuning
Inner loop model	(MV flow - PV flow)	(MV flow - PV flow)
	$\frac{0.66e^{-0.5s}}{3.5s + 1}$	$\frac{0.66e^{-0.5s}}{3.5s + 1}$
Outer loop model	(MV temp - PV temp)	(MV flow - PV temp)
	$\frac{-0.13e^{-20s}}{74s + 1}$	$\frac{-0.08e^{-43s}}{103s + 1}$

3.2 Correlation Tunings for Sequential and Simultaneous Tuning Methodology

Table 3 depicts all PID settings for single loop and cascade loop temperature control. For single loop temperature control, PID settings are applied for Moderate and Aggressive Mode tuning. In contrast, cascade loop temperature control has setting PID values by using both sequential and simultaneous tuning methodologies. For simultaneous tuning methodology, Multi-Scale Control [8] and Enhanced Cascade Control [9] are used for determining the PID setting values.

Table 3. PID settings for temperature control system

Tuning methodology	Loop	Proportional band, PB	Integral time, τ_I	Derivative time, τ_d
Single moderate	Single	125%	128	10.1
Single aggressive	Single	18.5%	128	10.1
Cascade tuning-sequential tuning	Outer	–26%	–84	–8.8
	Inner	85%	3.5	0
Cascade tuning-enhanced	Outer	–4.9%	–103	–19.8
	Inner	13.5%	3.7	0
Cascade tuning-multiscale	Outer	–4.7%	–105	–17.7
	Inner	14.2%	3.2	0
Cascade tuning-optimized	Outer	–4%	–85	–8.8
	Inner	85%	3.5	0

3.3 Model of Cascade Control System

Figure 5 illustrates developed cascade control system with Integral Error Measurement. It is noted that the inner loop controller is direct-acting control loop with positive process gain. Therefore, the applied Proportional Band, PB is with positive value. Meanwhile, the outer loop is a reverse-acting control loop thus the applied Proportional Band, PB is with negative value.

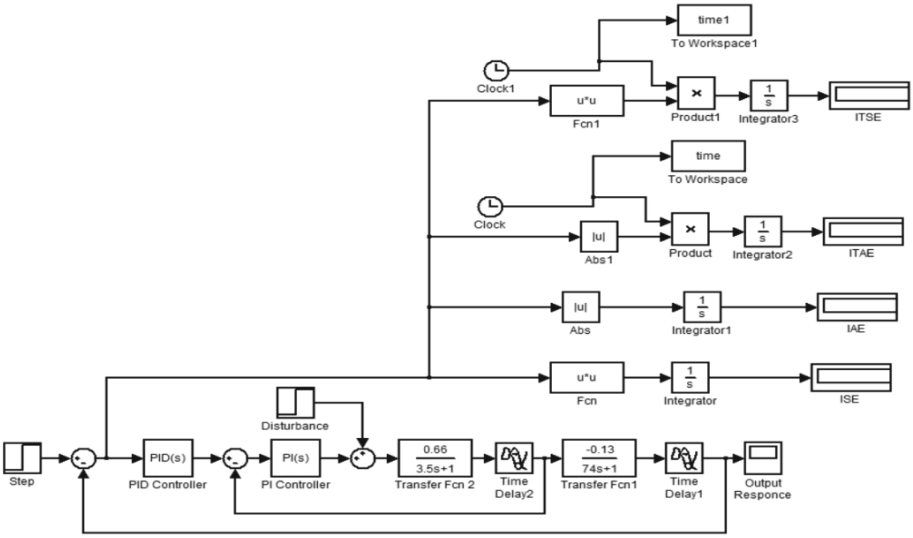
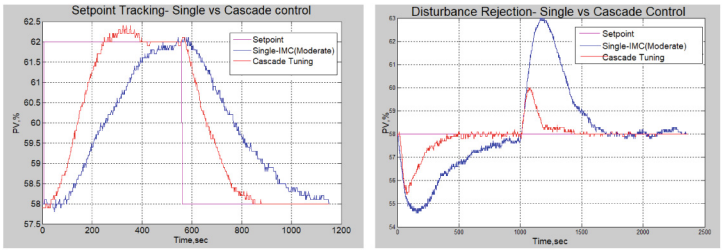


Fig. 5. Block diagram of cascade control system in Matlab Simulnk

3.4 Relative Performance of Single Loop Temperature Control Compared to Cascade Loop Temperature Control

Figure 6(a) and (b) depicts improvement in terms of system response by using cascade control methodology, which is relatively fast compared to the single loop temperature control. It is experimentally shown that the cascade control overall improves servo and regulatory control performances and this can be well-explained due to the inner loop able to regulate the airflow that enhances manipulation to response to the temperature control.

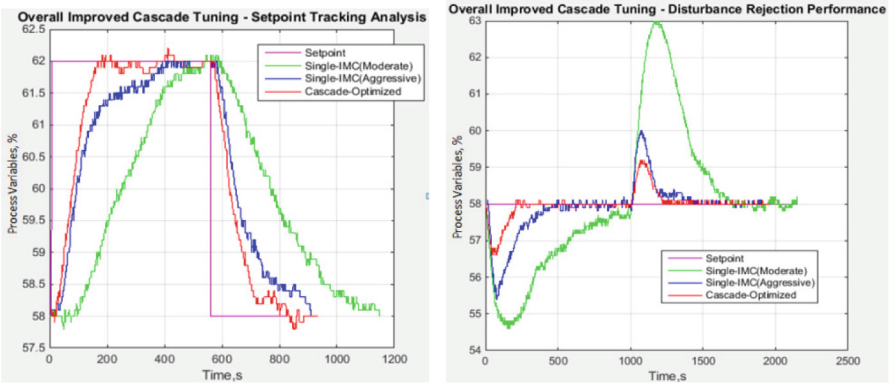


(a) (b)

Fig. 6. Set point tracking analysis and disturbance rejection performance optimized cascade control loop tuning.

3.5 Relative Performance Index Using Minimum Integral Error

As noted, one of the main purposes to apply cascade control loop is for improving the area of integral error under the curve of output response as compared to setpoint over duration of time. Figure 7, shows the block diagram of cascade control loop, which is linked to function of Integral Error Measurement.



(a) (b)

Fig. 7. Improvement on disturbance rejection performance for moderate-IMC tuning, aggressive-IMC tuning and cascade control optimized tuning.

Relative performance of temperature control system can be measured through IAE, ITAE, ISE and ITSE criterion. The simulation result is shown in Table 4.

Experimental result showed that optimized cascade control tunings gives the smallest value of integral errors for all IAE, ITAE, ISE and ITSE, which means the greatest stability control towards the changes of setpoint, SP and disturbance, d. In contrast, moderate tuning of single loop temperature control system gives the largest value for all IAE, ITAE, ISE and ITSE measurement, implies the least controllability, which expects the output response with higher overshoot and longer settling time.

Table 4. Minimum integral error measurement using IAE, ISE, ITAE, ITSE

Heading level	Minimum integral tuning measurement			
	IAE	ISE	ITAE	ITSE
Single-moderate	262.8	161.3	6.024e + 04	3.226e + 04
Single-aggressive	89.95	73.2	2.295e + 04	1.827e + 04
Cascade tuning-sequential tuning	81	58.22	5964	2463
Cascade tuning-enhanced	76.09	54.59	5442	2199
Cascade tuning-multiscale	71.51	53.01	4637	2054
Cascade tuning-optimized	40.43	26.65	2865	711.8

3.6 Relative Output Response of Optimized Cascade Control Compared with Single Loop Control

As noted, one of the main purposes to apply cascade control loop is for improving the closed-loop system performance. Overall, an improved response can be justified through comparing relative performance for both cascade and single loop temperature control.

Figure 7(a) and (b) shows relative performance for both setpoint tracking analysis and disturbance rejection performance for Single-IMC (Moderate) tuning, Single-IMC (Aggressive) tuning and optimized cascade control tuning. It is obvious to note that the temperature of process with cascade loop is driven more quickly to the new setpoint. Higher controllability as a consequence of the cascade control to produce smaller overshoots in coping with external disturbances, and thus the impact to the system is significantly reduced.

Improvements of optimized cascade control in terms output response are shown in the Table 5. It is showed that optimized cascade control improves 67.4% of settling time for setpoint tracking analysis, and reduces 76% overshoot for disturbance rejection performance when is compared to single loop moderate tuning control.

Table 5. Output response of single loop and cascade loop temperature control

Control loop	Set point tracking analysis		Disturbance rejection performance	
	Settling time(s)	Improvement (%)	Overshoot (%)	Improvement (%)
Single loop (Moderate)	478	–	5	–
Single loop (Aggressive)	377	21.1	2	60.0
Optimized cascade control	156	67.4	1.2	76.0

4 Conclusion

Cascade control system is designed to improve the closed-loop control performance in terms of improving the response speed, robustness and controllability of the temperature control system.

Research methodology was justified by case studies through Process Control Simulator, SE201. All findings were recorded and rigorously compared. The developed FOPDTs showed that the process gain for the equipment dictates that a forward-acting control for the inner loop and reverse-acting control for outer loop. Both sequential tuning and simultaneous tuning were recommended for analysing relative performance of process towards new setpoint tracking and imposed disturbance.

It has been found that sequential tuning is more dependent on personnel's tuning skills by using state-of-art approach for reasonable control. In contrast, the simultaneous tuning applies the respective developed formulas, which requires less skill and reduces time for tuning task.

Moreover, a study of the optimization of cascade control was carried out and re-fined tuning was performed so to ensure the real-time process performed at optimum level. The optimized setting for cascade control based on the experimental rig was found to be PB = 4%, $\tau_i = 88$ and $\tau_d = 8.8$ for outer loop and PB = 85%, $\tau_i = 3.5$ for inner loop. Noteworthy, it can be concluded that the cascade control loop has improves 67.4% of servo control and 76.0% of regulatory control based on the real system under investigation.

References

1. Riggs, J.B., Karim, M.N.: Chemical and Bio-process Control. 3rd edn. Pearson Education International, Boston (2007)
2. Nise, N.S.: Control System Engineering, 4th edn., pp. 255–263. Wiley, Hoboken (2004)
3. Thirumarimurugan, M., Mahalakshmi, D., Sivakumar, V.M., Dinesh Kumar, V.: Evaluation of series and parallel cascade using optimal controller. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. (IJAREEIE) **5**(Special Issue 1), 122–126 (2016)
4. Nandong, J., Zang, Z.: Generalised multi-scale control scheme for cascade processes with time-delays. J. Process Control. **24**, 1057–1067 (2014)
5. Smith, C.A., Corripio, A.B.: Principles and Practice of Automatic Process Control, 2nd edn., pp. 439–459. Wiley, New York (1997)
6. Cooper, D.J.: Practical Process Control using LOOP-PRO Software, pp. 160–192. Control Station Inc., USA (2005)
7. Ericks, K.T., Hedrick, J.L.: Enhancements of Single Loop Regulatory Control. Wiley, New York (1999)
8. Nandong, J.: Analytical tuning method for cascade control system via multi-scale control scheme. Int. J. Autom. Control. **10**(2), 167–192 (2016)
9. Lee, Y., Oh, S., Park, S.: Enhanced control with a general cascade control structure. Ind. Eng. Chem. Res. **41**, 2679–2688 (2002)
10. SOLTEQ: Operating and maintenance manual – Process Control Simulator SE 201, Malaysia (2015)
11. Shah, A.K., Aniljumar, M., Parikh, N.N.: Performance analysis of IMC based PID controller tuning on approximated process model. Nirma Univ. J. Eng. Technol. **1**(2), 51–54 (2010)
12. Ahmed, S., Huang, B., Shah, S.L.: Parameter and delay estimation of continuous-time models using a linear filter. J. Process Control **16**(4), 323–331 (2006)
13. The Mathwork Inc.: Design optimization to meet step response requirements (GUI) (1997–2014). https://www.mathworks.com/help/sldo/gs/optimize-controller-parameters-to-meet-step-response-requirements-gui.html

GOW-LDA: Applying Term Co-occurrence Graph Representation in LDA Topic Models Improvement

Phu Pham[1], Phuc Do[1(✉)], and Chien D. C. Ta[2]

[1] University of Information Technology (UIT), VNU-HCM, Ho Chi Minh City, Vietnam
phamtheanhphu@gmail.com, phuc.do@uit.edu.vn
[2] Industrial University of Ho Chi Minh City (IUH), Ho Chi Minh City, Vietnam
tdcchien@gmail.com

Abstract. In this paper, we demonstrate a novel approach in topic model exploration by applying *word co-occurrence* graph or ***g**raph-**o**f-**w**ords (GOW)* in order to produce more informative extracted latent topics from a large document corpus. According to the ***L**atent **D**irichlet **A**llocation* (LDA) algorithm, it only considers the words occurrence independently via probabilistic distributions. It leads to the failure in term's relationship recognition. Hence in order to overcome this disadvantage of traditional LDA, we propose a novel approach, called *GOW-LDA*. The GOW-LDA is proposed that combines the GOW graph used in document representation, the frequent subgraph extracting and distribution model of LDA. For evaluation, we compare our proposed model with the traditional one in different classification algorithms. The comparative evaluation is performed in this study by using the standardized datasets. The results generated by the experiments show that the proposed algorithm yields performance respectably.

Keywords: Topic models · LDA · Word co-occurrence graph
Graph-of-Word (GOW) · Frequent subgraph mining · Classification
Large-scale graph

1 Introduction

In recent researches, topic modeling is considered as an interesting field which addresses the approach to represent the documents in multiple latent themes or topics corresponding with the probabilistic distributions over words in each document of corpus. There are many applications in regard to topic modeling, such as information retrieval, large-scale document corpus indexing, text mining, classification, clustering, etc. The most popular topic modeling algorithm is *Latent Dirichlet Allocation* (LDA) [1, 2]. In general, the LDA also is considered as **b**ag-**o**f-**w**ords (BOW) model with the probabilistic term weight. The disadvantage of this method is the failure of tracking the document's term semantic relationship and word's combination representation which are most informative factors in document representation.

There is no doubt that every single human written document is a collection of words which occur in the specific related contexts – might belong to one or many

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 420–431, 2018.
https://doi.org/10.1007/978-981-10-8276-4_40

topics/categories, the terms' distributions inside a document rely on the semantic relations and natural language's grammar. In other words, the neighboring terms in a document represented for the contextual topic which a document belongs to.

The ultimate goal of our work in this paper is to find a novel approach for resolving the problem of tracking the word's relationships in document representation as well as the latent topic extraction. We propose the LDA-based graph-of-word (GOW-LDA) method, which is a combination between graph-based document representation and LDA algorithm in order to tackle the aforementioned issues of classical LDA algorithm.

2 Literature Review

There are several notable improvement in LDA topic models for improving the informative of extracted knowledge such as supervised annotation topic extraction [3], investigating the semantic relationship among words inside a document via "***n**atural **l**anguage **p**rocessing*" (NLP) techniques in text processing, analyzing the common-sense concepts from input document via appropriate knowledge-bases as the topic's features for LDA [4], *SemLDA* [5] applying *WordNet*, *SemCor*, *WSD* (word-sense disambiguation). In general, most previous methods require knowledge-based repository such as WordNet, lexical database, domain ontologies, etc. However, in some cases those methods are not suitable, such as:

- The words' combination, abbreviation, synonyms are vary and specific which can't be looked up in available lexical or dictionary (especially in technical document).
- The input document's corpus and included concepts related to multiple knowledge-domain fields which is needed to apply multiple knowledge-source. Moreover, the input documents are in combination of different data types (image, charts) such as scientific articles, academic reports, etc. also cause the difficulties in applying lexical lookup.

In order to avoid the burdensome obtained from previous approaches, we propose the approach of applying the graph-of-word representation model and "*frequent sub-graph mining*" (FSM) techniques in order to improve the quality of extracted latent topics in classical LDA algorithm.

3 Proposed Solutions and Models

3.1 Document to Graph (doc2graph) Transformation

Definition 1. *graph-of-words (GOW)* in document representation is defined as a graph-based structure having vertices correspond to unique terms of the document and whose edges represent co-occurrences between these terms.

Constructing the word co-occurrence graph from texts in document is considered as the simplest way to transform a document to graph-based structure. In order to eliminate self-looping (the edge connecting same vertex) – the (k) sliding window's size

parameter is defined to control the document's term sequential iterating. The selection of (k) depends on the processing language. The value of (k) might be set 2 (for capturing bigrams) or 3 (for capturing trigrams). With English document corpus, the (k) is normally set to 3, in order to capture the trigrams and in order to avoid English bigrams which are composed by duplicated words, (i.e., "*so so*", "*well well*", etc.) [6–8]. All steps are described in Algorithm 1 as shown follows:

Algorithm 1. Pseudo code for document to GOW transformation

Input: document (d) with number of unique terms (N_d) extracted from (d) in text-preprocessing phrase, with sliding window's size ($k = 3$)[*]
Output: GOW-based structure of (d).

```
Function doc2GraphTransformation()
    Create G_d = (V,E), where V = N_d, E
    For word (w_i) in (N_d)
        For (j) in range(1, k)
            If w_i == w_(i+j)
                continue; #checking duplicated vertices
            If e(w_i,w_(i+j)) not existed in (E)
                E ← e(w_i,w_(i+j))
            End if
        End For
    End For
    return G_d
End Function
```

3.2 The Proposed GOW-LDA Approach

Applying GOW and FSM in Document Transformation

Definition 2. Graph-based concepts ($\delta : G_{freq}$): is a set of GOW based structures which are extracted from the input documents and FSM technique.

The GOW-LDA model is our novel approach which combines the "*graph-based concepts*" (Definition 2) for document representation and classical LDA algorithm to extract the generative distributions.

Graph-based concept via FMS

In general, the graph-based concept extraction from input documents (as shown in Fig. 1) is the process of FMS. Because the input documents are transformed into graph-based structures (GOW), hence extracted graph-based concept ($\delta_i : G_{freq}$) is isomorphic to input documents' graph (G_d). For the set of graph-based transformed documents $G_D = \{G_j | j \in |D|\}$, the support value $\sigma(\delta_i : G_{freq}, G_D) = \sum_{G_j \in |D|}^{j} \varsigma(\delta_i : G_{freq}, G_j)$ donates to the frequency of concept $\delta_i : G_{freq}$ in G_D, the

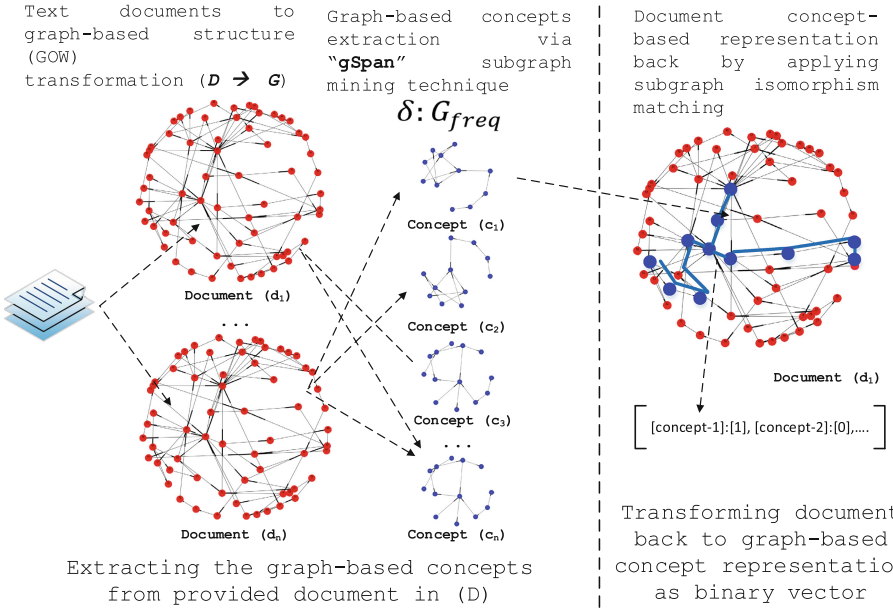


Fig. 1. Document dataset is transformed into graph-based concept in ($\boldsymbol{\delta} : \boldsymbol{G_{freq}}$) and represented back by applying subgraph isomorphic matching.

$\varsigma(\delta_i : G_{freq}, G_j) = 1$ if $\delta_i : G_{freq}$ is isomorphic to at least one subgraph of G_j and 0 when $\delta_i : G_{freq}$ is not isomorphic to any subgraph of G_j. The GOW-LDA introduces the input parameter ($\delta_i : G_{freq}$) - the $\delta_i : G_{freq}$ is the (i)-th frequent subgraph which extracted via (FSM) techniques such as gSpan in frequent graph-based pattern mining and the expanded version of DgSpan for directed graph structure [9], FFSM [10] applying isomorphic subgraph presence in FSM, and SPIN [11] focus on extracting "*maximal frequent subgraphs*" (MCS) from graph-based storage mechanism, etc.

Document representation by extracted graph-based concepts

After extracting the set of concepts, $\delta : G_{freq}$, each of the concepts will have a specific (*cid*) for identification. Then, a set of $\delta : G_{freq}$ will be used to represent the input documents into the concept-based binary vector (as shown in Fig. 1) by using the "*isomorphic subgraph matching*" (ISM) matching. For the ISM related problem, there are some notable proposed algorithms such as: VF2 in matching large graphs [12], QuickSI [13, 14].

Distributions of Topic–Concept in GOW-LDA

LDA topic model

LDA is considered as the foundation of existing probabilistic-based topic modeling [1]. It helps to generate the mixture of latent topic distribution and each given latent topic is

a probabilistic distribution over "*unique words/terms*". The $P(w|z) = \phi^{(t)}$ stand for the distribution of word (w) over the given topic (t) and every single document in the collection contained multiple topics which distributed probabilistically - $P(z) = \theta^{(d)}$ for the distribution of topic z in the specific document (d) (Figs. 2 and 3).

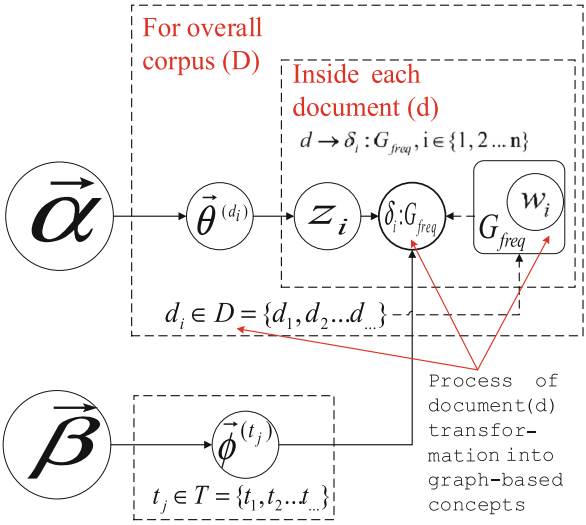


Fig. 2. Overall generative model processes of GOW-LDA

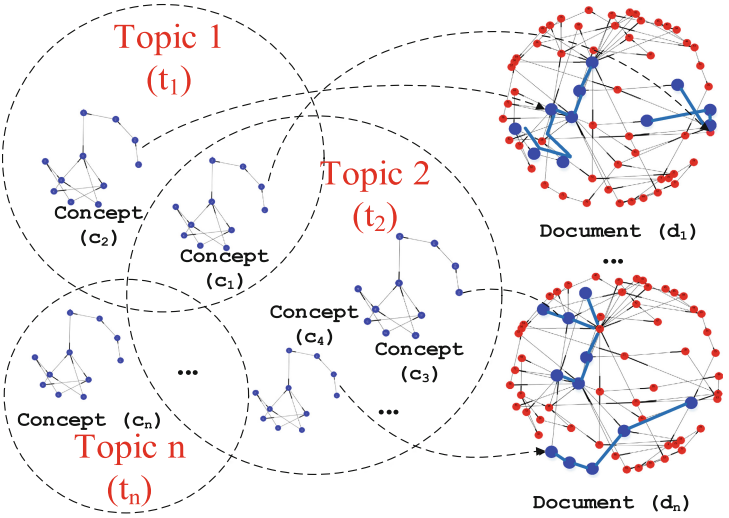


Fig. 3. Graphical illustration of [topic]-[concept] distributions in GOW-LDA via graph-based concepts

Similar to the previous classical one, in GOW-LDA, we transform the input documents into "*unique concepts*" of $\delta_i : G_{freq}$. Hence, the model becomes the probabilistic distribution of [topic]-[concept], the $P(w|z) = \phi^{(t)}$ becomes $P(\delta : G_{freq}|z) = \phi^{(t)}$ (as shown in Fig. 2). In the other word, the GOW-LDA is the distribution of [topic]-[group of co-occurrence words] (as shown in Fig. 2). Finally, we apply the *Gibbs sampling* [2] to compute the mixture probabilistic distributions of (ϕ^t) (the distributions of [topic (t_j)]-[concept ($\delta_i : G_{freq}$)]) and [topic]-[document] (θ^d) (as shown in Fig. 2). The parameters (α) and (β) are the selected hyper-parameters. The GOW-LDA generative processes can be summarized as Algorithm 2.

Algorithm 2. Generative processes of GOW-LDA inspired from classical LDA with Gibbs sampling

Input: the set of document collection (D) and selected hyper-parameters (α) and (β).
Output: the distribution of topic and concept($\delta: G_{freq}$)] (ϕ^t) and [topic-document] (θ^d).
Step 1: transforming the input documents ($D, D = \{d_1, d_1 \ldots d_n\}$) into set of $\delta: G_{freq}$ concepts (subgraphs): $D \rightarrow \delta: G_{freq}$.
Step 2: choosing the topic-document proportions ($\theta^d|\alpha$) by Dirichlet distribution: $\theta^d|\alpha \backsim Dir(\alpha)$.
Step 3: Iterating through set of concepts, each (i)-th $\delta: G_{freq}$ (subgraphs) in each (j)-th document:
  Step 3.1: Sampling the topic of [topic]-[document] distribution by multinomial distribution on θ^{d_j}: $t_{ij}|\theta^{d_j} \backsim Multinomial(\theta^{d_j})$.
  Step 3.2: Sampling the concept $\delta_i: G_{freq}$ and topic (t_{ij}) distribution $\delta_j: G_{freq}|t_{ij}$ by multinomial distribution on $\phi^{t_{ij}}$: $\delta_j: G_{freq}|t_{ij} \backsim Multinomial(\phi^{t_{ij}})$.

The Advantages of GOW-LDA in Resource Consuming

The main advantage of GOW-LDA is to keep the co-occurrence, semantic relationships and combination of terms within documents. Besides, our proposed approach prevents the problem related to BOW problem in classical LDA algorithm. Furthermore, based on our experiments (as shown in Fig. 2), the GOW-LDA also reveals the effectiveness in consuming time as well as resource usage for generating distributed models due to the replacement of set of unique term (V) by defined graph-based concepts ($\delta : G_{freq}$) which absolutely much smaller in size ($|V| \gg |\delta : G_{freq}|$).

We conducted our experiments on the *WebKB* dataset with different number of extracted latent topics and extracted concepts (GOW-LDA) and terms (for classical LDA algorithm) (Table 1).

Table 1. Experimental results on execution time between GOW-LDA and classical LDA with different number of extracted topics and concepts/terms

Number of extracted latent topics	Number of distributed concepts/terms per topic	Execution time (in seconds)	
		GOW-LDA	Classical LDA
5	10	7.799	10.026
10	20	8.564	12.790
15	30	10.424	13.395
20	40	13.881	14.426
50	100	17.551	22.486

4 Experiments and Evaluations

In order to estimate the effectiveness and potential applications of our proposed GOW-LDA model, we conducted several experiments related to evaluations in co-occurrence terms associated with extracted concept from GOW-LDA with classical.

The second, evaluation is focused on the capabilities in classification implemented with multiple classifying algorithms. For the fair testing, all experiments will be conducted in the provided standardized dataset.

4.1 Evaluation's Metric Usage

For the second evaluation on model capabilities of classification, three main metrics including Precision (P), Recall (R) and F-measure (F_1) are used in order to evaluate the experimental results:

$$\boldsymbol{P} = \frac{\mathrm{TP}}{TP+FP}, \boldsymbol{R} = \frac{TP}{TP+FN}, \boldsymbol{F}_1 = 2.\frac{P.R}{2(P+R)} \quad (1)$$

Where (TP) is the number of topic found and correct, (FP) and (FN) are for the classifier designated the expected topic but not correct and not the expected topic by actually belong to that topic, respectively.

4.2 Experimental Results and Evaluations

The Term/Concept Distributions over Topic

In order to evaluate the extracted terms in the concepts which probabilistic distributing over topic from dataset – After generating the [concept]-[topic] (ϕ^t) distribution, we map back the terms from previous concepts (in graph-based structure – {term1, term2…} following the sequential order).

It is also edges' directions in GOW to compare with the extracted terms in classical LDA model. The experiments are taken in WebKB (4 classes) and Reuters-R8 (8 classes) datasets. We also choose the frequent support (σ) threshold ($minSup$) being 15% for WebKB and 25% for Reuters-R8 test cases (Tables 2 and 3).

Table 2. Experiment for concept/term distributions between GOW-LDA and classical LDA algorithm on WebKB (4 classes) dataset

Topic	GOW-LDA (nTopic = 4, nConcept = 5) (minSup = 15%)	Classical LDA (nTopic = 4, nTerm = 5)
1	{fax, wide}, {principl}, {brian, scienc, comput, principl, depart, updat, colorado, madison,, appli}, {hilton, depart, russel, impagliazzo}, {nation, comput}	student, offic, graduat, phone, engin
2	{brian, comput, depart, arpa}, {brian, nation}, {brian, comput}, {gordon, content}, {hilton, depart,, comput, russel, impagliazzo}	program, assign, class, lectur, homework
3	{brian, comput, depart}, {fax, wide}, {wisconsin, dayton}, {brian, depart, cremer, dayton, comput, scienc}, {denni, perus, cours}	program, design, project, softwar, parallel
4	{principl}, {jolla, colorado, cpsc, comput, fall}, {brian, scienc, cours, comput, depart}, {brian, fax, wide}, {gordon, content}	learn, work, intellig, network, group

Table 3. Experiment for concept/term distributions between GOW-LDA and classical LDA algorithm on Reuters-R8 (8 classes) dataset

Topic	GOW-LDA (nTopic = 8, nConcept = 5) (minSup = 25%)	Classical LDA (nTopic = 8, nTerm = 5)
1	{net, electronics, obod, three}, {net, shr, jan}, {completes, inc}, {cbco, shr}, {plan, four, net, shr}	said, shares, pct, stock, company
2	{entry, inc}, {completes, inc}, {cbco, international}, {completes, inc, split}, {shr, profit}	cts, April, record, dividend, div
3	{net, electronics}, {cbco, year, net, shr}, {net, four, electronics, obod}, {net, shr, profit}, {plan, four, net}	said, company, merger, offer, canadian
4	{net, shr, four}, {cbco, shr, profit}, {cbco, net, year, shr, five}, {cbco, net, year, shr, five}, {net, shr, four, electronics}	said, trade, bank, japan, pct
5	{year, net, shr, five}, {cbco, net, shr, note, year}, {plan, four, net, electronics}, {inc, said}, {cbco, net, jan}	mln, dlrs, said, company, year
6	{year, net, shr}, {shr, profit, five}, {cbco, international, year}, {cbco, international, note, year}, {cbco, year, net, shr}	oil, said, prices, mln, production
7	{shr, jan}, {net, shr, jan}, {loans, note, jan}, {profit, five}, {cbco, shr, year, net}	vs, mln, cts, net, loss
8	{cbco, year, net, shr, five}, {cbco, net, shr, year}, {net, four, shr}, {cbco, net, shr, year}, {cbco, net, shr}	billion, pct, said, year, mln

Evaluation on Text Classification

In these experiments, the GOW-LDA and classical LDA algorithm are implemented simultaneously in different classifying algorithms, including *SVM, Naïve Bayes (MultinomialNB)*, *Decision Tree (J48)* and *K-nearest Neighbors* (Figs. 4 and 5).

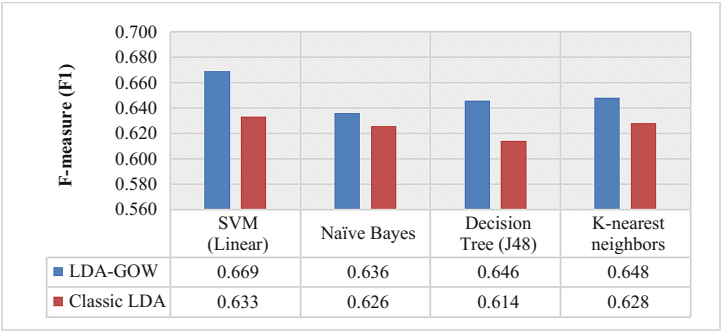


Fig. 4. Comparison between GOW-LDA and LDA in classifying capability with small number of taken topics and concepts/terms

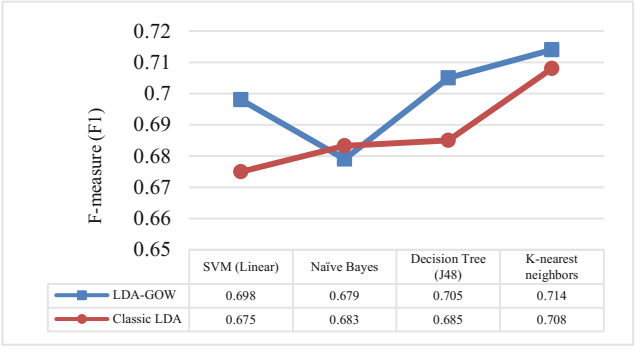


Fig. 5. The illustration for classifying comparison between GOW-LDA and classical LDA algorithm in Reuters-R8 dataset

The output of topic and document distributions is used to generate the vectors' weight which used to feed to classifier. We used W*ebKB* and *Reuters-R8* dataset for these experiments (Table 4 and 5).

4.3 Experimental Concluding Remarks

With our proposed model we get slightly increases in some classifiers such as SVM (*3.295%*), J48 (*2.837%*) and K-nearest (*0.84%*) - for testing on predictive quality of GOW-LDA. In the future, we believed that the measures such as the precision and the recall will be better when we apply NLP and syntactic dependency graph parsing for improving of document representation by GOW-LDA model.

Table 4. Experimental results on classification accuracy between GOW-LDA and classical LDA algorithm with WebKB dataset (4 classes)

Topic	GOW-LDA (nTopic = 4, nConcept = 10, minSup = 15%)			Classical LDA (nTopic = 4, nTerm = 10)		
	P	R	F	P	R	F
SVM (Linear)						
student	0.661	0.792	0.721	0.640	0.870	0.738
faculty	0.743	0.692	0.717	0.518	0.475	0.496
project	0.562	0.212	0.308	0.417	0.05	0.089
course	0.684	0.702	0.693	0.926	0.917	0.921
total/avg	0.677	0.683	0.669	0.644	0.669	0.633
Naïve Bayes (MultinomialNB)						
student	0.620	0.863	0.721	0.616	0.915	0.737
faculty	0.768	0.643	0.700	0.498	0.442	0.468
project	0.500	0.035	0.066	0.700	0.068	0.124
course	0.707	0.656	0.680	0.965	0.854	0.906
total/avg	0.666	0.671	0.636	0.598	0.656	0.626
Decision Tree (J48)						
student	0.694	0.685	0.690	0.646	0.658	0.652
faculty	0.686	0.692	0.689	0.438	0.443	0.440
project	0.405	0.353	0.377	0.380	0.365	0.373
course	0.634	0.689	0.660	0.888	0.861	0.875
total/avg	0.639	0.653	0.646	0.625	0.604	0.614
K-nearest neighbor						
student	0.661	0.723	0.690	0.645	0.794	0.712
faculty	0.702	0.683	0.693	0.456	0.398	0.425
project	0.452	0.388	0.418	0.429	0.239	0.307
course	0.662	0.609	0.634	0.945	0.934	0.939
total/avg	0.643	0.653	0.648	0.626	0.644	0.628

Table 5. Experimental results on classification accuracy between GOW-LDA and classical LDA algorithm with Reuters-R8 dataset (8 classes)

GOW-LDA (nTopic = 30, nConcept = 20, minSup = 25%)			Classical LDA (nTopic = 30, nTerm = 20)		
P	R	F	P	R	F
SVM (Linear)					
0.669	0.743	0.698	0.622	0.749	0.675
Naïve Bayes (MultinomialNB)					
0.624	0.732	0.679	0.631	0.745	0.683
Decision Tree (J48)					
0.701	0.717	0.705	0.685	0.685	0.685
K-nearest neighbor					
0.722	0.723	0.714	0.692	0.731	0.708

5 Conclusion and Future Works

Our ultimate purpose of proposing GOW-LDA model is to overcome the shortcoming of classical LDA algorithm based on BOW model. The proposed novel approach considers applying GOW and frequent subgraph mining for generating the graph-based concepts to feed the classical LDA algorithm instead of using the independently extracted terms within documents. Therefore, the uncovered topic taken from GOW-LDA should be more informative than traditional one.

In the future, we will continue to research thoroughly on the solutions for constructing large-scale complex type of GOW by applying NLP to represent the whole document as syntactic dependency structure. Since the number of vertex and edge of GOW can be hundred thousand so the proposed approach related to distributed and parallel computing should be taken in consideration.

Acknowledgement. This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCMC) under the grant number B2017-26-02.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012). https://doi.org/10.1145/2133806.2133826
3. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 262–272 (2011)

4. Rajagopal, D., Olsher, D., Cambria, E., Kwok, K.: Commonsense-based topic modeling. In: Proceedings KDD WISDOM, vol. 6. ACM (2013)
5. Ferrugento, A., Oliveira, H.G., Alves, A.O., Rodrigues, F.: Can topic modelling benefit from word sense information? In: LREC (2016)
6. Rousseau, F., Vazirgiannis, M.: Graph-of-word and TW-IDF: new approach to ad hoc IR. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM 2013), pp. 59–68. ACM, New York (2013). http://dx.doi.org/10.1145/2505515.2505671
7. Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., Vazirgiannis, M.: Degeneracy-based real-time sub-event detection in Twitter stream. In: ICWSM 2015, pp. 248–257 (2015)
8. Rousseau, F., Kiagias, E., Vazirgiannis, M.: Text categorization as a graph classification problem. In: ACL, vol. 1, pp. 1702–1712 (2015)
9. Yan, X., Han, J.: gSpan: graph-based substructure pattern mining. In: Proceedings 2002 IEEE International Conference on Data Mining, ICDM 2003, pp. 721–724. IEEE (2002)
10. Jun, H., Wei, W., Jan, P.: Efficient mining of frequent subgraphs in the presence of isomorphism. In: Third IEEE International Conference on Data Mining, 2003, ICDM 2003. IEEE, pp. 549–552 (2003)
11. Huan, J., Wang, W., Prins, J., Yang, J.: SPIN: mining maximal frequent subgraphs from graph databases. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 581–586. ACM, New York (2004). http://dx.doi.org/10.1145/1014052.1014123
12. Cordella, L.P., Foggia, P., Sansone, C., Vento, M.: A (Sub)graph isomorphism algorithm for matching large graphs. IEEE Trans. Pattern Anal. Mach. Intell. **26**(10), 1367–1372 (2004). https://doi.org/10.1109/TPAMI.2004.75
13. Shang, H., Zhang, Y., Lin, X., Yu, J.X.: Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. Proc. VLDB Endow. **1**(1), 364–375 (2008). https://doi.org/10.14778/1453856.1453899
14. Lee, J., Han, W.-S., Kasperovics, R., Lee, J.-H.: An in-depth comparison of subgraph isomorphism algorithms in graph databases. Proc. VLDB Endow. **6**(2), 133–144 (2012). https://doi.org/10.14778/2535568.2448946

Topic Discovery Using Frequent Subgraph Mining Approach

Tri Nguyen and Phuc Do(✉)

University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
{tringuyen,phucdo}@uit.edu.vn

Abstract. The topic modeling has long been used to check and explore the content of a document in dataset based on the search for hidden topics within the document. Over the years, many algorithms have evolved based on this model, with major approaches such as "bag-of-words" and vector spaces. These approaches mainly fulfill the search, statistics the frequency of occurrences of words related to the topic of the document, thereby extracting the topic model. However, with these approaches the structure of the sentence, namely the order of words, affects the meaning of the document is often ignored. In this paper, we propose a new approach to exploring the hidden topic of document in dataset using a co-occurrence graph. After that, the frequent subgraph mining algorithm is applied to model the topic. Our goal is to overcome the word order problem from affecting the meaning and topic of the document. Furthermore, we also implemented this model on a large distributed data processing system to speed up the processing of complex mathematical problems in graph, which required many of times to execute.

Keywords: Co-occurrence graph · Frequent subgraph mining
Distributed gSpan · Big data

1 Introduction

Since the early 1990s, the topic model has been proposed and widely researched. One of the most common approaches for modeling the topic of a document is the "bag-of-words" approach. This is a suitable method for calculating the occurrence frequency of words, but the structure and semantics of information are ignored and are not used for calculations. Today, the Latent Dirichlet Allocation (LDA) algorithm [1] is gaining in popularity. However, in addition to efficiency, this algorithm is difficult to work with short texts, or massive data. Besides, the calculation does not use the order of word is also confusing some cases. So, the problem is to find a new approach to modeling topic that preserves the structure of information, and more effectively supports the problem of topic discovery.

Meanwhile, the graphs are mathematical structures that can be used to model relationships between components and to represent the structural information of the data efficiently. A document that can be represented as a graph in which a vertex represents a term and its attributes, the edge of the graph represents the relationship between the

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 432–442, 2018.
https://doi.org/10.1007/978-981-10-8276-4_41

terms. Through representing a document using a graphical model, it provides different solutions regarding the relationship between words or terms in the document such as semantic relationships, co-occurrence… Since there are many different graphical methods for the same purpose of representing documents such as co-occurrence graphs, co-occurrence graphs based on POS tags, semantic graphs, conceptual graphs [2].

In [3], the conceptual graphs (CGs) were used to describe the semantic relationship between concepts in the tutoring system. In the work of Rao and Devi [4], the authors use CGs to create a semantic network between all the sentences in patent document for concept mining and abstractive summarization. Undirected graph based on frequency [5] with vertices and edges are labeled as their occurrence frequency. In addition, there are some other graphical representation models such as non-labeled directed graphs; directed graph with edge is distance n between two words in document (simple n distance model). In this paper, we have chosen a co-occurrence graph for document representation because its structure con-forms to the simple criterion, making it easy to use for big data systems, which is sufficient to store the structure of the information. For more information about the co-occurrence graph, please read in section below.

In order to mine frequent sub-graph in the graphical representation of documents made up of the co-occurrence graph, we use the gSpan algorithm. In a review by Wörlein et al. [6] on the effectiveness of MoFa, gSpan, FFSM and Gaston algorithm, show that the Gas-ton algorithm has the fastest execution time on the test data set, followed by gSpan, FFSM, and MoFa. However, for large datasets with up to tens of thousands of graphs, the gSpan algorithm shows its advantage in using memory and processing speeds. Therefore, we chose this algorithm and will elaborate in section following. The next after finding the frequent subgraphs of each topic, it is necessary to filter out the graphs that can accurately characterize each topic. We used the distance measure between graphs given by Bunke [7] in 1997 to calculate the distance between graphs in the entire frequent subgraph set, thereby identifying which graphs are most characteristic for each topic.

In addition to offering a new approach to discover topics in document, we propose an improvement over the gSpan algorithm, which is parallelization of the algorithm using Apache Spark. During the implementation of the method, we encountered many difficulties in processing large number of graphs, in which each problem on the graph has high computational complexity, requiring improve the parallelization of the gSpan algorithm, in order to provide higher speeds, as well as better use of computing resources.

The rest of this paper is structured as follows: In Sect. 2, we introduce briefly the theory of co-occurrence graph, the graph-based substructure pattern mining algorithm, and the method of measuring the distance between two graphs. Based on that, in Sect. 3, we propose a new approach and explain the implementation step by step. Finally, in Sect. 5, we give some conclusions and future works.

2 Related Work

2.1 Co-occurrence Graph

Co-occurrence graph is graph based on the coexistence of words, terms, or sentences in a document. At present, there have been many studies dealing with the way to build co-occurrence graph, in which typical representations are the way of Hassan et al. [8] introduced in 2007. Each word in the document will be a vertex of the graph and an undirected edge will be added between the two vertices if two words co-occur in a certain window size.

Given a text: "*In an effort to help tackle the stigma surrounding HIV and AIDS, the Queen's grandson, Prince Harry has been tested for HIV. He wants to encourage more people to come forward to find out if they have the virus and his goal is to raise awareness about HIV and AIDS. He confessed he had been a bit nervous before having the simple procedure done.*" *("BBC Learning English", Oct 7th, 2016)*. Figure 1 show the graph constructed for this text, assuming a window size of 2, corresponding to two consecutive terms in the text (e.g. **help** is linked to **tackle**).

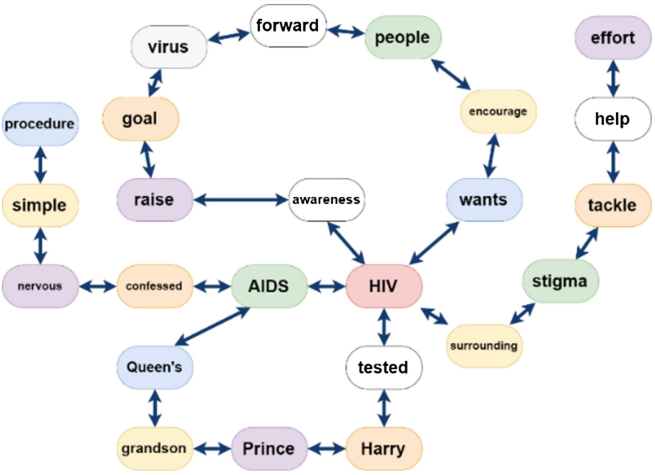


Fig. 1. Example graphs

2.2 Graph-Based Substructure Pattern Mining (gSpan) Algorithm

Frequent Subgraph Mining [9]

Given graph G is a labeled graph, represented by four sets $G = (V, E, L, l)$, where: V is the vertices set of the graph, $E \subseteq V \times V$ is the edge set of the graph, L is the set of vertices and edges label, $l{:}V \cup E \rightarrow L$ is the function that labels the vertices and edges of the graph. An isomorphism between graph G and G' is a binary function $f{:}V(G) \rightarrow V(G')$, such that:

- $\forall u \in V(G), l_G(u) = l_{G'}(f(u))$ and
- $\forall (u,v) \in E(G), (f(u),f(v)) \in E(G)$ and $l_G(u,v) = l_{G'}(f(u),f(v))$

For the graph dataset $\mathbb{GS} = \{G_i | i = 0..n\}$.

$$\varsigma(g, G) = \begin{cases} 1 \textit{ if } g \textit{ is isomorphic to a subgraph of } G, \\ 0 \textit{ if } g \textit{ is not isomorphic to any subgraph of } G. \end{cases}$$

$$\sigma(g, \mathbb{GS}) = \sum_{G_i \in \mathbb{GS}} \varsigma(g, G_i)$$

$\sigma(g, \mathbb{GS})$ denotes the frequency of occurrence of g in $\mathbb{GS}$, or the support of g in $\mathbb{GS}$.

Frequent subgraph mining is to find all graphs, g, such that $\sigma(g, \mathbb{GS})$ is greater than or equal to a given minimum support threshold [9].

gSpan Algorithm

The gSpan [9] algorithm, short for graph-based Substructure pattern mining, was introduced by Yan and Han in 2002. The algorithm is based on a pattern-growth approach, using the depth-first search strategy to browse graphs, find candidates and check for frequent subgraphs. Since then, there have been many scientific works using this algorithm or its derived algorithms to explore frequent subgraphs in a given graph set.

DFS Lexicographic Order. The gSpan algorithm proposes a sorting search code (DFS Code) method to build search code tree (DFS Code Tree), which find out the minimum search code. To be able to determine whether the graphs are isomorphic to each other, find the minimum search code of two graphs and compare them, if the two are equal, then the two graphs are isomorphic.

```
sort labels of the vertices and edges in GS by their frequency;
remove infrequent vertices and edges;
relabel the remaining vertices and edges in descending frequency;
S¹ ←all frequent 1-edge graphs in GS;
sort S¹ in DFS lexicographic order;
S ←S¹
for each edge e ∈ S¹ do
  initialize s with e, set s.GS = {g|∀g ∈ GS, e ∈ E(g)}; (only graph ID
is recorded)
  Subgraph_Mining(GS, S, s);
  GS ←GS – e;
  if |GS| < minSup;
    break;
```

In the pseudo code of the gSpan algorithm above [9], this algorithm has four basic steps. Step 1, from line 1 to line 6, refines and arranges frequent vertices and edges, relabels, adds all 1-edge graphs to S^1 and then sorts them in DFS lexicographic. Next in

Step 2, line 8 and line 9, with each 1-edge graph, `Subgraph_Mining` function creates and checks the candidate from the 1-edge graph. `Subgraph_Mining` function will search the graph in depth first to check all candidates. Step 3, line 10, the algorithm removes the 1-edge graph that has been calculated from all graphs in `GS`. Step 4, line 7 and 11, is the stop condition of the algorithm, when all 1-edge graphs have been checked or graphs do not have enough quantity to exceed the given support threshold. Although not the fastest execution algorithm yet, gSpan is still widely used in research because of its advantages in using resources. During the search process, this algorithm removes candidates from unsuitable search branches of the DFS Code Tree to effectively reduce the search space, avoiding the explosion of candidate numbers, thus dramatically reduce program execution time.

2.3 Distance of Two Graphs

Method of Measuring Distance

After finding a set of frequent subgraphs of each topic, the next issue is to figure out which graphs are specific to a particular topic without confusing other topics. In this paper, the author uses the method Bunke [7] proposed to calculate the distance between two graphs, thereby determining the similarity between the graphs to find out the characteristic graph. Given two graphs G_1 and G_2, the distance between two graphs G_1 and G_2 denoted by $d(G_1, G_2)$ is calculated by the following formula:

$$d(G_1, G_2) = 1 - \frac{\left|mcs\left(G_1, G_2\right)\right|}{\max\left(\left|G_1\right|, \left|G_2\right|\right)}$$

Where $mcs\left(G_1, G_2\right)$ is the maximal common subgraph of the two graphs G_1 and G_2, $|G|$ is the size of the graph, in this paper the size of the graph is the total number of vertices and the number of edges of that graph.

Finding Maximal Common Subgraph

A graph is called a common subgraph of the two given graphs if it is a subgraph isomorphism of both graphs. A common subgraph G of G_1 and G_2 is maximal if there exists no other common subgraph G' of the G_1 and G_2 that has more vertices and edges than G. In Fig. 2a and b, to measure the distance between two graphs, we need to find the maximal common subgraph. The graph in Fig. 3 is the maximal common subgraph because it has the largest number of vertices and edges. The size of the subgraph in Fig. 3 is 5. Then, the distance between these two graphs in Fig. 2a and b will be:

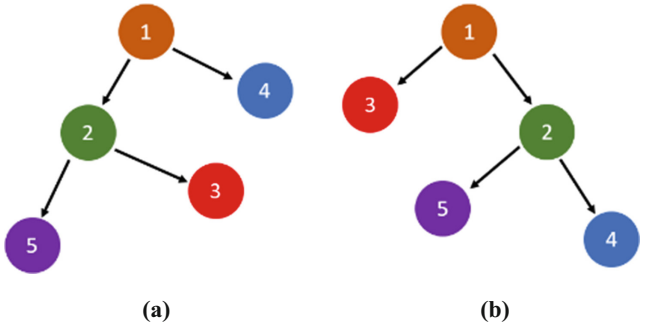


Fig. 2. Example graphs

Fig. 3. Maximal common subgraphs of our example graph a and b in Fig. 2.

$$d(G_1, G_2) = 1 - \frac{\left|mcs(G_1, G_2)\right|}{\max\left(\left|G_1\right|, \left|G_2\right|\right)} = 1 - \frac{5}{\max(9,9)} = 1 - \frac{5}{9} \approx 0{,}444$$

Since it is a NP-complete problem, we use the gSpan algorithm to find the maximal common subgraph. Details of implementation will be described later.

3 Proposed Method

3.1 Overview

In this paper, we propose a new approach to modeling the topic of document datasets. This approach includes the basic processing steps described in the following Fig. 4. These steps will be elaborated later.

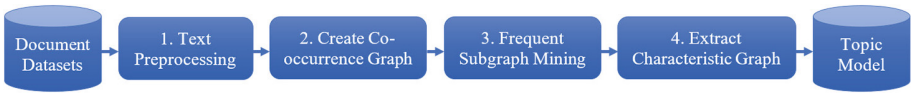


Fig. 4. Process flow diagram of the new approach

In the figure above, we describe the process of implementing the method as follows. The text input is preprocessed and produces co-occurrence graphs, and then uses the distributed gSpan algorithm to explore the frequent subgraph contained in it. Similarly,

for each topic, we will have a set of subgraphs corresponding to those topics. Next, to extract subgraphs specific to each topic, we calculate the distance between each graph in the subgraph of it with all subgraphs of the remaining topics. If its largest distances to other graphs are less than a given threshold, we will discard that graph. Only graphs that have sufficiently large distances, which means that the similarity between them and other graphs is small, makes that graph unique, which may characterize the topic under consideration, will selected into the characteristic subgraph set.

3.2 Representing a Document Using a Graphical Model

In this paper, we use a directed graph to preserve the order of the words in the document, in which the co-occurrence graph is a simple, easy-to-build graph but fully responsive applicable for the purpose of storing the order of words.

First, an input document will be tokenized, then, the document will be removed the stop words and special characters. Next, we give a sliding window size of 2 sliding through the document. Which words occur in the same window will have an edge connect them, in the direction of the first word to the next. So, the vertices of graph are all the word of the document. This will help us to represent the order of the words in the document by the direction of the edges of the graph. The sliding window has a size of 2 because our goal is to create an adjacency graph. If the window size increases, then the number of edges created between the vertices will increase dramatically, making the computation complexity much higher.

For example, we have a text: "*Khosrowshahi, who led travel-booking site Expedia Inc. for 12 years, made the remarks as he introduced himself to Uber's workforce on Wednesday during an all staff meeting at its San Francisco headquarters." (Heather S. and Tom H., "Reuters Technology News", August 30th, 2017)*. This text will be preprocessed before being used to create the graph. It will become: *"Khosrowshahi, led travel-booking site Expedia Inc. 12 years, remarks introduced Uber's workforce Wednesday staff meeting San Francisco headquarters"*.

In Fig. 5, we can see a sliding window with size of 2 sliding along the words in the text, with each sliding, the words appearing in the same window will become the vertices of the graph. And an edge between the two words will be added, which the direction is from the word before to the one after.

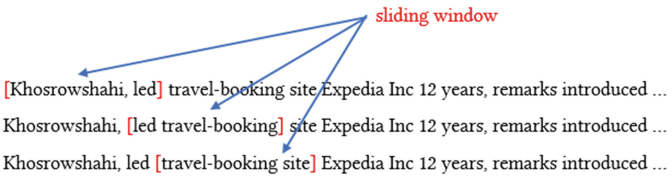


Fig. 5. Sliding window size of 2 slide through the given text.

After performing the sliding of the window, we obtain the co-occurrence graph shown in Fig. 6. Thereby, we can see, by this way, the order of the words in the text is stored correctly.

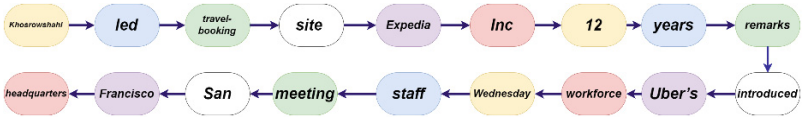


Fig. 6. Co-occurrence graph is created by sliding the window along the words in the example text

3.3 Distributed gSpan for Frequent Subgraph Mining

From the set of graphical representation of the document obtained in the preceding steps, we apply the gSpan algorithm to explore the frequent subgraphs in it. The original gSpan algorithm was created by the author to handle undirected graphs, but this algorithm is considered to be easily scalable to work with directed graphs. In addition to this improvement, we also parallelized the gSpan algorithm using Apache Spark, which aims to accelerate computing, as well as process large data. Because most of the problems on the graphs are extremely complex, rendering the processing time very slow, as well as not handling large amounts of data.

We split the processing steps of the gSpan algorithm for parallel distributed processing as illustrated in the diagram in Fig. 7. We can see, first of all, that input graphs will be processed parallel on the slave, which will remove infrequent vertices, reassign labels and sort in descending frequency, after that, the frequent 1-edge subgraphs are created and sorted in DFS lexicographic order.

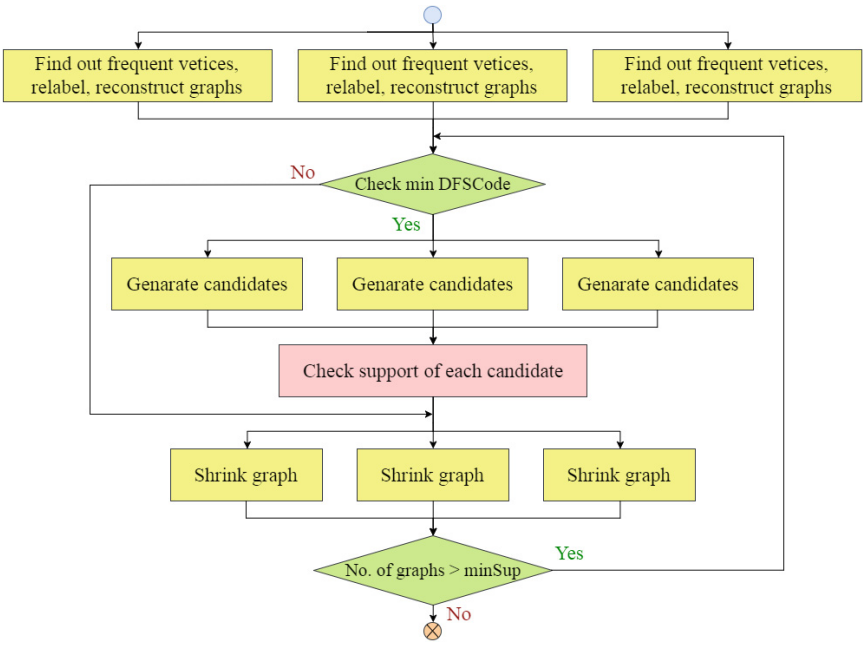


Fig. 7. The parallel distributed gSpan algorithm

Then, for each frequent subgraph, it will be checked if it is a minimum DFS Code on the master. Next, from the frequent subgraph that has been checked, the slave will simultaneously generate candidate graphs. All that subgraphs will be checked for a frequent subgraph. Finally, after searching in the 1-edge subgraphs, these edges are shrunk from the graphs in the dataset. This process will be performed simultaneously in parallel on the slave. All results will be returned to the master to check the termination criterion.

3.4 Modeling Topic by Extract Characteristic Graphs

After figuring out the frequent subgraph set for each topic of the document dataset, we proceeded to measure the distance between the graphs to eliminate graphs that are similar to each other, cannot be used as characterizes a particular topic. To find the maximal common subgraph of the two graphs for the distance calculation problem, we use the gSpan algorithm, with a minimum support threshold of 1; we find the subgraphs appear in all the given two graphs, it is the set of common subgraphs, then proceed to calculate the size of all the subgraphs in that set to find the maximal common subgraph. Based on the above measuring distance formula, we remove the graphs with high similarity, all the remaining graphs will represent a topic.

4 Experiments

To evaluate the value of the proposed approach, we have implemented the program with Reuters-21578 Text Categorization Collection Data Set of David D. Lewis at AT&T Labs - Research. The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. and Carnegie Group, Inc. in 1987. We run this algorithm in UIT-Cloud with 9 nodes and 1 master.

We discovered topics after implementation ours approach with this data and we have some set of frequent subgraphs in the Table 1 below.

Table 1. Number of frequent subgraphs of each topic

Topic	No. of frequent subgraphs	Topic	No. of frequent subgraphs	Topic	No. of frequent subgraphs
Bop	58	Nat-gas	38	Lei	205
Coffee	72	Orange	84	Lin-oil	101
Gnp	83	Palm-oil	59	Trade	118
Gold	2	Reserves	45	Wpi	107
Heat	168	Retail	131	Yen	104
Interest	90	Rubber	82	Money-supply	55
Ipi	65	Ship	37		

In Figs. 8 and 9, we show some topical frequent subgraphs that we have discovered in topic Orange and Heat from Reuters dataset. We manually check the words and the relationships between words to confirm that they define a good meaningful concept.

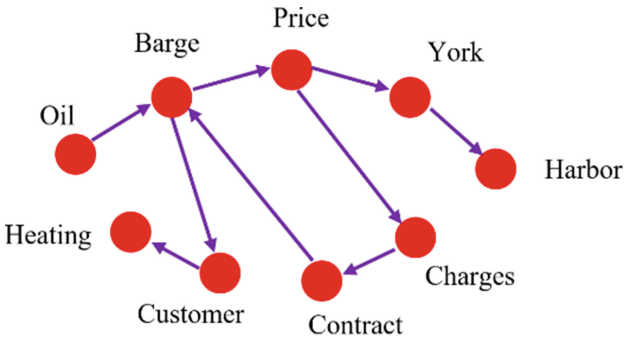


Fig. 8. The frequent directed subgraph of topic Heat

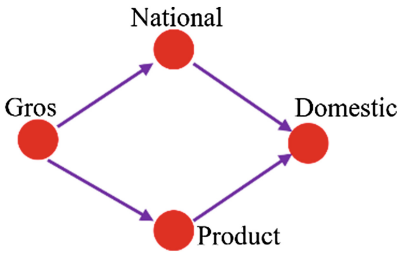


Fig. 9. The frequent directed subgraph of topic Orange

5 Conclusion and Future Works

In this paper, we propose a new approach for discovering topics of document. First, we represent document by co-occurrence graph, then we use distributed gSpan to mine frequent subgraphs and extract characteristic graphs for representing each topic. Based on our method, we can discover the representation of topics by graph in which we have the words and the semantic relation between words. This way, we can solve the weakness of representing document by bag of word with missing the order of words. In the future, we plan to use this approach to discover the latent topic in copus. We also build an ontology for labeling the discover topic.

Acknowledgments. This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCMC) under the grant number B2017-26-02.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Sonawane, S.S., Kulkarni, P.A.: Graph based representation and analysis of text document: a survey of techniques. Int. J. Comput. Appl. **96**(19), 1–8 (2014)
3. Khodeir, N.: Graphical representation in tutoring systems. Int. J. Comput. Sci. Inf. Technol. **9**(3), 107–116 (2017)
4. Rao, P.R., Devi, S.L.: Patent document summarization using conceptual graphs. Int. J. Nat. Lang. Comput. **6**(3), 15–32 (2017)
5. Tomita, J., Nakawatase, H., Ishii, M.: Graph-based text database for knowledge discovery. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers and Posters, New York, NY, USA (2004)
6. Wörlein, M., Meinl, T., Fischer, I., Philippsen, M.: A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM, and Gaston. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 392–403. Springer, Heidelberg (2005)
7. Bunke, H., Shearer, K.: A graph distance metric based on the maximal common subgraph. Pattern Recog. Lett. **19**(3), 255–259 (1998)
8. Hassan, S., Mihalcea, R., Banea, C.: Random walk term weighting for improved text classification. Int. J. Semant. Comput. **1**(4), 421–439 (2007)
9. Yan, X., Han, J.: gSpan: graph-based substructure pattern mining. In: Proceedings of the IEEE International Conference on Data Mining, 2002, ICDM 2003, pp. 721–724. IEEE (2002)

Creating Prior-Knowledge of Source-LDA for Topic Discovery in Citation Network

Ho Duy Tri Nguyen, Trac Thuc Nguyen, and Phuc Do(✉)

University of Information Technology, VNU-HCM,
Ho Chi Minh City, Vietnam
{trinhd, thucnt, phucdo}@uit.edu.vn

Abstract. Discovering and understanding the development of research topics in the community is useful for identifying important milestones and prominent researches. Recent works related to detect topics from scientific corpus also used the latent Dirichlet Allocation (LDA) to explore topics of papers. These systems usually used abstract of papers as the corpus instead of full papers. However, the LDA is based on the bag-of-words model so with such short texts it will give low accuracy. The tendency for improvement is to add prior knowledge to the analysis process with the latest algorithm, Source-LDA, which was presented by Justin Wood et al. at UCLA in 2017. We found that the Source-LDA has some shortcomings to overcome. Firstly, it is also based on counting method as LDA so short text will decrease the accuracy. Secondly, the knowledge source mentioned in the algorithm is constructed manually from labeled text data. This make Source-LDA becomes a supervised method. Therefore, we propose an approach to automatically construct knowledge source for Source-LDA from unlabeled data with an assumption that a specific paper will often cite papers which contain related topics. This approach both helps to integrate source knowledge in an unsupervised manner and resolve the issue of short text by using information from citation network. In the first stage, the propound method has achieved encouraging results.

Keywords: Citation network · Topic modeling · Source-LDA
Knowledge source · LDA model

1 Introduction

One of the difficulties in research productivity is that as the community grows, it will be very difficult for researchers to see the complete picture of how a field develops, given the fact that new works are being written on previous works. Young researchers can often get lost in a large number of papers. Researchers move on to a new topic that will take a long time to find related works in the field. All of this clearly hindered the advancement of scientific research. And if developing mining techniques that make it easier and more efficient for researchers to understand research topics will be very beneficial. Therefore, leveraging information technology to improve the productivity of scientific research is a very important challenge that has a huge impact on society. With the development history presented below, the computer was able to support the analysis of related documents,

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 443–453, 2018.
https://doi.org/10.1007/978-981-10-8276-4_42

thereby discovering the topic of research and the theme evolution in a community. However, the tools and models are still limited and have disadvantages [1].

Since 1990s, the Topic Modeling of Deerwester et al. [4] has enabled computers to mine text corpora based on mathematical statistical tools. It allows one to examine and exploit documents based on identify and statistical frequency of the occurrence of words related to topic in each document. Over 27 years of development, from the single-topic hypothesis (a document has only one topic) to the multi-topic hypothesis (a document contains a set of topics mixture), this model is more efficiency in exploring the hidden topics in the set of documents. One of the most effective topic modeling algorithms is the LDA. So far this model has been studied, applied and improved by many researchers. Because the LDA is a model assembled from multiple modules, so each part can easily be replaced or expanded. Researchers focus on extending largely the modeling of relationships between topics. The most thoughtful improvement is replacing the Dirichlet distribution with another distribution. Typically, the Correlated Topic Model [5] adopts this alternative, which uses standard logistic distribution instead of Dirichlet distribution. Another innovative approach is hierarchical LDA (hLDA) [6]. In this model, topics are linked together in a hierarchy using the Chinese Restaurant process. The paper [7] presents a Hidden Factors as Topics (HFT) method. It is a combination of hidden factor model and the LDA to fully exploit the user's judgment, both scores and feedback, on the user's product rating file. The hidden factor model helps analyze the customer score for a product. The LDA model helps find keywords and topics that express the customer's perspective on the product through the feedback received. Thenceforth, it is possible to develop a recommendation system to introduce more accurate products to the user's tastes and improve the efficiency of e-commerce services.

One of the latest improvements to the LDA model is Source-LDA [3]. This algorithm applies prior knowledge through knowledge sources into the topic sampling process to improve model stability and accuracy. However, this work only provides a method for building knowledge sources manually on labeled data sets. To address these limitations, in this paper, we will attempt to propose an approach to build automated knowledge sources with unlabeled data through the relationship of papers in the citation network. The idea of the citation network used to explore topics more effectively derived from its characteristics, besides the abstract of the paper, the summary from cited documents will also contain topics that paper mentioned. We rely on these characteristics to hypothesize the most topics mentioned by the paper contained in the topic set of the documents it cites. The Source-LDA analysis will find more hidden topics in the paper and filter out inappropriate topics that has in the knowledge source. Therefore, after the implementation, we will retain the most appropriate topics. The goal of this work is to automatically discover hidden topics in scientific publications. A summary of the contributions are:

1. We propose a novel technique to topic modeling in unsupervised fashion which uses the prior knowledge from the references.
2. We introduce an approach that allows automatically generating knowledge sources of Source-LDA from unlabeled data.

3. We show how to find the appropriate topics in a corpus based on the content that the document references.
4. We explain how to use the characteristics of the citation networks to explore topics more effectively.

The rest of this paper is structured as follows: In Sect. 2, we give a brief overview of citation network, the theory of knowledge source and the Source-LDA algorithm. Based on that, we introduced our proposed approach step by step with details description in Sect. 3. In Sect. 4, we tested and compared a result of LDA and Source-LDA models using knowledge sources built on the proposed methodology. Finally, Sect. 5 gives the conclusion and future works of this paper.

2 Related Works

2.1 Citation Network

Citation network is a graph which contains information of paper in each vertex and an edge is citation relationship between them. A vertex attributes are details about the paper such as id, publication year, abstract, keywords, content… And when the paper p_i references to paper p_j, we will have an arrow going from the vertex representing p_i to the vertex representing p_j. Therefore, the citation network has some following characteristics:

- The citation network is a directed graph which each edge is an arrow go from one paper to the other.
- All citation arrows almost always point backwards in time to older papers. Therefore, the graph of citation network is acyclic and it also shows the development of research field over time.
- The most important characteristic of the citation network is the close relevance between topics of paper and mentioned topics of the others which it cites.

Figure 1 is an example of the citation network, which p_1, p_2,…, p_{11} are scientific papers. p_1 is in the center of the graph and it references to papers p_2, p_3,…, p_{11} on around it. So we have arrows pointing from p_1 to p_2, p_3,…, p_{11}. We can know that p_1 is the latest paper because it references all the other ones in the network and the topics contained in p_1 are closely related to the topic in p_2, p_3,…, p_{11}.

In this paper, we use topic related features among papers in the citation network to assist in finding hidden topics in a paper.

2.2 Knowledge Source and Source-LDA

Published in 2017 at UCLA, Source-LDA is the effective improvement model of LDA by integrating prior knowledge in sampling from Dirichlet distribution of topic-to-word. Since we are going to develop a new method to build the knowledge source for Source-LDA algorithm, it is worth giving a brief overview of this model and the theory of knowledge source.

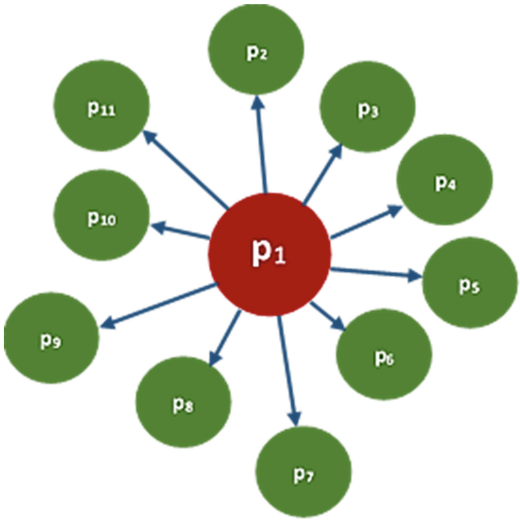


Fig. 1. Example of the citation network

Knowledge Source. By using additional prior knowledge through knowledge sources, Source-LDA has the ability to explore hidden topics (unknown topics) and known topics in the text corpora. This additional knowledge includes a list of topics that is represented via bag-of-words and their probability distribution. Source-LDA will filter out subsets of known topics and explore more latent topic mixtured in the document creation.

In [3], the author provides three definitions for the construction and integration of knowledge sources into the topic model:

Definition 2.1 (Knowledge Source): A knowledge source is a collection of documents that are focused on describing a set of concepts.

Definition 2.2 (Source Distribution): The source distribution is a discrete probability distribution over the words of a document describing a topic. The probability mass function is given by

$$\forall w_i \in W, f(w_i) = \frac{n_{w_i}}{\sum_j^G n_{w_i}} \tag{1}$$

where W is the set of all words in the document, G = |W|, and n_{w_i} is the number of times word w_i appears in the document.

Definition 2.3 (Source Hyperparameters): For a given document in a knowledge source the knowledge source hyperparameters are defined by the vector $X_1, X_2, \ldots, X_V$ where $n_{w_i} + \epsilon$ and ϵ is a very small positive number that allows for non-zero probability draws from the Dirichlet distribution. V is the size of the vocabulary of the corpus for which we are topic modeling, and n_{w_i} is the number of times the word w_i from the corpus vocabulary appears in the knowledge source document.

Source-LDA. Expanded from LDA, the Source-LDA [3] is also a three-level hierarchical Bayesian model which uses Dirichlet priors to approximate the intractable latent variables. Accordingly, each document is a collection of randomly topics and each topic is a discrete distribution of a set of vocabulary. The formal definition of the Source-LDA generative model over a corpus is:

1. **for each** topic $t \in [1, T]$ **do**
2. **if** $t \leq K$ **then**
3. Choose $\phi_t \sim Dir(\beta)$
4. **else**
5. $\delta_t \leftarrow (X_{t,1}, X_{t,2}, \ldots, X_{t,V})$
6. Choose $\phi_t \sim Dir(\delta_t)$
7. **for each** document $d \in D$ **do**
8. Choose $N_d \sim Poiss(\zeta)$
9. Choose $\theta_d \sim Dir(\alpha)$
10. **for each** word $w_{n,d}$ **do**
11. Choose $z_{n,d} \sim Multinomial(\theta)$
12. Choose $w_{n,d} \sim Multinomial(\phi_{z_{n,d}})$

The topic sampling process will be divided into two stages. In the first stage (line 1 to 6), unlike the LDA, for unknown topics (1 to K), the word-to-topic distribution ϕ_t is drawn from Dirichlet distribution and for known topics (K+1 to T), ϕ_t is drawn from Dirichlet distribution of source hyperparameters of a respective topic. And then, in the second stages (line 7 to 12), the algorithm will execute Gibbs sampling like LDA [2, 8].

3 Proposed Approach

The method that Wood et al. used to build knowledge source and source distribution has disadvantages: (1) selected documents for knowledge source are manually chosen so the final quality depends on experts; (2) the proposed probability mass function which is based on counting methods may be too simple to stand for a topic. Therefore, we propose a novel method to build knowledge source and source distribution for Source-LDA. Our approach which is based on citation network to choose the knowledge source will make this process unsupervised. This method supposes that references of a paper usually consist of papers that are contained related topics.

3.1 Data Extraction to Build Knowledge Sources

In order to build an effective knowledge source, to better support the topic modeling process, we must grasp the similarities between the topics that are being addressed. If the topics contained in the knowledge source are not mentioned in the documents, this makes the use of prior knowledge to modeling more costly than the normal LDA.

Therefore, in this work, we selected scientific paper data from the citation network. Because of the very important nature of this one, there is a similarity between topics of paper and mentioned topics of the others which it cites. This feature helps us to assert

the closeness in the topics between the papers being analyzed and the documents it references. The abstract content of the cited papers is prior knowledge dataset. They are used to automatically create the knowledge source, used in the Source-LDA algorithm during the topics exploration.

The data used to build the knowledge source are those referenced by the papers in the test data set and those ones must also be in the same research field. A small example of test data and knowledge source data in citation network is depicted in Fig. 2.

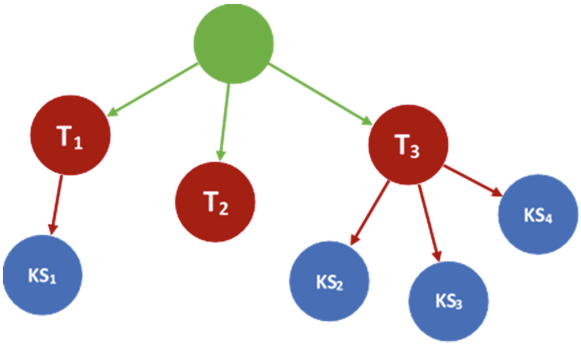


Fig. 2. A small example of test data and knowledge source data in citation network

For example, in Fig. 2, all vertices are articles, but the test data is denoted as T_1, T_2, T_3 and the knowledge source data is denoted as KS_1, KS_2, KS_3, KS_4. Then, we use T_1, T_2, T_3 for testing and KS_1, KS_2, KS_3, KS_4 for building knowledge source.

3.2 The Method of Automatically Building Knowledge Sources

The prior knowledge dataset extracted from the citation network as described above can be treated as a corpus that has not been labeled for the topic. We propound exploring hidden topics from this dataset using the LDA algorithm. The input to the automated knowledge-building model is the set of documents that contains the prior knowledge, the number of topics to be explored, and the input parameters of the LDA algorithm. At the output of this algorithm, the topics will be represented by a bag-of-words. Each topic will include a set of keywords belonging to that topic, each word is followed by the corresponding probability Dirichlet distribution. This result is the source distribution of words in the knowledge source as mentioned in Definition 2.2. From this source distribution we can construct source hyperparameter based on the Definition 2.3 as follows:

- A source hyperparameter is a collection of vectors.
- Each vector in the source hyperparameter will deputize a topic.
- The composition of each vector is the word and its probability distribution in a pertained topic.

Thus, we can store the source hyperparameter that represents the knowledge source in the structure described in Fig. 3.

$t_1\ w_{11}\ \phi_{t_1,w_{11}}\ w_{21}\ \phi_{t_1,w_{21}}\ w_{31}\ \phi_{t_1,w_{31}}\ w_{41}\ \phi_{t_1,w_{41}} \cdots w_{n1}\ \phi_{t_1,w_{n1}}$

$t_2\ w_{12}\ \phi_{t_2,w_{12}}\ w_{22}\ \phi_{t_2,w_{22}}\ w_{32}\ \phi_{t_2,w_{32}}\ w_{42}\ \phi_{t_2,w_{42}} \cdots w_{n2}\ \phi_{t_2,w_{n2}}$

$t_3\ w_{13}\ \phi_{t_3,w_{13}}\ w_{23}\ \phi_{t_3,w_{23}}\ w_{33}\ \phi_{t_3,w_{33}}\ w_{43}\ \phi_{t_3,w_{43}} \cdots w_{n3}\ \phi_{t_3,w_{n3}}$

$\vdots$

$t_k\ w_{1k}\ \phi_{t_k,w_{1k}}\ w_{2k}\ \phi_{t_k,w_{2k}}\ w_{3k}\ \phi_{t_k,w_{3k}}\ w_{4k}\ \phi_{t_k,w_{4k}} \cdots w_{nk}\ \phi_{t_k,w_{nk}}$

Fig. 3. Structure of the file store the source hyperparameter that represents the knowledge source

In Fig. 3, each line represents a topic t, the first element of each line is the topic name t_i followed by pairs of a word w_{ji} and a probability distribution $\phi_{t_i,w_{ji}}$ separated by a whitespace. The Source-LDA algorithm will read this file to analyze a testing corpus.

3.3 The Source-LDA Model Uses Automatic Knowledge Sources

After obtaining the source hyperparameter, we execute the topic modeling with the Source-LDA algorithm using the knowledge source that has been built automatically. The output of the algorithm will be the matrix of the topic-to-document distribution and word-to-topic distribution. So that, the pseudo-code of our proposed approach is:

```
for each topic t ∈ [1,T] do
    if t ≤ K then
        Choose φ_t ~ Dir(β)
    else
        δ_t ← auto_build_knowledge_soure()
        Choose φ_t ~ Dir(δ_t)
for each document d ∈ D do
    Choose N_d ~ Poiss(ζ)
    Choose θ_d ~ Dir(α)
    for each word w_{n,d} do
        Choose z_{n,d} ~ Multinomial(θ)
        Choose w_{n,d} ~ Multinomial(φ_{z_{n,d}})
```

The *auto_build_knowledge_source*() function is:

```
1.  for each topic k ∈ [1,T] do
2.      Choose φ_k ~ Dir(β)
3.  for each document d ∈ D_cited do
4.      Choose N_d ~ Poiss(ζ)
5.      Choose θ_d ~ Dir(α)
6.      for each word w_{n,d} do
7.          Choose z_{n,d} ~ Multinomial(θ)
8.          Choose w_{n,d} ~ Multinomial(φ_{z_{n,d}})
9.  ks ← format(φ_z)
10. return ks
```

In the above code, α, β are the pre-defined parameters of Dirichlet distribution, K is the number of unknown topic, $T - K$ is number of known topics and θ is an array of probability distribution of topic-document, ϕ is an array of probability distribution of word-topic, ks is represented the knowledge source.

Comparing with the traditional Source-LDA, we make changes in line 5. In this line, we use an *auto_build_knowledge_source* function to automatically create the knowledge source based on reference papers of citation network.

The *auto_build_knowledge_source* function use LDA model to analyze topics of abstract data D_{cited} which is cited by papers in D. And then, we use a probability distribution of word-topic that is the output of LDA algorithm to build knowledge source.

The whole proposed model is represented by Fig. 4.

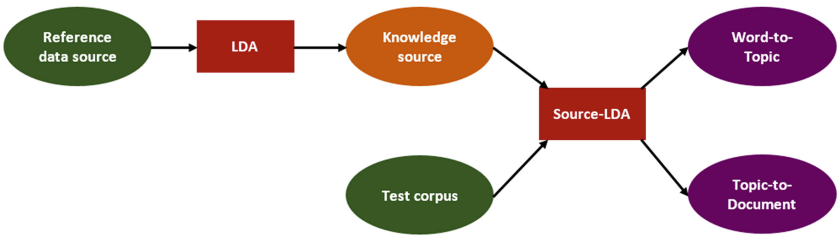


Fig. 4. The Source-LDA model uses automatic knowledge sources

According to Fig. 4, we have the following steps:

- Step 1: use references of documents in the test corpus as the input data source of the LDA algorithm. The output of this process is the structured knowledge source described in Fig. 3.
- Step 2: apply the knowledge source in step 1 to topic modeling the test corpus using the Source-LDA algorithm.

4 Experiments

To evaluate the effectiveness of the model, we tested and compared a result of LDA and proposed model using knowledge sources built on our methodology. In this experiment, we use dataset of academic social network (ASN) which is published at https://aminer.org/data by Aminer research group. This dataset comprises of 2,092,356 papers together with 8,024,869 citations. We conducted experiments on the training data of 14.327 abstracts paper extracted from the ASN. These data will be used to build the knowledge source for the Source-LDA algorithm. The number of topics chosen to test is 50.

The test data includes 806 abstract papers also extracted from the ASN; these are ones that reference to the 14327 papers in the training data set.

The paper uses the Kullback–Leibler divergence measurement to examine the variability of the word-to-topic probability distribution when using LDA and the proposed method. And we uses perplexity to measure a how well a probability model predicts a sample. The perplexity is a measure of language model performance based on average probability can be developed within the field of information theory [9].

With the number of iterations increasing from 100 to 2000, Kullback–Leibler divergence from proposed model (Source-LDA) to LDA is represented in Fig. 5.

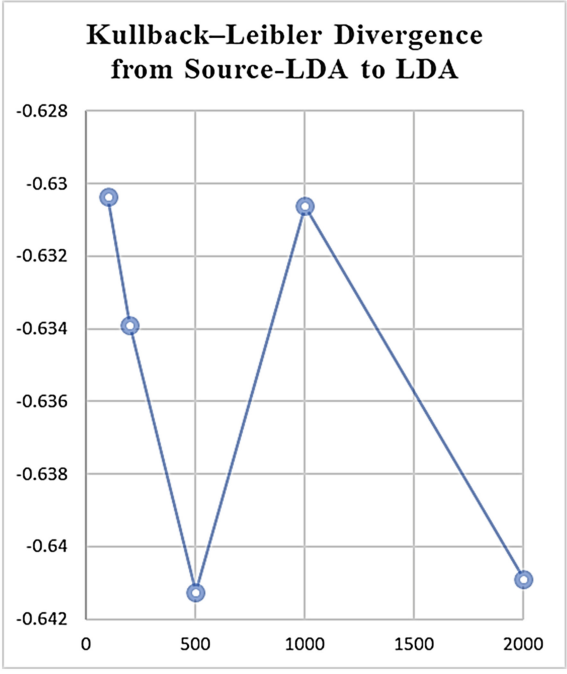


Fig. 5. Kullback–Leibler divergence from proposed model (Source-LDA) to LDA

The value of Kullback–Leibler divergence from proposed model (Source-LDA) to LDA is approximately equal to –0,635. This shows a large difference in the word-to-topic probability distribution in the results of two models.

Then, the perplexity of LDA and proposed model is represented in Table 1.

Table 1. Comparision of perplexity of LDA and proposed model (Source-LDA)

No. of iterations	LDA	Source-LDA
100.00	2831.437784	610.8790888
200.00	2817.622729	608.3100205
500.00	2838.181254	604.2790949
1000.00	2773.8472	604.2475382
2000.00	2788.372726	603.3165104

The perplexity value of the proposed approach is always between 603 and 611, while the perplexity value of the normal LDA algorithm ranges from 2773 to 2839. This proves the high stability of our model and its effectiveness in analyzing papers' topic on citation networks.

All of these results suggest that our working produces significantly different results and the model are much more stable than traditional LDA.

5 Conclusion and Future Works

In this paper, we propose a method for constructing a knowledge source for the Source-LDA algorithm. This method can automatically create knowledge source by using citation network. The data used to build the knowledge source are those referenced by the papers in citation network. Our proposed method has significant results when comparing with the source-LDA algorithm in citation network.

In the future, we plan to use our method for topic discovery of papers in very large citation network. In this case, a version of Spark based source-LDA will be developed with knowledge source created from our proposed method.

References

1. Wang, X., Zhai, C., Roth, D.: Understanding evolution of research themes: a probabilistic generative model for citations. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 11–14 August 2013 (2013)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Wood, J., et al.: Source-LDA: enhancing probabilistic topic models using prior knowledge sources. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 411–422. IEEE (2017)

4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
5. Blei, D., Lafferty, J.: Correlated topic models. In: Advances in Neural Information Processing Systems, vol. 18, p. 147 (2006)
6. Griffiths, D., Tenenbaum, M.: Hierarchical topic models and the nested chinese restaurant process. In: Advances in Neural Information Processing Systems, vol. 16, p. 17 (2004)
7. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: 2013 Proceedings of the 7th ACM Conference on Recommender Systems, pp. 165–172. ACM, New York (2013)
8. Griffiths, T.: Gibbs sampling in the generative model of latent dirichlet allocation. Standford University (2002)
9. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948). 623–656

The Study of Genetic Algorithm Approach to Solving University Course Timetabling Problem

Kuan Yik Junn[1], Joe Henry Obit[1(✉)], and Rayner Alfred[2(✉)]

[1] Universiti Malaysia Sabah, Labuan International Campus, W. P. Labuan, Malaysia
kyj_anderson@hotmail.com, joehenryobit@gmail.com
[2] Knowledge Technology Research Unit (KTRU), Faculty of Computing and Informatics, Universiti Malaysia Sabah (UMS), Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
ralfred@ums.edu.my

Abstract. This research presents the metaheuristic strategy to solve educational timetabling problem. The metaheuristic described in this research highlight the role of Genetic Algorithm (GA) when the algorithm improves the quality of solution by performing genetic operators. Two datasets of university course timetabling are used whereby the datasets are obtained from Universiti Malaysia Sabah Labuan International Campus (UMSLIC). The research experiment is conducted by comparing the quality of solutions produced by Genetic Algorithm with other metaheuristics which have been done in the past researches. The experimental results suggest that Genetic Algorithm manages to produces good solutions in this domain although other algorithms are able to improve the quality of the solutions.

Keywords: Genetic Algorithm · Meta-heuristics
University course timetabling problem

1 Introduction

Timetabling is among the common scheduling problems which are difficult to solve by any conventional method in polynomial time as the problem size increase exponentially against the optimal solutions [2]. A general algorithm used for a particular problem may be incapable for another problem due to certain unique constraints which are required in a specific problem instances. In the university course timetabling problem (UCTP), an event (or the term subject or course) has to be allocated into a limited number of time slots and room, at the same time various constraints need to be satisfied. It is very difficult to search for a general and effective approach for this problem because of multiple variances of constraints, diversification of the problems and special requirements from each institution [11]. Hence, there is no deterministic polynomial time algorithm for this problem since it is an NP-hard combinatorial optimization problem. A guided local search GA has been proposed to solve UCTP in the past few years [8]. One of the important concepts of GA is the population of chromosomes. This algorithm relies on a population of candidate solutions [1]. Therefore, the guided search strategy relies on

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 454–463, 2018.
https://doi.org/10.1007/978-981-10-8276-4_43

the data structures which use to store the chromosomes and extract useful information such as best individual from the populations.

2 Related Works

The enhanced performance of GA by modifying genetic operator or heuristics operators tend to improve the quality of solutions with various combinations of local-area-based and global-area based algorithms [17]. Some believes that by adapting the suitable crossover rate and mutation rates in GA [15], the fitness function may increase exponentially across generations.

Most of the crossover operator tends to create damaged chromosomes during crossover between chromosomes. Hence, repair function is necessary in order to fix the damaged chromosomes. Several repair techniques are introduced [3] such as remove infeasible timetables, apply high penalty in fitness function, so that the chromosome will not survive and also repairing from the genes themselves (timetable). The repairing techniques [3] consist of four important steps which begin with finding the free slots in course timetabling. Next, it find the free time slots for each event and then search for feasible time slots for rooms and events without conflict. Finally, the last movement involve with repairing infeasible course timetable by evaluating the fitness function of the chromosomes.

The repair function of GA in this research involves cross checking the genes of each chromosomes before it swap the genes to another chromosome. The evaluation of fitness function will be performed after the crossover and mutation operators. The experiment is further investigated by comparing the performance between other meta-heuristics search methodology such as Great Deluge (GD) [5] and Simulated Annealing (SA) algorithm [4]. The purpose of this research is to study the performance between specific metaheuristics in UCTP. In order to test the performance of the algorithm, an experiment is carried out with Analysis of Variance (ANOVA) test.

3 Problem Background

The domain studied in this research focus on UMSLIC. The sets of constraints are usually divided into two categories: hard constraints and soft constraints. Generally, hard constraints are important than soft constraints. The objective of this research is to satisfy all the hard constraints and minimize the violation of soft constraints. The datasets are obtained from the Academic Service Division (ASD) of UMSLIC.

The aim of this paper is to produce a feasible and high quality timetabling solution for the institution. The hard constraints for the UMSLIC course timetabling are as follow:

1. (H_{C_1}) No student should be attending more than one class at a time.
2. (H_{C_2}) The total number of students enrolled in each course should be smaller than or equal to the capacity of the room assigned.
3. (H_{C_3}) Not more than one course is assigned to a specific room at the same time.

4. (H_{C_4}) Each type of course offered should be assigned to a predefined period of timeslots only.

Meanwhile, the soft constraints for the UMSLIC course timetabling are as follow:

1. (S_{C_1}) A student should not attend more than two consecutive classes.
2. (S_{C_2}) The size of course enrolment should be utilized as much as possible into the capacity of the room by placing the particular course into the specific room.

The course timetabling problem is completely varies from one institution to another in which every institution has their own sets of constraints in order to fully utilize the resources [9]. For instance, the UMSLIC's datasets summary is stated as follow (Table 1):

Table 1. Summary of datasets in UMSLIC.

Datasets	No. of students	No. of courses	No. of timeslots	No. of rooms
Sem 2 s2014/2015	2248	112	35	18
Sem 1 s2015/2016	2371	125	35	18

4 Mathematical Model of Problem Instance

In this section, the formal model of UMSLIC course timetabling is presented. Let C be the total number of courses, S be the total number of students, T be the total number of timeslots and R be the total number of rooms. The courses offered in UMSLIC can be further divided into four categories which are categorized as follow: main courses; knowledge promotion courses; language courses and co-curricular courses. The four categories of courses are accumulated to get the total number of courses, C.

Let C1 be the total number of main courses, C2 be the total number of knowledge enhancement courses, C3 be the total number of language courses and finally C4 be the total number of co-curriculum courses. Thus the total number of courses, C, will be the sum of all courses, C = C1 + C2 + C3 + C4. The objective is to satisfy all the hard constraints, $\mathrm{H_C} = H_{C_1}, H_{C_2}, H_{C_3}, H_{C_4}$ at the same time minimize the soft constraints, S_{C_1}, S_{C_2} subject to Hc = 0 where the cost function F, can be calculated as follow:

$$F = H_C + S_{C_1} + S_{C_2} \tag{1}$$

The penalty for hard constraint is set with the weight of 100,000 so that the cost function can easily identify if hard constraints exist in case the value is too high, noted that there are 4 hard constraints (λ_1 *until* λ_4) while λ_5 and λ_6 are both soft constraints identified in UMSLIC as stated previously. The penalty for every weight is summarized in Table 2:

Table 2. Penalty assignation of constraint violations.

Weight	Penalty	Description
λ_1 *to* λ_4	100,000	Hard constraints, *Hc*
λ_5	5	Soft constraint, Sc_1
λ_6	2	Soft constraints, Sc_2

The formulation model for both hard constraints and soft constraints are presented below. For hard constraint 1 (H_{C_1}) shown in Eq. (2), student attends more than one lecture at the same time is given by

$$H_{C_1} = \lambda_1 \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} CC_{ij} \tag{2}$$

where C is the total number of courses, λ_1 is the weight, CC_{ij} is the number of students who have at least two classes i and j at the same time.

For hard constraint 2 (H_{C_2}) shown in Eq. (3), the size of the course enrolment should be smaller than or equal to the capacity of the assigned room is given by

$$H_{C_2} = \lambda_2 \sum_{i=1}^{C} \sum_{j=1}^{R} CR_{ij} \tag{3}$$

where C is the number of courses, R is the number of rooms, λ_2 is the weight, CR_{ij} is the number of rooms which the course i larger than the capacity of that room j.

For the hard constraint 3 (H_{C_3}) in Eq. (4), the condition of not more than one course is assigned to a particular timeslot and room at the same time is given by

$$H_{C_3} = \lambda_3 \sum_{i=1}^{R} \sum_{j=1}^{T} RT_{ij} \tag{4}$$

where R is the number of rooms, T is the number of timeslots λ_3 is the weight, RT_{ij} is the number of courses which have different lectures, room i and timeslot j at the same time and classroom.

For the hard constraint 4 (H_{C_4}) shown in Eq. (5), the type of course should be assigned into a specific timeslot is given by

$$H_{C_4} = \lambda_4 \left(\sum_{i=1}^{c_1} \sum_{t=1}^{T} M_{it} + \sum_{j=1}^{c_2} \sum_{t=1}^{T} P_{jt} + \sum_{k=1}^{c_3} \sum_{t=1}^{T} L_{kt} + \sum_{l=1}^{c_4} \sum_{t=1}^{T} C_{lt} \right) \tag{5}$$

where c_1, c_2, c_3, c_4 are the total number of main courses, knowledge enhancement courses, language courses, and co-curricular courses respectively, λ_4 is the weight, T is the number of timeslot and M_{it}, P_{jt}, L_{kt} and C_{lt} are violation of the particular course i, j, k, l to be assigned in timeslot t.

Meanwhile, in terms of soft constraint 1 (S_{C_1}) shown in Eq. (6), student has more than two consecutive courses is given by

$$S_{C_1} = \lambda_5 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij} \tag{6}$$

where n is the number of class activities, λ_5 is the weight, C_{ij} is the consecutive classes attend by student of course i and j.

For soft constraint 2 (S_{C_2}) shown in Eq. (7), the number of student in a particular course, the course should be assigned into a room close to or equal to its size.

$$S_{C_2} = \lambda_6 \sum_{i=1}^{n} \sum_{j=1}^{n} C_{i-j} \tag{7}$$

where n is the number of activities, λ_6 is the weight, C_{i-j} is the difference between the size of room i and course enrolment of course j.

5 Genetic Algorithm

The Genetic Algorithm is a parallel and evolutionary search algorithm based on Darwinian evolutionary theory [6]. It is used to perform search on large, non-linear solution space whereby the knowledge is tough to be encoded and does not require any gradient of information which can evolve from one population to a new population across generation [17]. Therefore, those characteristics make GA a well-suited meta-heuristic search methodology for UCTP. The GA process is called genetic operators and the process starts from a population of random individuals produced in the initial solution. In every generation, the fitness of the population is evaluated. The representation of GA in course timetabling is as follow:

- An individual chromosome represents the solution of the timeslot whereby each index of the timeslot contains the genes
- The genes represent a list of courses, C_j, selected at particular timeslots, T_i, as shown in the Fig. 1.

Timeslots	T(i)	T(i+1)	…	T(n)
Courses	C(j)	C(j+1)	…	C(m)

Fig. 1. Chromosome representation

The application of genetic operators consists of several important operators:

1. Initialization – a population of chromosomes are randomly generated according to the size of population. In this case, the size of population is set to 50 in order to produce enough chromosomes in the pool for each generation.
2. Selection – guides the algorithm to produce optimal solution by selecting the chromosomes with the best fitness. The chromosomes will evolve themselves by

performing tournament selection method which selects K individual from the population (K-tournament). The best two fitness chromosomes in this selection will be used to perform crossover to form an offspring [3].

3. Crossover – one-point crossover is implemented which performs single point recombination between two chromosomes selected from the selection operator [15]. The crossover point is selected at random between the chromosome genes ranging from 1 until $T_n - 1$. The gene before the crossover point will be inherited to the offspring from the first chromosome while the gene after the crossover point will be inherited into the offspring from the second chromosome.
4. Mutation – mutation allow a large search space to be explored in order to prevent the solution from being stuck in local optima due to convergence towards the fittest chromosome [15]. In this research, a random mutation is used which iterates 50 times and the mutation applies only on the offspring produced from crossover operator with small probability. Hence, the mutation rate for this research is set with minimum value of 0.025 or 2.5%.

There are two termination conditions for this algorithm conducted in this research. Those termination conditions allow the final solution to be produced based on: (1) the maximum number of generations and (2) the total time taken to perform the operators.

In order to conduct the experiment, parameter setting of the algorithm needs to be determined. Table 3 summarized the parameter setting of GA conducted in this research which has been discussed previously in the genetic operators.

Table 3. Parameter setting of GA.

Attributes	Parameter
Initialization	Constraint Programming (CP)
Population size	50
Generations	100,000
Selection type	K-tournament selection
Crossover type	One-point crossover
Mutation method	Random mutation
Mutation rate	2.5%
Maximum time allow	300 s (5 min)

In the initialization operator, individual chromosomes are generated by using CP algorithm, which produce a feasible solution and represent as single chromosome. This algorithm has been previously done in the research which only search for feasible solution without taking the consideration of soft constraints violation. The implementation of this algorithm is refers to hybrid constructive heuristics [10]. Hence, 50 chromosomes are generated this way to form a population and each of the chromosomes fitness is evaluated by calculating the cost functions.

In the selection operator of GA, K-tournament is used in order to select best fitness chromosomes to perform crossover and produce offspring. 5 random chromosomes are selected with the best 2 fitness will be chosen to perform crossover in the K-tournament selection.

In crossover operator, one-point crossover is implemented between two parent chromosomes selected in K-tournament. The algorithm checks whether it is feasible to perform crossover in the gene before proceed with the next gene in the chromosome. If it is feasible, the crossover for that gene is successful or else the gene will remain and proceed with the next gene. Hence, prevention function is implemented in order to avoid producing damaged chromosomes before crossover [7].

In mutation operator, random mutation applied in this research perform neighborhood search which involves two types of movements which are:

- Simple Search Neighborhood (SSN) – movement selects a course at random which have been scheduled and assign into an empty time slot and room.
- Swap Search Neighborhood (SWN) – movement randomly selects two courses which have been scheduled and swap their respective time slot and room with one another.

The process of genetic operators is illustrated in Fig. 2. The experimental design follows the parameter setting based on Table 3. The proposed investigation is tested on two datasets obtained from ASD of UMSLIC and 50 iterations are performed for each datasets used in this work. These datasets include the Semester 2 session 2014/2015 dataset and the Semester 1 session 2015/2016 dataset. Two types of experiments are conducted whereby the first experiment tests on the feasibility of GA to produce feasible

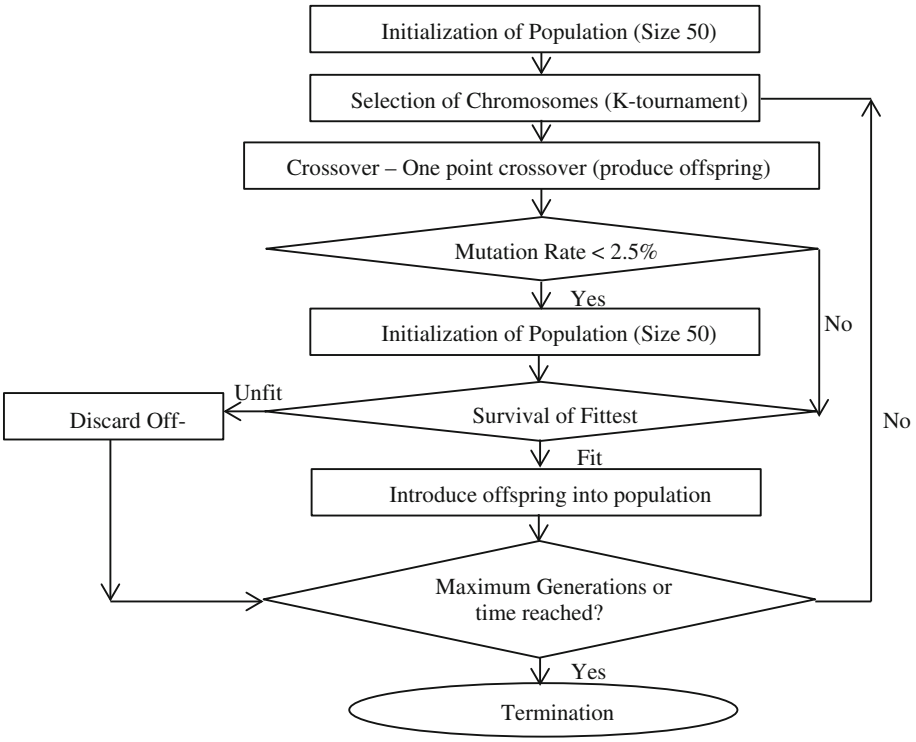


Fig. 2. Process of GA

and good quality of solution. The second experiment conducted to compare the performance between other metaheuristics search which have been done in the past researches [12, 13]. Those algorithms are GD and SA [4]. The experiments are further investigated by performing ANOVA test.

6 Experimental Results

In the first experiment which is to investigate its performance for both semesters and perform analysis based on the quality of the solution produced. Table 4 shows the results of GA for both semesters. The average cost function is the average penalty cost of both hard constraints and soft constraints in this domain. The Max Cost indicates the worst solution produced while the Min Cost indicates the best solution produced among the 50 runs in the experiment. The standard deviation, SD shows the consistency of the cost function produced while the percentage improvement indicates the percentage of improvement from initial solution. The average generations indicates the number of generations have been performed in the cycle of genetic operators. The results obtained can be considered as premature as the ANOVA test will be conducted in the second experiment.

In this experimental results, clearly showed that GA meets the second termination condition which have been discuss previously whereby the algorithm stops when it reaches 300 s instead of reaching the maximum number of generations allowed. The algorithm is able to perform of an average thirty-two thousands generations within five minutes and this shows the importance of data structure in the algorithm. In terms of performance, the algorithm manages to improve the quality of average more than 35% while there is not much difference in both semesters.

Table 4. Experimental result of GA.

Experiments	Semester 2 session 2014/2015	Semester 1 session 2015/2016
Average cost	23847.2	24682.02
Max cost	27044	29563
Min cost	20956	21709
Average time, s	300.05	300.06
Std deviation, sd	1487.537	1403.478
Improvement, %	37.67	35.33
Average generations	33747.9	31915.14

The second experiment compares the performance between GA, GD and SA as shown in Table 5. The results for SA and GD are obtained from the experiment in previous researches. The overall results proved GA performs much better than other meta-heuristics. Although the improvement done by SA or GD is less than of GA, the results showed significant improvement for both semesters with GA manage to improve 35% while both SA and GD only manage to improve 25%.

Table 5. Experimental results of meta-heuristics.

Experiment	Sem 2 S2014/2015			Sem 1 S2015/2016		
Algorithm	GA	GD	SA	GA	GD	SA
Avg cost	23847	29081	30139	24682	29400	30475
Max cost	27044	30745	32991	29563	31003	32872
Min cost	20956	27159	28635	21709	28050	28526
Avg time	300.05	301.47	301.37	300.06	301.01	301.21
SD	1487.5	806.2	820.2	1403.5	603.8	767.8
Improvement	37.67	25.87	23.62	35.33	25.44	22.41

For further statistical analysis ANOVA test is performed on the same experiment in order to verify the significant difference between the qualities of solutions between those meta-heuristics as shown in Table 6. The null hypothesis, H_0 states the qualities of solutions improved by both algorithms are equal while the alternative hypothesis, H_A states otherwise. The p-value for both of the datasets is very small which is smaller than the alpha value of 0.05 and this makes it more likely to reject the null hypothesis. Hence, it means that there is significant difference between the performances of each meta-heuristics search.

Table 6. ANOVA test in meta-heuristics.

Datasets	Variation source	F-value	p-value	F crit
Sem 2 session 2014/2015	Between groups	497.788680	3.49E–66	3.05762
Sem 1 session 2015/2016	Between groups	465.488436	2.5179E–64	3.05762

7 Conclusion

The quality of solutions produced by GA managed to outperform other algorithms in this course timetabling domain. There are improvements in each semester to enhance the quality of initial solutions, which show much better improvement than the other algorithms. As there are significant differences between the performances of those meta-heuristics, this indicates that not all algorithms can produce good solutions in a specific domain [16]. Hence, there is no guarantee the proposed algorithm will perform well in other scheduling domains.

For future research, a multi-agent framework is proposed whereby the agent incorporates the GA which may communicates with other agents in order to search for improving solutions. This may involves several important factors which need to be study such as the communication and coordination between agents. Aside from that, a negotiation protocol is required to be establishes when the agent communicate with one another. Potentially, the implementation of GA and other meta-heuristics into the multi-agent system can increase the quality of solution due to its efficiency and reliability of the system [14].

References

1. Abdullah, S., Shaker, K., McCollum, B., McMullan, P.: Construction of course timetables based on great deluge and tabu search. In: Proceedings of MIC: VIII Metaheuristic International Conference, pp. 13–16 (2009)
2. Abdullah, S., Turabieh, H.: Generating university course timetable using genetic algorithms and local search. In: Third International Conference on Convergence and Hybrid Information Technology, ICCIT 2008, vol. 1, pp. 254–260 (2008)
3. Abdullah, S., Turabieh, H., McCollum, B., Burke, E.K.: An investigation of a genetic algorithm and sequential local search approach for curriculum-based course timetabling problems. In: Proceedings of the Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2009), Dublin, pp. 727–731 (2009)
4. Aycan, E., Ayav, T.: Solving the course scheduling problem using simulated annealing. In: Advance Computing Conference 2009, IACC 2009, pp. 462–466. IEEE International (2009)
5. Dueck, G.: New optimization heuristics: the great deluge algorithm and the record-to-record travel. J. Comput. Phys. **104**(1), 86–92 (1993)
6. Goldberg, D.: Genetic Algorithms in Search Optimization, and Machine Learning. Addison-Wesley, New York (1989)
7. Jat, S.N., Yang, S.: A guided search non-dominated sorting genetic algorithm for the multi-objective university course timetabling problem. In: European Conference on Evolutionary Computation in Combinatorial Optimization, pp. 1–13. Springer, Heidelberg (2011)
8. Jat, S.N., Yang, S.: A guided search genetic algorithm for the university course timetabling problem (2009)
9. Kahar, M.N.K., Kendall, G.: The examination timetabling problem at Universiti Malaysia Pahang: comparison of a constructive heuristic with an existing software solution. Eur. J. Oper. Res. **207**(2), 557–565 (2010)
10. Landa-Silva, D., Obit, J.H.: Comparing Hybrid Constructive Heuristics for University Course Timetabling (2011)
11. Obit, J.H.: Developing novel meta-heuristic, hyper-heuristic and cooperative search for course timetabling problems, Ph.D. thesis School of Computer Science, University of Nottingham (2010)
12. Obit, J.H., Landa-Silva, D., Ouelhadj, D., Sevaux, M.: Non-linear great deluge with learning mechanism for solving the course timetabling problem. In: 8th Metaheuristics International Conference, MIC (2009)
13. Obit, J.H., Landa-Silva, D., Sevaux, M., Ouelhadj, D.: Non-linear great deluge with reinforcement learning for university course timetabling. In: Metaheuristics–Intelligent Decision Making. Operations Research/Computer Science Interfaces, pp. 1–19. Springer (2011)
14. Oliveira, E., Fischer, K., Stepankova, O.: Multi-agent systems: which research for which applications. Robot. Auton. Syst. **27**, 91–106 (1998)
15. Lin, W.-Y., Lee, W.-Y., Hong, T.-P.: Adapting crossover and mutation rates in genetic algorithms. J. Inf. Sci. Eng. **19**, 889–903 (2003)
16. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (2006)
17. Yang, S., Jat, S.N.: Genetic algorithms with guided and local search strategies for university course timetabling. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **41**(1), 93–106 (2011)

Author Index

R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 465–466, 2018.
https://doi.org/10.1007/978-981-10-8276-4