

## About this publication

<b>Title:</b>	The Workshop Proceedings of the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI Workshop 2016), August 22-26, 2016, Phuket, Thailand
<b>Editor-in-chief:</b>	Thanaruk Theeramunkong (SIIT, TU, Thailand)
<b>Editors:</b>	Thepchai Supnithi (NECTEC, Thailand) Chuleerat Jaruskulchai (Kasetsart University, Thailand) Sanparith Marukatat (NECTEC, Thailand) Mahasak Ketcham (KMUTNB, Thailand) Narit Hnoohom (Mahidol University, Thailand) Patiyuth Pramkeaw (KMUTT, Thailand) Masayuki Numao (Osaka University, Japan) Manabu Okumura (Tokyo Institute of Technology) Merlin Teodosia Suarez (De La Salle University, Philippines) Richard Booth (Cardiff University, UK) Min-Ling Zhang (Southeast University, China)
<b>Workshop Co-Chairs:</b>	Masayuki Numao (Osaka University, Japan) Boonserm Kijsirikul (Chulalongkorn University, Thailand) Sanparith Marukatat (NECTEC, Thailand)
<b>Production Assistants:</b>	Mahasak Ketcham (KMUTNB, Thailand) Narit Hnoohom (Mahidol University, Thailand) Patiyuth Pramkeaw (KMUTT, Thailand) Choermath Hoongakkrapha (SIIT, Thammasat University) Wiwit Suksangaram (Phetchaburi Rajabhat University, Thailand) Worawut Yimyam (Phetchaburi Rajabhat University, Thailand)
<b>Cover Designer:</b>	Wasan Na Chai (NECTEC, Thailand)
<b>Printing Production:</b>	VS 8 Inter Limited Partnership, 2/484 Moo 1, Klong Luang, Pathum Thani, 12120, Thailand
<b>Date Published:</b>	August 2016
<b>Supporters and Sponsors:</b>	Thammasat University (TU, Thailand) Prince of Songkla University (PSU, Thailand) Artificial Intelligence Association of Thailand (AIAT), 22/1, Soi Phutta Bucha 30, Phutta Bucha Rd. Bang mot, Tungkru, Bangkok, 10140, Thailand, Email: office@aiat.in.th, URL: <a href="http://aiat.in.th/">http://aiat.in.th/</a>
<b>Publishers:</b>	Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU), 131 Moo 5 Tiwanont Road, Bangkadi, Muang Pathumthani 12000, Thailand, Email: thanaruk@siit.tu.ac.th Tel. +66-2-501-3505(-20) Fax +66-2-501-3524
<b>Contact:</b>	

© 2016 by AIAT, Artificial Intelligence Association of Thailand (PRICAI 2016)

Printed in Thailand

ISBN 978-616-92700-0-3

## Preface

Established in 1990, The Pacific Rim International Conference on Artificial Intelligence (PRICAI) is a biennial international event which concentrates on AI theories, technologies and their applications in the areas of social and economic importance for countries in the Pacific Rim, and has become a major venue for scholars and researchers in the Pacific Rim region to showcase their work in artificial Intelligence. The aim behind organizing these workshops is to bring together researchers of various interests to present, discuss and explore the state of applying information technology in various aspects of learning. The volume contains the supplementary proceedings of the 14<sup>th</sup> The Pacific Rim International Conference on Artificial Intelligence (PRICAI: <http://aiat.in.th/pricai2016/>) held from August 22<sup>nd</sup> through August 26<sup>th</sup>, 2016 in Phuket, Thailand.

This year, we accepted six workshop proposals with the goal of exploring focused issues across various themes. There are 46 papers were submitted to all workshops. Each proposal in these proceedings was peer-reviewed by international reviewers in their respective areas to ensure the highest quality work. We believe that the workshops provide a valuable venue for researchers to share their work and have the opportunity to collaborate with likeminded individuals. The workshop papers spanning various topics will certainly stimulate more interesting research in respective areas in Pacific Rim countries. We hope that readers will find the ideas and lessons presented in the proceedings relevant to their research.

Finally, we would like to thank The Pacific Rim International Conference on Artificial Intelligence (PRICAI 2016) Executive Committees, Program Co-Chairs for entrusting us with the important task of chairing the workshop program, thus giving us an opportunity to grow through valuable academic learning experiences. We also would like to thanks all workshop Co-Chairs for their tremendous and excellence work.

### Organizers

Masayuki Numao (Osaka University, Japan)

Boonserm Kijisirikul (Chulalongkorn University, Thailand)

Sanparith Marukatat (National Electronics and Computer Technology Center, Thailand)

Thepchai Supnithi (National Electronics and Computer Technology Center, Thailand)

Thanaruk Theeramunkong (SIIT, Thammasat University, Thailand)

## **I3A**

### **International Workshop on Image, Information, and Intelligent Applications**

The adoption of Artificial Intelligent (AI) technology, and in particular of its most challenging components like Image and Information which can constitute the basic building blocks for a variety of applications within the intelligent world. The combination of the image and emerging information technologies such as image retrieval, computer vision, expert systems, social network analysis, and big data analytics lets us transform everyday information into smart knowledge applications.

International Workshop on Image, Information, and Intelligent Applications (I3A) is at our 1st edition to collocate with the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2016), held in Phuket, Thailand, August 22-26, 2016. I3A brings together image, information and intelligent applications from a diverse group of people from industry and academia. As artificial intelligence matures, it is gradually embedded in the environment and goes beyond conventional information to cover a complex and dynamic environment composed of multiple artifacts.

We received 13 submissions from all over the world. After an intense reviewing process with at least three reviewers for all papers from different countries. Seven full papers and two short papers are accepted to present in I3A with 54% acceptance ratio. All accepted papers have illustrated interesting research projects, results and industrial experiences that describe significant advances in image, information and intelligent computing. We hope to provide opportunities for all participants for beneficial discussion and future research resource for you.

We look forward to seeing all of you in Phuket for this workshop.

#### Workshop Organizers

Narit Hnoohom (Mahidol University, Thailand)

Tanasanee Phienthrakul (Mahidol University, Thailand)

Mingmanas Sivaraksa (Mahidol University, Thailand)

Anuchit Jitpattanakul (King Mongkut's University of Technology North Bangkok, Thailand)

Sakorn Mekruksavanich (University of Phayao, Thailand)

Ghita Berrada (Twente University, Netherlands)

Rozlina Mohamed (University Technology Malaysia, Malaysia)

## **AIED**

### **International Workshop on Artificial Intelligence for Educational Applications**

The “Artificial Intelligence for Educational Applications” (AIED) Workshop at the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2016) has organized as a forum for researchers who are interested in research and development of interactive and adaptive learning environments for learners of all ages, across all domains, to exchange several experiences and results from applications of Artificial Intelligent approaches in Education.

This is the first time the AIED 2016 Workshop has been organized. It received ten submissions which are passed the full double-blind refereeing process reviewed by at least three program committee. Three full papers and five short papers are accepted.

We would like to thank all people that assisted to make this Workshop possible. Our grateful thanks to all the members of Program Committee for timely and excellent reviews. Finally, we thank to all AIED 2016 participants and we hope this Workshop will give some valuable ideas and inspiration to make practical applications for education in the near future.

Workshop Organizers  
Thepchai Subnithi (National Electronics and Computer Technology Center, Thailand)  
Rachada Kongkachandra (Thammasat University, Thailand)

## **AI4T**

### **International Workshop on Artificial Intelligence for Tourism**

The tourism industry has become one of the fastest growing industries in the world. Furthermore, the travel and tourism industry has always been open to new technologies.

With the increased interest in the travel and tourism industry in the world, developing new information and communication technologies (ICT) in the travel and tourism industry has become quite important in these days. Since artificial intelligence technologies can be considered one of the key technologies in ICT, they should play a key role in the travel and tourism industry.

Today, AI-based developments in the field are at the forefront. In fact, AI developments and researches have induced much change in this industry. We can expect this innovation to continue: at both the industrial level and the academic level.

This workshop offers a worldwide and unique forum for attendees from academia, industry, government, and other organizations to actively exchange, share, and challenge state-of-the-art researches and industrial case studies on the application of artificial intelligence technologies to travel and tourism.

I hope you will enjoy the workshop. Once again, thank you very much for joining our workshop.

#### Workshop Organizers

Manabu Okumura (Tokyo Institute of Technology, Japan)

Hidetsugu Nanba (Hiroshima City University, Japan)

Kazutaka Shimada (Kyushu Institute of Technology, Japan)

Fumito Masui (Kitami Institute of Technology, Japan)

ThanarukTheeramunkong (SIIT, Thammasat University, Thailand)

## **IWEC**

### **International Workshop on Empathic Computing**

The goal of the International Workshop on Empathic Computing (IWEC) has been to bring together researchers in Asia who wish to solve interesting problems of human-machine interaction that considers social signals and emotions via real-world modalities of body movement and gestures, facial and audio expressions, and non-vocal utterances, as well as wearable devices to measure physiological signals to measure the effects of such interactions.

This year's workshop focuses on the brain signals and the emotion. The workshop is honored to have Associate Professor Kenji Tanaka, M.D., Ph.D. from the Department of Neuropsychiatry of Keio University School of Medicine. This is the first time a brain expert will be delivering a talk at IWEC.

The workshop received a number of papers from all over the world. Each paper was reviewed by international experts in these fields, and only 50% of the submissions were accepted as full papers. The papers accepted to the workshop range from processing physiological signals, to the effects of emotion during a student's learning episode, as well as understanding emotional laughter expressions. The participants will learn a breadth of modalities, how these are processed, and how emotions and its expressions play a large part to make user experiences worthwhile.

The workshop organizers would like to thank all the advisers and the members of the program committee, including all the authors who submitted papers. In addition, we would like to thank the PRICAI Organizers for their support to IWEC-16. We look forward to a successful workshop.

Workshop Organizers  
Merlin Teodosia Suarez (De La Salle University, Philippines)  
The Duy Bui (Vietnam National University - Hanoi, Vietnam)  
Ma. Mercedes Rodrigo (Ateneo de Manila University, Philippines)  
Masayuki Numao (Osaka University, Japan)

## RSAI

### Workshop on Research Student Symposium on the Artificial Intelligence and Applications

This is an exciting time to be a part of an Artificial Intelligence Research Student Symposium. AI technologies and applications have truly entered our everyday lives, with AI systems in use throughout society. Against this backdrop of AI's remarkable success, the First International Research Student Symposium on the Artificial Intelligence and Application (RSAI-2016), to be held in Phuket, Thailand between 22 and 26 August 2016, is the first time the flagship international AI conference has been held in Kingdom of Thailand, and provide the opportunity for Master Students and Ph.D. candidates and researchers to share and discuss on research work and idea.

These proceedings collect some of the most exciting research taking place in AI today and offer a window into the future. The theme of the workshop this year is "Artificial Intelligence and Application." Being held in Phuket, the Pearl of Andaman Ocean, the RSAI workshop will feature invited talks, performances and a technical track dedicated to the exploration of AI's growing role in Application and research update, both in enriching and producing AI work and injecting AI into application to make it an elegant and more accessible scientific discipline.

All accepted papers are included in these proceedings and are invited to present in workshop sessions and posters. Authors of all papers are invited to give oral presentations, where a distinction is made between long talks and short talks based on the papers' quality, clarity, and potential relevance to a wider audience. The Program Committee did a thorough and highly professional job in reviewing all papers submitted to the conference. Every paper received at least two reviews, which were provided by members of the program committee (PC). The review process for each paper was overseen by at least one senior program committee (SPC) member, who monitored reviews and initiated discussion before and after the author feedback period. Each paper was also managed by the workshop chairs (WC), who engaged in discussions and gave recommendations on the final decision based on the PC and SPC input. During the review process, author feedback was taken into account for the final discussion. When necessary, additional reviews were obtained. Finally, Workshop Chairs, PCs, SPCs and the Reviewers worked closely in making the final decisions. The end result is a RSAI-2016 program of outstanding quality. We wish to express our deep appreciation to the Workshop Chairs and Program Committee Members for their outstanding and very professional organization of the review process.

We also wish to express our sincere thanks to all reviewers for their valuable dedication. Our deep gratitude extends to the organizers of many other workshop programs of the PRICAI-2016 conference. We are indebted to their great effort and professionalism. The conference organizing committee provided the much-needed support for the review process. Finally, we wish to thank all authors of submitted technical papers for contributing to this great RSAI Workshop Symposium. With your high quality work and devotion, RSAI will continue its tradition of excellence and leadership in advancing the Artificial Intelligence Application for the upcoming Symposia in future.

#### Workshop Organizers

Vincent Shin-Mu Tseng (National Cheng Kung University, Tainan, Taiwan)

Mahasak Ketcham (King Mongkut's University of Technology North Bangkok, Thailand)

Thaweesak Yingthawornsuk (King Mongkut's University of Technology Thonburi, Thailand)

# Organizing Committee

## Honorary Co-Chairs

- Wai Kiang (Albert) Yeap (*AUT University, New Zealand*)
- Abdul Sattar (*Institute for Integrated and Intelligent Systems, Griffith University, Australia*)
- Hiroshi Motoda (*Osaka University, Japan*)
- Vilas Wuwongse (*Mahidol University, Thailand*)
- Somnuk Tangtermsirikul (*SIIT, Thammasat University, Thailand*)
- Sarun Sumriddetchkajorn (*NECTEC, Thailand*)
- Pun Thongchunum (*Prince of Songkla University, Thailand*)

## General Co-Chairs

- Dickson Lukose (*MIMOS Berhad*)
- Thanaruk Theeramunkong (*SIIT, Thammasat University, Thailand*)

## Technical Program Co-Chairs

- Richard Booth (*Cardiff University, United Kingdom*)
- Min-Ling Zhang (*Southeast University, China*)

## Workshop Co-Chairs

- Masayuki Numao (*Osaka University, Japan*)
- Boonserm Kijsirikul (*Chulalongkorn University, Thailand*)
- Sanparith Marukatat (*NECTEC, Thailand*)

## Workshop Organizers

### PeHealth

- Chuleerat Jaruskulchai (*KU, Kasetsart University, Thailand*)
- Ornuma Thesprasith (*KU, Kasetsart University, Thailand*)
- Rey-Long Liu (*Tzu Chi University, Hualien, Taiwan, R.O.C.*)

### I3A

- Narit Hnoohom (*Mahidol University, Thailand*)
- Tanasanee Phienthrakul (*Mahidol University, Thailand*)
- Mingmanas Sivaraksa (*Mahidol University, Thailand*)
- Anuchit Jitpattanakul (*King Mongkut's University of Technology North Bangkok, Thailand*)
- Sakorn Mekruksavanich (*University of Phayao, Thailand*)
- Ghita Berrada (*Twente University, Netherlands*)
- Remi Barillec (*Aston University, United Kingdom*)
- Rozlina Mohamed (*University Technology Malaysia, Malaysia*)

### AIED

- Thepchai Supnithi (*NECTEC, Thailand*)
- Rachada Kongkrachandra (*Thammasat University, Thailand*)
- Tsukasa Hirashima (*Hiroshima University, Japan*)

### AI4T

- Manabu Okumura (*Tokyo Institute of Technology, Japan*)
- Hidetsugu Nanba (*Hiroshima City University, Japan*)
- Kazutaka Shimada (*Kyushu Institute of Technology, Japan*)
- Fumito Masui (*Kitami Institute of Technology, Japan*)
- Thanaruk Theeramunkong (*SIIT, Thammasat University, Thailand*)

**IWEC**

- Merlin Teodosia Suarez (*De La Salle University, Philippines*)
- The Duy Bui (*Vietnam National University - Hanoi, Vietnam*)
- Ma. Mercedes Rodrigo (*Ateneo de Manila University, Philippines*)
- Masayuki Numao (*Osaka University, Japan*)

**RSAI**

- Vincent Shin-Mu Tseng (*National Cheng Kung University, Tainan, Taiwan*)
- Mahasak Ketcham (*King Mongkut's University of Technology North Bangkok, Thailand*)
- Thaweesak Yingthawornsuk (*King Mongkut's University of Technology Thonburi, Thailand*)

**Special Session and Doctoral Symposium Co-Chairs**

- Abdul Sattar (*Institute for Integrated and Intelligent Systems, Griffith University, Australia*)
- Chuleerat Jaruskulchai (*Kasetsart University, Thailand*)
- Rachada Kongkachandra (*Thammasat University, Thailand*)
- Mahasak Ketcham (*King Mongkut's University of Technology North Bangkok, Thailand*)
- Narit Hnoohom (*Mahidol University, Thailand*)
- Pokpong Songmuang (*Thammasat University, Thailand*)
- Patiyuth Pramkeaw (*KMUTT, Thailand*)

**Financial Co-Chairs**

- Chutima Beokhaimook (*Rangsit University, Thailand*)
- Nongnuch Ketui (*RMUTL, Nan, Thailand*)
- Choermath Hongakkaranphan (*SIIT, Thammasat University, Thailand*)

**Local Organizing Co-Chairs**

- Rattana Wetprasit (*Prince of Songkla University, Thailand*)
- Virach Sortlertlamvanich (*SIT, TU., Thailand*)
- Thepchai Supnithi (*NECTEC, Thailand*)
- Nattapong Tongtep (*Prince of Songkla University, Thailand*)

**Secretary Generals**

- Thatsanee Chareonporn (*Burapha University, Thailand*)
- Choermath Hongakkaranphan (*SIIT, Thammasat University, Thailand*)
- Kiyota Hashimoto (*Prince of Songkla University, Thailand*)

**PRICAI Steering Committee**

- Tru Hoang Cao (*Ho Chi Minh City University of Technology, Vietnam*)
- Aditya Ghose (*University of Wollongong, Australia*)
- Byeong-Ho Kang (*University of Tasmania, Australia*)
- Dickson Lukose (*MIMOS Berhad, Malaysia*)
- Hideyuki Nakashima (*Future University Hakodate, Japan*)
- Seong-Bae Park (*Kyungpook National University, Korea*)
- Duc Nghia Pham (*MIMOS Berhad, Malaysia*)
- Abdul Sattar (*Griffith University, Australia*)
- Toby Walsh (*NICTA, Australia*)
- Chengqi Zhang (*University of Technology, Sydney, Australia*)
- Zhi-Hua Zhou (*Nanjing University, China*)

**Webmasters**

- Kobkrit Viriyayudhakorn (*SIIT, Thammasat University, Thailand*)
- Thanasan Tanhermhong (*SIIT, Thammasat University, Thailand*)
- Wirat Chinnan (*SIIT, Thammasat University, Thailand*)

## Reviewers

- Aiba Eriko
- Amphawan Komate
- Azcarraga Judith
- Baker Ryan
- Bermudez Thomas
- Bevacqua Elisabetta
- Bui Duy The
- Cabredo Rafael
- Choosumrong Sittichai
- Chumuang Narumol
- Cu Jocelynn
- Dekok Iwan
- D'Mello Sidney Beck Joseph
- Echizen Isao
- Ganokratanaa Thittaporn
- Hagad Lorenzo Juan
- Herzallah Randa
- Hirashima Tsukasa
- Hnoohom Narit
- Inoue Masashi
- Inventado Salvador Paul
- Jaruskulchai Chuleerat
- Jeefoo Phaisarn
- Jensuttiwetchakul Tanapon
- Jitpattanakul Anuchit
- Kashihara Akihiro
- Ketcham Mahasak
- Kiewkanya Matinee
- Kongkachandra Rachada
- Liu Rey-Long
- Lutfi Lebai Syaheerah
- Marcos Nelson
- Marukatat Rangsipan
- Masui Fumito
- Matam Rajeswari
- Nanba Hidetsugu
- Niewiadomska Radoslaw
- Numao Masayuki
- Okumura Manabu
- Ono Isao
- Otani Noriko
- Otani OchsNoriko Magalie
- Phokharatkul Pisit
- Plaimas Kitiporn
- Ponsawat Jiradej
- Pramkeaw Patiyuth
- Pravesjit Sakkayaphop
- Reidsma Dennis
- Rimcharoen Sunisa
- Robert Pummakarnchana Ornprapa
- Rodrigo Mercedes Ma.
- Sangkavichitr Chalermsub
- Sawatnatee Amnat
- Scott Kirk
- Shimada Kazutaka
- Sison Raymund Azcarraga Arnulfo
- Songmuang Pokpong
- Srisawat Anantaporn
- Suarez Teodosia Merlin
- Suksangaram Wiwit
- Supnithi Thepchai
- Suratanee Apichat
- Thammasan Nattapong
- Theeramunkong Thanaruk
- Thesprasith Ornuma
- Tran Hong Ngoc
- Truong Khiet
- Urbain Jerome
- Wanvarie Dittaya
- Wohlgemuth Sven
- Yamagata MasashiInoue
- Yingthawornsuk Thaweesak

## Table of Contents

<b>Pacific Rim International Workshop on eHealth Mining</b>		<b>Page</b>
PRICAI-W-0026	<b>Exploring the Distributional Semantic Relation for ADR and Therapeutic Indication Identification in EMR</b>  Siriwon Taewijit, Thanaruk Theeramunkong	1
<b>International Workshop on Image, Information, and Intelligent Applications</b>		
PRICAI-W-0020	<b>Desktop Tower Defense is NP-Hard</b>  Vasin Suttichaya	14
PRICAI-W-0021	<b>Learning Latent Word Representations for Enhanced Short Text Classification</b>  Luepol Pipanmaekaporn, Suwatchai Kamolsantiroj	21
PRICAI-W-0022	<b>Variant Annotation and Clinical interpretation software for Cancer (VARCIN): Report generating software for targeted therapy method</b>  Mingmanas Sivaraksa, Pitinat Asawasutsakorn, Benjamard Meeboon, natini Jinawath	33
PRICAI-W-0023	<b>MEDICINE RECOGNITION USING INTRINSIC GEOMETRIC PROPERTY FROM PILL IMAGE</b>  M Ashraful Amin, Md. Zakir Hossan, Tanjina Piash Proma	46
PRICAI-W-0025	<b>A Regression-based SVD Parallelization using Overlapping Folds for Textual Data</b>  Uraiwan Buatoom, Thanaruk Theeramunkong, Ware Kongprawechnon	52
PRICAI-W-0031	<b>Virtual Reality System with Smartphone Application for Height Exposure</b>  Suppanut Nateeraitaiwa, Narit Hnoohom	64
PRICAI-W-0040	<b>Classification of diabetic retinopathy stages using image segmentation and an artificial neural network</b>  Narit Hnoohom, Ratikanlaya Tanthuwapathom	78

## International Workshop on Artificial Intelligence for Educational Applications

Page

PRICAI-W-0014	<b>Examination timetabling using prey predator algorithm</b> Surafel Tilahun, Jean Medard Ngnotchouye	90
PRICAI-W-0016	<b>K-Mean Algorithm for Finding Students' Proficiency with a Framework's Item Examination</b> Nongnuch Ketui, Prasert Luekhong	101
PRICAI-W-0030	<b>Building a Semantic Ontology for Virtual Peers in Narrative-Based Environments</b> Ethel Ong	108
PRICAI-W-0033	<b>Automatic Question Generation on SQL Language Using Template-Based Method</b> Jittima Janphat, Orawan Chaowalit	120
PRICAI-W-0035	<b>TSCS Monitor: Generation of Time Series Cross Section Tables from Moodle Logs for Tracking In-Class Page Views Using Excel Macros</b> Konomu Dobashi	133
PRICAI-W-0045	<b>Development of Salary Prediction System to Improve Student Motivation using Data Mining Technique</b> Pornthep Khongchai, Pokpong Songmuang	139
PRICAI-W-0054	<b>Contents Organization Support for Logical Presentation Flow</b> Tomoko Kojiri	145
PRICAI-W-0057	<b>A Framework to Generate Carrier Path Using Semantic Similarity of Competencies in Job Position</b> Wasan NA Chai, Taneth Ruangrajjitpakorn, Marut Buranarach, Thepchai Supnithi	157

## International Workshop on Artificial Intelligence for Tourism

PRICAI-W-0029	<b>Two stage travel salesman model of world tourism</b> Surafel Tilahun, Jean Medard Ngnotchouye	166
PRICAI-W-0037	<b>Extracting and Characterizing Functional Communities in Spatial Networks</b> Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda, Kazuhiro Kazama	182
PRICAI-W-0055	<b>Travellers' Behaviour Analysis Based on Automatically Identified Attributes from Travel Blog Entries</b> Hidetsugu Nanba	194
PRICAI-W-0059	<b>Inferring Tourist Behavior and Purposes of a Twitter User</b> Yuya Nozawa	206

International Workshop on Empathic Computing		Page
PRICAI-W-0034	<b>Application of Annotation Smoothing for Subject-independent Emotion Recognition based on Electroencephalogram</b> Nattapong Thammasan, Ken-ichi Fukui, Masayuki Numao	218
PRICAI-W-0043	<b>Modeling Negative Affect Detector of Novice Programming Students using Keyboard Dynamics and Mouse Behavior</b> Larry Vea, Ma. Mercedes Rodrigo	231
PRICAI-W-0047	<b>Multimodal Latent Feature Learning for Psycho-Physiological Stress Modelling and Detection</b> Juan Lorenzo Hagad, Ken-ichi Fukui, Masayuki Numao	243
PRICAI-W-0056	<b>Affective Laughter Expressions from Body Movements</b> Jocelynn Cu, Ma. Beatrice Luz, McAnjelo Nocum, Timothy Jasper Purganan, Wing San Wong	253
PRICAI-W-0060	<b>Computational Model for Affect Detection in Learning</b> Najlaa Sadiq Mokhtar, Syaheerah Lebai Lutfi	261

#### Symposium on the Artificial Intelligence and Applications

PRICAI-W-0007	<b>Verifying Properties of Multi-agent Systems via Bounded Model Checking</b> Agnieszka Zbrzezny	272
PRICAI-W-0011	<b>MOEPSO for Multi-objective Optimization</b> Ittikon Thammachantuek, Mahasak Ketcham	278
PRICAI-W-0013	<b>Enhancement of Palm-Leaf Manuscript for Segmentation</b> Siriya Phattarachairawee, Montean Rattanasiriwongwut, Mahasak Ketcham	286
PRICAI-W-0015	<b>Comparison of Edge Detection Algorithms for Coastline Detection in Satellite Imageries</b> Chutiwan Boonarchatong Sucha Smanchat, Mahasak Ketcham, Nawaporn Wisitpongphan	298
PRICAI-W-0018	<b>Real-time Snoring Sound Detecting U Shape Pillow System using Data Analysis Algorithm</b> Patiyuth Pramkeaw, Penpichaya Lertritchai, Nipaporn Klangsakulpoontawee	304
PRICAI-W-0019	<b>A multi-objective adaptive Invasive Weed Optimization intelligence approach for solving DNA sequence design</b> Qiang Zhang, Gaijing Yang, Changjun Zhou, Bin Wang	315
PRICAI-W-0038	<b>Fatigue Classification of Military Mission by EEG signals via Artificial Neural Network (ANN)</b> Worawut Yimyam, Mahasak Ketcham	327
PRICAI-W-0041	<b>Arrival Time Prediction and Train Tracking Analysis</b> Somkiat Kosolsombat, Wasit Limprasert	337
PRICAI-W-0046	<b>The Limb Leads ECG Signal Analysis in Myocardial Infarction Patients</b> Anchana Muankid, Mahasak Ketcham	346
PRICAI-W-0058	<b>Estimating PSD Characteristics of ECG in Comparison between Normal and Supraventricular Subjects</b> Thaweesak Yingthawornsuk, Siriphan Phetnuam	352

PRICAI-W-0049	<b>Evolving Public Opinion Mining Methods on Decision Support System in Thai E-Government</b> Jeerana Noymanee, Wimol San-Um, Thanaruk Theeramunkong	360
<b>Author Index</b>		368

## Exploring the Distributional Semantic Relation for ADR and Therapeutic Indication Identification in EMR

Siriwon Taewijit<sup>12</sup> and Thanaruk Theeramunkong<sup>1</sup>

<sup>1</sup> Sirindhorn International Institute of Technology, Thammasat University,  
Pathum Thani, Thailand

<sup>2</sup> Japan Advanced Institute of Science and Technology, Ishikawa, Japan  
siriwont@jaist.ac.jp, thanaruk@sjit.tu.ac.th

**Abstract.** Extraction of relations and their semantic relations from a clinical text is significant to comprehend the actionable harmful and beneficial events between two clinical entities. Particularly to implement drug safety surveillance, two simplest but most important semantic relations are *adverse drug reaction* and *therapeutic indication*. In this paper, a method to identify such semantic relations is proposed. A large scale of nearly 1.6 million sentences over 50,998 discharge summary from Electronic Medical Records were preliminary explored. Our approach provided the three main contributions; (i) Electronic Medical Records characteristic exploration; (ii) OpenIE examination for clinical text mining; (iii) automatic semantic relation identification. In this paper, the two complementary information from public knowledge base were introduced as a comparative advantage over expert annotation. Then the set of relation patterns were qualified with  $0.05$  significant level. The experimental results show that our method can identify the common *adverse drug reaction* and *therapeutic indication* with the high lift value. Additionally, a novel *adverse drug reaction* and alternative drug for a specific symptom therapy are reported to support the comprehensive further drug safety surveillance. The paper clearly illustrates that our method is not only effortless from expert annotation, automatic pattern-specific semantic relation extraction, but also effective for semantic relation identification.

**Keywords:** adverse drug reaction, electronic medical records, semantic relation extraction, therapeutic indication, text mining

### 1 Introduction

Exponential growth of Electronic Medical Records (EMR), with replacement of paper-based records enables us to utilize information in effective and efficient ways. Recently several works have been focused on semantic relation extraction

from EMR to conspicuously benefit to *drug-symptom* network [1, 2] for comprehensive drug safety surveillance [3–5]. Basically, there are two simply complementary semantic (meaning) relations existing along with the network; *drug-induce-symptom* for the harmful perspective and *drug-treat-symptom* for the beneficial one. The former is namely as adverse drug reaction (ADR) and the latter refers to therapeutic indication. To deal with EMR, text preprocessing (e.g. sentence boundary detection (SBD), name entity recognition (NER), etc.) and the machine-readable form for text representation are the mandatory tasks which highlight the challenges.

On the one hand, in order to explore associations between arbitrary two entities in huge EMR contexts, the relation extraction is a fundamental process. There are two main paradigms for relation extraction [6]. Firstly, the traditional relation extraction primarily recognises a relation-specific between two or more entities in text. The process is fallen into drawback of labouring tasks due to hand-labeled training and infeasible less efforts when shifting to a new relation. The cost and time-consuming of this manual tasks are linearly expensive along with the number of relations. Another paradigm, open information extraction (OpenIE) [7], is rather new and feasible for a large scale corpora due to relation independent extraction. Hence, a vast number of diversity of relational tuples between arbitrary two clinical entities are extracted simultaneously. Putting semantic into these relational tuples can further enhance the comprehensive regarding the actionable harmful and beneficial events. Practically, the semantic relations [8–10] are not always expressed with explicit words such as *treat* or *cause* etc. Mostly, they are also frequently expressed with combined and complex expressions [11] and need interpretation. In our research contexts, the semantic relation identification in narrative text from EMR corresponds to automatically annotate the relation type (e.g. *ADR*, *indication*). Considering on the one of our discovered relation, “*be hold in*” and “*be appropriately controlled with*”, they can be annotated into *ADR* and *indication* semantic relations respectively.

In this work, we provided three contributions; (i) to explore the characteristic of discharge summary from EMR; (ii) to examine the powerful of OpenIE on narrative text; (iii) to automatically identify semantic relation. We initially analyzed a huge number of nearly 1.6 million sentences over 50,998 discharge summary from EMR. The two rich contexts from the brief hospital course (BHC) and the history of present illness (HPI) sections in discharge summary were selected for investigation. Then the rather simple but efficient method by exploring distributional semantic relation was employed in order to capture the key pattern-specific semantic relation for *ADR* and *indication* identification. To avoid suffering from hand-labeled training, the comparatively existing knowledge base from SIDER and DrugBank were enriched as temporary label. Our hypothesis is that the pattern-specific semantic relation is significantly related to its label. The conditional entropy was computed to examine the uncertainty of semantic relation given a pattern. Lately, the key pattern-specific semantic relation was qualified by hypothesis testing of contingency tables with 0.05 significant level.

We organise the remaining of this paper into five sections: the related work is given in the next section. The text mining for data preprocessing and distributional semantic relation exploration are described in Section 3. The experimental results and conclusion are summarised into Section 4 and 5 respectively.

## 2 Related Work

Detection of underlying relations to further comprehend *drug-symptom* network, multidisciplinary approaches are extensive study. Traditional approach, co-occurrence statistic is favour for decade [12–14] due to simplest and less effort. The method relies on the co-occur of a pair of drug and symptom entities in the specified boundary such as with in a window size, sentence, abstract, paragraph, or document. Unfortunately, this method is loosely to capture the true semantic relation such as *drug-induce-symptom* or *drug-treat-symptom*.

The considering of distinct between two complementary semantic relations of *ADR* and *indication* has been the significant of comprehensive *drug-symptom* network. Wang et al. [15] incorporated omic data (i.e. chemical structures and protein targets) and the two complementary semantic relations. The *ADR* and *indication* semantic relations are interchangeable as a feature representation for themselves predictive model, for example, *ADR* with omic data as feature representation to predict *indication*, and *vice versa*. Then two interdependent models were constructed to estimate the probability of *ADR* and *indication* by logistic regression. Recently, the preliminary study of ADR and therapeutic indication on social media, Segura-Bedmar et al. [16] employed co-occurrence of *drug-symptom* pairs by varying  $n$  window size from 10 to 50 on 400 Spanish user comments. Hence, the semantic relation was assigned based on the appearance of *drug-symptom* pairs according to its sections describing (ADR or indication) which was derived from drug package leaflets.

Lately, two works but complementary are reported by Xu et al. The former [17] aimed to derive new drug therapeutic indication for drug repurposing by explore *drug-symptom* pairs from 20 million MEDLINE abstracts. To learn the drug indication pattern, *drug-symptom* pairs were extracted from Clinicaltrials.gov, then, the contexts between all extracted pairs were examined for *indication* pattern. Latter work [18] employed dependency parse tree to retrieve ADR-specific syntactic patterns for ADR detection. The large scale of 119 million MEDLINE sentences were investigated. Different from previous work, the SIDER knowledge base was deployed to derive *drug-symptom* pairs regarding ADR. Finally, all extracted patterns were ranked based on their associated pattern scores and co-occurrence frequencies. Then, the manual selection process was manipulated in order to remove irrelevance patterns and used to retrieve unknown *drug-symptom* pairs. The example patterns regarded to *ADR* and *indication* semantic relations are “*induced*”, “*associated*”, “*related*”, etc. and “*in*”, “*for the treatment of*”, “*in the management of*”, etc. respectively.

### 3 Materials and Methods

The entire experimental process was divided into four main tasks: (i) data pre-processing; (ii) information extraction; (iii) the key pattern-specific semantic relation identification; (iv) semantic relation inference. Our proposed method was shown in Figure 1.

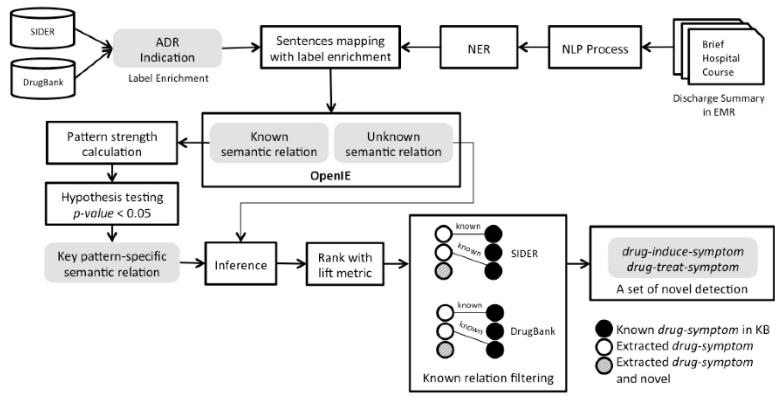


Fig. 1: The framework for semantic relations identification. The two public knowledge bases are comparative enrichment as labeled relations rather than domain expert annotation. The distributional semantic relations are examined in order to capture pattern-specific semantic relation, then the key pattern-specific semantic relation was qualified by hypothesis testing of contingency tables with 0.05 significant level. Lastly, the inference process was employed to derive novel semantic *drug-symptom* pairs.

#### 3.1 Data Preprocessing

In order to explore the relation between arbitrary two clinical entities to comprehend harmful and beneficial actions corresponding *drug-symptom* network, we set up the experimental study by utilizing the public source of EMR from MIMIC-III [19]. The database is introduced by *National Institute of Biomedical Imaging and Bioengineering* and available at *PhysioNet*<sup>3</sup>. Two sections of BHC and HPI over 58,000 observed hospital admissions were extracted. The preprocess tasks were a considerable prerequisite for further text analysis.

The sentence boundary detection (SBD) is comparatively fundamental task in Natural Language Processing (NLP), but significantly important regarding the text quality. There were a number of noise-prone in the narrative text and essentially needed to manipulate. For instance, in MIMIC-III, the period (“.”) in the narrative text can be employed as sentence boundary marker, abbreviations, level number in laboratory results, medication dosage, etc. The discontinuous text over a line is an irritated ill-form as well. The capital letter along with the punctuation mark such as colon (“:”) for content section expression also

<sup>3</sup> <https://mimic.physionet.org>

increased the challenges. Figure 2 depicted a natural format of the narrative text from MIMIC-III. We tackled such noise-prone by developing an in-house SBD rather than utilizing a *state-of-the-art* method to compatible with the narrative text from EMR. The heuristic patterns were predefined to carry out the unregulated linguistic form. Eventually, nearly 1.6 million sentences were extracted from BHC and HPI sections in the discharge summary.

In addition, highly accurate semantic relations identification is strongly related to clinical entities extraction. This is a common task in text mining, which corresponds to NER in Information Retrieval. We accomplished NER for a given narrative text using notable MetaMap tool [20]. The tool recognizes a clinical term in the given narrative text and results its standard term provided by Unified Medical Language System (UMLS). Our post processing was employed on the *out-of-the-box* MetaMap results to overcome the ambiguous NER. The two semantic group codes of *chemical* and *disorder* were considered for drug and symptom entities respectively. The summary of statistical number was placed in Table 1.

Table 1: The statistical number of contexts derived from the brief hospital course (BHC) and the history of present illness (HPI) in EMR.

	EMR Corpus			Knowledge base (KB)		
	Total	BHC	HPI	Total	BHC	HPI
Discharge summary	49,271	36,907	49,092	-	-	-
Sentences	1,580,628	980,795	599,833	-	-	-
Sentences ( <i>drug-symptom</i> )	218,135	124,074	94,061	-	-	-
Sentences/document						
+ min/max	1/251	1/248	1/118	-	-	-
+ avg./std.	31/23.7	26/19.1	12/8.7	-	-	-
Drug terms ( <i>matched to KB</i> )	3,231	2,637	2,184	-	-	-
Symptom terms ( <i>matched to KB</i> )	9,960	7,648	7,406	-	-	-
<b>Relation extraction</b>						
All open relations	1,210,501	675,664	534,837	-	-	-
Drug terms	1,142	639	977	192	168	78
Symptom terms	3,080	2,143	2,111	190	171	78
<i>drug-induce-symptom</i>	77,652	43,088	34,564	589	480	109
<i>drug-treat-symptom</i>				732	553	179

# ARF: On arrival the patient's creatinine was 1.5 (up from baseline 1.2-1.3). Over the course of his stay his creatinine increased as high as 2.2. This was believed to be multifactorial, likely due to decreased renal perfusion in setting of his arrest, infection, and possibly due to gentamycin, although his trough level never exceeded 1.9. He was followed by the renal team during his stay and had a bland urine, negative for eosinophils. His creatinine remained stable at about 2.2 for 5 days before discharge. His outpatient ramipril was held in the setting of acute renal insufficiency and may be restarted as an outpatient when his renal function is more stable.

Fig. 2: An example of narrative text in discharge summary from MIMIC-III

### 3.2 Information Extraction

A new paradigm OpenIE is a generalization of typical information extraction (IE). OpenIE provides the potential effort to deal with the large-scale corpora without manual tagging of relations [21], while the traditional one fully requires precisely target relation beforehand. Early of OpenIE [22] aimed to extract an unknown relation in advance on highly scalable Web corpus. The evident achievements on web mining let to an extensive paradigm shift in medical text mining. Recently, the Stanford CoreNLP developed OpenIE [23] in order to reduce a large pattern set for canonical sentences and excerpt self-contained clauses from longer sentences as well.

In our work, given a set of sentences  $S$  from the preprocessing process, the Stanford OpenIE was carried out to examine the powerful on clinical text mining (Figure 3). As the results, 1.2 million of the massive number of domain independent relational tuples  $\langle arg_1, pattern, arg_2 \rangle$  was reported. Unfortunately, not all but some of tuples can represent *drug–symptom* relationship. We, hence, filtered irrelevance relations and derived 77,652 *drug–symptom* relational tuples.

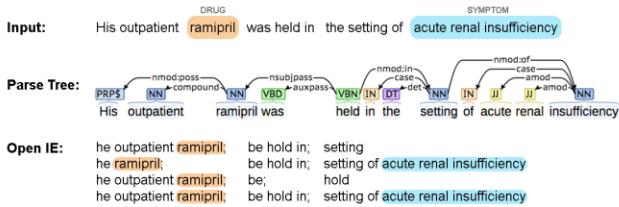


Fig. 3: An illustration of toy example of relation extraction using the Stanford OpenIE. The top of the figure is a preprocessed input. The sentence contains *ramipril* drug (concept id C0072973) and *acute renal insufficiency* symptom (concept id C0022660). The grammatical structure from parse tree of the given sentence was placed in the middle of the figure. The set of clauses from the Stanford OpenIE was shown in the bottom of the figure. It contains diverse self-contained clauses, which preserves both of syntactic and semantic entailing by the original sentence.

### 3.3 The Key Pattern-Specific Semantic Relation Extraction

The identification of semantic relation between arbitrary two clinical entities of drug and symptom is necessary for further harmful and beneficial analysis. Our hypothesis is that a mediated relation between drug and symptom entities should specific to its semantic relation (e.g. *ADR*, *indication*). This relation dependency analysis needed the known pair of *drug–induce–symptom* and *drug–treat–symptom* beforehand. To avoid the expensive of expert domain acquisition, the public and freely accessible knowledge bases from DrugBank and SIDER were comparatively incorporated as such sources of annotated data. While SIDER was utilized as harmful relation (*ADR*) and DrugBank was obtained for beneficial relation (*indication*) subsequently. The considering of annotated data from knowledge base are well-known as population base while our corpus is instance level.

The utilization of the fact from knowledge base, *firstly*, the known semantic relations were temporarily annotated into each relational tuple derived from

OpenIE. For example, *drug–symptom* pair of *amiodarone–hypotension* existed in SIDER. We, then, annotated the label of *ADR* into all relational tuples that contained *amiodarone* drug and *hypotension* symptom. Conversely, *drug–symptom* pair of *labetalol–hypertension* existed in DrugBank, therefore, the label of *indication* was annotated into all relational tuples that contained *labetalol* drug and *hypertension* as well. However, when we considered on the mediated context between *drug–symptom* pair for each sentence, the temporary annotation might not be always true.

Secondly, the pattern-specific semantic relation was investigated. Based on our hypothesis, we observed the distribution of the extracted pattern along with its semantic relation. The conditional entropy (Eq.1) was examined to quantify the degree of uncertainty for each pattern. After that, the pattern strength was obtained by conditional entropy adjustment (Eq.2), the higher score, the stronger pattern strength.

Finally, we filtered out unreliable patterns by examining the hypothesis testing of association between the semantic relation given a specific extracted pattern. The statistical Fisher's exact test at 0.05 significant level was considered. The qualified patterns were resulted as the key pattern-specific semantic relation. In summary, we derived 353 key pattern-specific semantic relation; 216 for *ADR*; 137 for *indication*. Difference from the previous study by Xu et al. [17,18], our proposed method is automatic identification, non redundant, and feasible for large amount of the key pattern-specific semantic relation derivation. We exhibited the ranking of the key pattern-specific semantic relation in Figure 4. Additionally, Table 2 illustrated the top 5 key pattern-specific semantic relation and sample sentences for *drug–induce–symptom* and *drug–treat–symptom* relations.

Given an extracted pattern  $x_i$  and semantic relation  $y_j \in Y$  where as  $Y = \{ADR, indication\}$ , the conditional entropy and pattern strength were defined as follows:

$$H(Y|X = x_i) = - \sum_{j=1}^C P(y_j|x_i) \log_2 P(y_j|x_i) \quad (1)$$

$$P_{strength}(X = x_i) = (1 - H(Y|X = x_i))(P(y_j|x_i) - (1 - P(y_j|x_i))) \quad (2)$$

### 3.4 Semantic Relation Inference

The extracted relational tuples derived from OpenIE were queried through all key pattern-specific semantic relation in order to infer the semantic relation. Then we computed the lift metric to evaluate the likelihood of a *drug–symptom* pair against the co-occurrence by chance. The lift value over than 1 implies the stronger association between drug and symptom over the chance (lift = 1).

$$lift(drug, symptom) = \frac{P(drug, symptom)}{P(drug)P(symptom)} \quad (3)$$

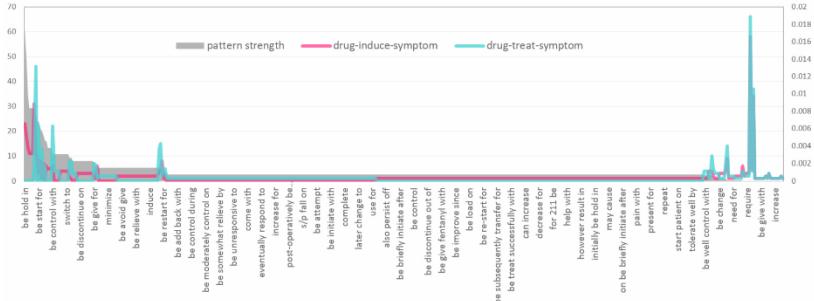


Fig. 4: The set of key pattern-specific semantic relation is ranked by the pattern strength (grey area), the higher score, the stronger pattern strength. The frequency of *drug-induce-symptom* and *drug-treat-symptom* are exhibited with red and green lines accordingly.

## 4 Experimental Results

We investigated the performance of our proposed method and reported into two parts; (i) analysis of the characteristics of discharge summary; (ii) evaluation of the key pattern specific semantic relation.

### 4.1 Analysis of the characteristics of discharge summary

In order to comprehend the data characteristics of EMR, nearly 1.6 sentences of 50,998 discharge summary was explored. The seven main sections were narrated in discharge summary (Figure 5). The maximum number of sentences was located in the BHC section. On the one hand, the HPI and the discharge medications sections contained equally number of sentences, however, contexts in the discharge medications were mostly written as a list of drug prescription regardless symptom description. Unfortunately, not all sections described the purpose of drug prescription or adverse reaction from drug usage, which can contribute to further research of *drug-symptom* network. Our work initially investigated on BHC and HPI sections that their information were closely related to *drug induce-symptom* and *drug-treat-symptom*.

From the Table 1, the BHC section contained the number of sentences more than the HPI around 1.3 times and the average sentences per document are 26 and 12 respectively. Then drug and symptom entities were extracted, it was astonished that the both sections contained equally the same number of drug and symptom entities. This is because the HPI section is permeated with clinical contents that are directly related to patient's symptom and remedy, while, the BHC narrates patient health status before, during, and after admission including treatment courses in more details. Afterwards, the relational tuples from all sentences were extracted by using the Stanford OpenIE. We found that nearly 6.4% (77,652) of all extracted relational tuples contained both drug and symptom entities in the same sentence. In contrast, the remaining (93.6%) of extracted relational tuples contained only drug, only symptom, or not related to *drug-symptom*

Table 2: Top 5 of the extracted key pattern-specific semantic relation and the example sentences from EMR

Top 5 Key patterns	Example sentences
<i>drug-induce-symptom (ADR)</i>	
1 <i>be hold in</i>	<i>His outpatient ramipril was held in the setting of acute renal insufficiency</i>
2 <i>contribute to</i>	<i>Morphine contributed to urinary retention as seen in high PVR so foley placed</i>
3 <i>be think</i>	<i>His rash was thought secondary to the nafcillin</i>
4 <i>improve with</i>	<i>Patient's dysphagia improved with iv pantoprazole</i>
5 <i>cause</i>	<i>Propofol caused mild hypotension to 95</i>
<i>drug-treat-symptom (indication)</i>	
1 <i>continue</i>	<i>Depression - continue outpatient fluoxetine</i>
2 <i>be start for</i>	<i>Phenylephrine drip was started for hypotension</i>
3 <i>be on</i>	<i>RHEUMATOID ARTHRITIS: The patient is on Methotrexate at home</i>
4 <i>be control with</i>	<i>The patient's blood pressure was controlled with labetalol</i>
5 <i>be add for</i>	<i>Norepinephrine was later added for persistent hypotension</i>

relation. Finally, nearly 1.7% (1,321) from 77,652 *drug-symptom* relation were derived as temporary semantic relations corresponding known relations from SIDER and DrugBank, hence, they were qualified for the key pattern-specific semantic relation extraction.

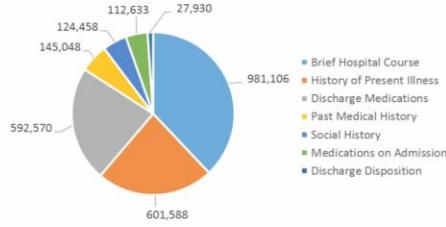


Fig. 5: The number of sentences of the seven sections in discharge summary from MIMIC-III.

Moreover, our key pattern-specific semantic relation were compared with the one from the two studies of Xu et al. The key pattern phrases from our method are rather different from Xu et al. due to different writing style (Table 3). The study of Xu et al. via MEDLINE corpus provided the shorter phrases such as only *verb* or *preposition*, and not express auxiliary verb. Our key pattern phrases derived from EMR using the Stanford OpenIE are longer with more complete phrase such as *be in* rather than *in* or *well control with* rather than *with*, etc. Mostly, the extracted key pattern phrases are fully contained *verb* and followed by *preposition*, which benefits to relation comprehension and expose theirs semantic corresponding *ADR* and *indication*. Moreover, unlike the report from two studies of Xu et al., our proposed method certainly discriminated patterns between *ADR* and *indication*, so no overlapped patterns were found.

The uncertain pattern regarding the overlapping is well-known as the cause of false positive.

Table 3: The key pattern-specific semantic relation comparison between our proposed method, which are derived by Stanford OpenIE, and the two studies of Xu et al.

	ADR		indication	
	drug-symptom	symptom-drug	drug-symptom	symptom-drug
Our method	be hold in be hold contribute to be hold for cause be discontinue be hold give ...	be in improve with be think be attribute to think feel hold for ...	continue be start for be add for be initiate for be give for be continue for be restart for ...	be continue on be control with be on well control with continue on be control on be treat with ...
Xu et al. [17, 18]	_induced induced _associated _related induces caused developed ...	induced_by after caused by following produced by after treatment in patients treated ...	in for the treatment of treatment of in the management of _resistant in a patient with to treat ...	with were treated with to after during in associate with ...

#### 4.2 Evaluation of the key pattern-specific semantic relation

The semantic relation identification of arbitrary *drug–symptom* pair can be derived by the key pattern-specific semantic relation inference. All 353 key pattern-specific semantic relation were employed. Our method successfully identified approximate to five times increasing (6,347 *drug–symptom* pair) from the known relation. We randomly selected four sets of *Metoprolol–symptom* and *drug–Hypotension* for *ADR* semantic relation, and *Amiodarone–symptom* and *drug–Pneumonia* for *indication* semantic relation. The lift metric was used to evaluate the likelihood of a *drug–symptom* pair against co-occurrence by chance (Table 4).

From FDA prescribing information, *Metoprolol* drug is indicated to treat chest pain, hypertension, and prevents heart attack. The key pattern-specific *ADR* semantic relation was used to query *Metoprolol–symptom* pair. We found that the frequent and common *ADR* caused by *Metoprolol* such as *AV block*, *Heart block*, *Hypotension* has the higher lift value, while the no frequent information or rare such as *rash* and *tachycardia* provided the small number over chance. The novel *Metoprolol–Pulmonary* pair which is not existing in the KB, was derived from our method. The RxList<sup>4</sup>, the premier Internet Drug Index resource, was used to verify the identified semantic relation. We found that *dyspnea of pulmonary origin* was reported as *ADR* of *Metoprolol* drug. Another one, the novel *Metoprolol–Kidney failure* pair was reported by Mayo Clinic<sup>5</sup>

<sup>4</sup> <http://www.rxlist.com/>

<sup>5</sup> <http://www.mayoclinic.org/>

as the possible *ADR* for long-term treatment. Identically, *drug-Hypotension* pairs regarding the key pattern-specific *ADR* semantic relation were queried. According to the fact from SIDER, hypotension is a common ADR of *Losartan* and *Metoprolol* drugs, thus, the lift value is placed in the high order. The *Ciprofloxacin-Hypotension* has the lower lift value relevance to the rare and uncommon ADR.

In the contrast, the lift value of *therapeutic indication* semantic relation is totally different from *ADR*. In which, the *indication* semantic relation provides the bigger number of lift due to the drug prescribing regularly with known indication for hospitalisation. Considering on *Amiodarone-symptom* pair, all relevance symptoms have the high lift score because *Amiodarone* drug is indicated to treat heart rhythm disorders. Another one, *Pneumonia* symptom, this symptom is an infection of the lungs and it can be caused by bacteria, viruses or fungi. All of the drugs relevance *Amiodarone-symptom* pair, that are listed in the Table 4, are indicated to treat bacterial infections. However, here in, only *Azithromycin* and *Cefepime* which were reported in DrugBank database. Our method can derive the alternative drug therapy (*Levofloxacin* and *ceftriaxone*) for *Pneumonia* with the lift of 24.95 and 14.40 respectively.

Table 4: The semantic identification regarding the key pattern-specific semantic relation. The lift value is used to evaluate the association between *drug-symptom* pair through our semantic relation identification over co-occurrence by chance. The KB column is marked *yes*, if *drug-symptom* pair exists in our knowledge base that are extracted from SIDER or DrugBank, and *vice versa*.

<i>Metoprolol-induce-symptom</i>	lift	KB	<i>drug-induce-Hypotension</i>	lift	KB
<i>AV block</i>	11.91	yes	<i>Losartan</i>	5.67	yes
<i>Heart block</i>	11.91	yes	<i>Metoprolol</i>	5.24	yes
<i>Pulmonary disease</i>	5.96	new	<i>Propofol</i>	4.99	yes
<i>Hypotension</i>	5.24	yes	<i>Carvedilol</i>	4.86	yes
<i>Asystolic</i>	3.40	yes	<i>Imdur</i>	4.86	new
<i>Sepsis</i>	2.70	new	<i>Furosemide</i>	3.60	yes
<i>Kidney failure</i>	2.38	new	<i>Nadolol</i>	3.24	yes
<i>Tachycardia</i>	2.09	yes	<i>Atenolol</i>	3.13	yes
<i>Rash</i>	1.54	yes	<i>Ciprofloxacin</i>	2.05	yes

<i>Amiodarone-treat-symptom</i>	lift	KB	<i>drug-treat-Pneumonia</i>	lift	KB
<i>Ventricular arrhythmia</i>	28.40	yes	<i>Levofloxacin</i>	24.95	new
<i>Rhythm</i>	19.88	yes	<i>Azithromycin</i>	17.43	yes
<i>Ventricular tachycardia</i>	17.04	yes	<i>Ceftriaxone</i>	14.40	new
<i>Atrial fibrillation</i>	11.09	yes	<i>Cefepime</i>	11.20	yes

## 5 Conclusions

We introduced the framework to identify semantic relation of *ADR* and *indication* from a large scale of narrative text from EMR. From our initial investigation, nearly 1.6 million sentences were examined by enrichment labeled data from the two sources of knowledge base SIDER and DrugBank. Consequently, the notable Stanford OpenIE was carried out to retrieve mediated re-

lations between two entities of *drug* and *symptom* as a result of tuple relation  $\langle arg_1, pattern, arg_2 \rangle$ . Henceforth, the conditional entropy with 0.05 significant level was presented to capture the pattern strength and automatically qualify the key pattern-specific semantic relation. Furthermore, the key pattern-specific semantic relation inference was employed in order to identify the semantic relation of new *drug-symptom* pair. Lastly, the lift metric was computed to measure likelihood of semantic association of *drug-symptom* pair. To derive the novel *drug-induce-symptom* pair and *drug-treat-symptom*, we filtered out the *drug-symptom* pair corresponding the known relations from SIDER and DrugBank respectively.

However, our method has some limitations that needs to be improved such as the low rate of recall due to small numbers of the key pattern-specific semantic relation, and the precision of drug and symptom NER. Moreover, OpenIE can retrieve diversity of relational tuples, but the method fails to discover partial or incomplete sentence especially the absent of *verb*, which is the natural pattern of narrative text in EMR.

From the experimental results, our work is not only effective and scalable for semantic relation identification, less expensive for expert annotation, but also promising framework to discover a novel harmful and beneficial drug therapeutic indication. This preliminary investigation of the utilization from EMR indicated that our contribution can support the further research of drug safety surveillance and drug repurposing as a screening method by systematic way.

### Acknowledgment

Authors thank for the research environment, which was supported by National Research University project (NRU) and the Thammasat center of Excellence (CILS), Thailand.

### References

1. X. Xu, C. Zhang, P. Li, F. Zhang, K. Gao, J. Chen, and H. Shang, “Drug-symptom networking: Linking drug-likeness screening to drug discovery,” *Pharmacological research*, vol. 103, pp. 105–113, 2016.
2. Y. Zhang, C. Tao, G. Jiang, A. A. Nair, J. Su, C. G. Chute, and H. Liu, “Network-based analysis reveals distinct association patterns in a semantic medline-based drug-disease-gene network,” *J. Biomedical Semantics*, vol. 5, p. 33, 2014.
3. I. Karlsson, J. Zhao, L. Asker, and H. Boström, “Predicting adverse drug events by analyzing electronic patient records,” in *Artificial Intelligence in Medicine*, pp. 125–129, Springer, 2013.
4. M. Y. Park, D. Yoon, K. Lee, S. Y. Kang, I. Park, S.-H. Lee, W. Kim, H. J. Kam, Y.-H. Lee, J. H. Kim, *et al.*, “A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database,” *Pharmacoepidemiology and drug safety*, vol. 20, no. 6, pp. 598–607, 2011.
5. S. Sohn, J.-P. A. Kocher, C. G. Chute, and G. K. Savova, “Drug side effect extraction from clinical narratives of psychiatry and psychology patients,” *JAMIA*, vol. 18, no. Supplement, pp. 144–149, 2011.

6. M. Banko, O. Etzioni, and T. Center, "The tradeoffs between open and traditional relation extraction," in *ACL*, vol. 8, pp. 28–36, 2008.
7. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in *IJCAI*, vol. 7, pp. 2670–2676, 2007.
8. S. Löbner, *Understanding semantics*. Routledge, 2013, 2002.
9. J. R. Hurford, B. Heasley, and M. B. Smith, *Semantics: a coursebook*. Cambridge University Press, 2007, 1983.
10. J. Lyons, *Linguistic semantics: An introduction*. Cambridge University Press, 1995.
11. A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of biomedical semantics*, vol. 2, no. 5, p. 1, 2011.
12. X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study," *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 328–337, 2009.
13. X. Wang, G. Hripcsak, and C. Friedman, "Characterizing environmental and phenotypic associations using information theory and electronic health records," *BMC bioinformatics*, vol. 10, no. Suppl 9, p. S13, 2009.
14. E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87–98, 2008.
15. F. Wang, P. Zhang, N. Cao, J. Hu, and R. Sorrentino, "Exploring the associations between drug side-effects and therapeutic indications," *Journal of biomedical informatics*, vol. 51, pp. 15–23, 2014.
16. I. Segura-Bedmar, S. De La Pena, and P. Martinez, "Extracting drug indications and adverse drug reactions from spanish health social media," in *ACL*, vol. 2014, p. 98, 2014.
17. R. Xu and Q. Wang, "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing," *BMC bioinformatics*, vol. 14, no. 1, p. 181, 2013.
18. R. Xu and Q. Wang, "Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature," *Journal of biomedical informatics*, vol. 51, pp. 191–199, 2014.
19. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).
20. A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.,," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.
21. S. Soderland, B. Roof, B. Qin, S. Xu, O. Etzioni, et al., "Adapting open information extraction to domain-specific relations," *AI magazine*, vol. 31, no. 3, pp. 93–102, 2010.
22. O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.
23. G. Angeli, M. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pp. 26–31, 2015.

# Desktop Tower Defense is $\mathcal{NP}$ -Hard

Vasin Suttichaya

Department of Computer Engineering,  
 Faculty of Engineering, Mahidol University,  
 999 Phutthamonthon 4 Road, Salaya 73170, Thailand,  
[vasin.sut@mahidol.ac.th](mailto:vasin.sut@mahidol.ac.th)

**Abstract.** This paper proves the hardness of the Desktop Tower Defense game. Specifically, the problem of determining where to locate  $k$  turrets in the grid of size  $m \times n$  in order to maximize the minimum distance from the starting point to the terminating point is shown to be  $\mathcal{NP}$ -hard. The proof applies to the generalized version of the Desktop Tower Defense.

**Keywords:** Graph theory, Hamiltonian path, Complexity

## 1 Introduction

Tower Defense (TD) is a type strategy video game that concentrates on protecting some part of the territory from waves of enemies, also known as “creeps”. Enemies always appear at the entrance and attempt to walk to the exit point. The player must plan defensive strategies for protecting their bases, usually achieved by placing turrets alongside the enemy’s road. Winning TD can be a painful task since the player needs to concurrently optimize many factors, such as resource, location, and enemy’s abilities.

Desktop Tower Defense (DTD) is a popular tower defense game. The major difference between classic TD and DTD is the player’s ability to control the enemy’s path from the starting point to the exit point. DTD allows the player to build turrets on any positions on the map. Turrets can be used as walls for blocking enemies, and force them to find the new shortest path to the exit point. The only limitation is the player cannot place turrets in the way such that they completely block the exit. Therefore, the best strategy for winning DTD is not only optimizing own resources, but also instantly extending the distance. The DTD’s gameplay is illustrated in Figure 1.

TD introduces many challenges in problem solving areas, such as resource allocation, and geometry problems. Therefore, it is interested by researchers in the artificial intelligence field. Avery et al. proposed a framework based on TD for testing artificial intelligence algorithms [3]. The dynamic difficulty adjustment of TD was proposed in [20]. The resource allocation algorithm for the turn-based game in [11] can also be applied to TD.

However, unlike other puzzle games such as Chess and Go, the hardness of TD and DTD is still unclear since they contain many sub-problems. This research



Fig. 1. Desktop Tower Defense Gameplay

attempts to formally prove one of the sub-problems in DTD, the problem of placing turrets on the map that maximizing the distance of opponent's path toward the exit point.

This paper is organized into 5 sections. Section 2, some mathematical notations are presented. The hardness of DTD is proved in Section 3. The analysis and discussion are elaborated in Section 4. The conclusion of this research is drawn in Section 5.

## 2 Preliminaries

This section starts by defining grid graph's terminology and the Hamiltonian path on the general grid graph. Then, the hardness of many games and puzzles are reviewed.

### 2.1 Grid Graph Terminology

Suppose  $G^\infty$  is the infinity graph with vertex set contains all points of the Euclidean plane with integer coordinates. Any two vertices in  $G^\infty$  are connected if and only if the Euclidean distance between them is 1. Let  $v = (v_x, v_y)$  be a vertex in  $G^\infty$  such that  $v_x$  and  $v_y$  are integer coordinates of  $v$  in  $G^\infty$ .

For any  $m, n \in \mathbb{Z}$ , let  $R(m, n)$  be the rectangular grid graph, the grid graph whose vertex set is  $V(R(m, n)) = \{v | 1 \leq v_x \leq m, 1 \leq v_y \leq n\}$ .

The arbitrary grid graph  $G$  is a finite vertex-induced subgraph of  $G^\infty$ . Clearly, each vertex in arbitrary grid graph has degree at most 4. In other words, the arbitrary grid graph  $G$  is the subgraph isomorphism of the rectangular grid graph  $R(m, n)$ .

Let  $G = (V, E)$  be an undirected graph; and let  $s, t \in V$  be distinct vertices of  $G$ . The Hamiltonian path problem, HAMPATH( $G, s, t$ ), has a solution if there exists a path from  $s$  to  $t$  that visits each node in  $G$  exactly once. In the decision

version,  $\text{HAMPATH}(G, s, t)$  is used to determine whether there is a path from  $s$  to  $t$  that visits each node in  $G$  exactly once.

The problem of finding Hamiltonian path in the arbitrary grid graph is known to be  $\mathcal{NP}$ -complete [10]. However, there exist linear-time algorithms for some special class of grid graphs [23, 21, 5, 14].

## 2.2 The Hardness of Games and Puzzles

Many classic board games were proven to be  $\mathcal{EXPTIME}$ -complete, such as Chess [19], Go [17], Chinese checkers [12], and draughts [18]. Several meta-theorems for proving the hardness of modern video games were established in [8, 22]. Some modern video games, such as Price of Persia and Doom, were proven to be  $\mathcal{PSPACE}$ -complete. Many video games, such as Tetris and Super Mario Bros, were proven to be  $\mathcal{NP}$ -hard as well [2, 4]. Kendal provided the survey of  $\mathcal{NP}$ -complete puzzles in [13].

## 3 $\mathcal{NP}$ -hardness of Desktop Tower Defense

This section starts by formally defining DTD in the term of mathematical modeling. Then, the hardness of DTD is proven.

### 3.1 Desktop Tower Defense Problem Definition

Let  $m, n, k$  be some positive integers. Suppose  $T$  is a rectangle grid of size  $m \times n$ . Without loss of generality, assume that each element in  $T$  is indexed by row-major order method (i.e., the square grid's index starts from  $T[1][1]$  at the upper left corner to  $T[m][n]$  at the lower right corner). Each element in  $T$  is marked by 0 or 1, which indicates a path and a wall respectively. Let  $W$  be a set of positions in  $T$  such that the position  $(x, y)$  in  $T$  is marked as a wall. Namely,

$$W = \{(x, y) | T[x][y] = 1 \text{ where } x \leq m \text{ and } y \leq n\}.$$

Let  $\mathbf{s} = (x_s, y_s)$  be a starting point and  $\mathbf{t} = (x_t, y_t)$  be a terminating point in  $T$ , for some  $1 \leq x_s, x_t \leq n$  and  $1 \leq y_s, y_t \leq m$ . The generalized DTD,  $\text{DTD}(T, W, k, \mathbf{s}, \mathbf{t})$ , is to determine where to locate  $k$  additional walls to the rectangle grid  $T$  so that they can maximize the shortest path from  $\mathbf{s}$  to  $\mathbf{t}$ . The only restriction is the position of all  $k$  walls must not completely block the terminating point  $\mathbf{t}$ . Therefore, there always has at least one path from  $\mathbf{s}$  to  $\mathbf{t}$ .

The generalized DTD can be also stated in the decision form. Let  $d$  be a shortest distance from  $\mathbf{s}$  to  $\mathbf{t}$ . The decision version, denote as  $\text{DDTD}(T, W, k, \mathbf{s}, \mathbf{t}, d)$ , is to determine if there is a way to place  $k$  walls in  $T$  such that the shortest path is increased to  $d$  or more. The output is yes if there is a path of length at least  $d$  after placing  $k$  additional walls, and no otherwise.

### 3.2 The hardness of Desktop Tower Defense

The  $\mathcal{NP}$ -hardness of DTD can be shown by transforming an instance of the Hamiltonian path for an arbitrary grid graph problem to an instance of the generalized DTD. Formally, the arbitrary grid graph  $G \subseteq R(m, n)$  is transformed to the rectangle grid of size  $2m - 1 \times 2n - 1$  such that the solution of the generalized DTD yields the Hamiltonian path in the arbitrary grid graph  $G$ .

**Theorem 1.** *The generalized DTD is  $\mathcal{NP}$ -hard.*

*Proof.* Suppose that  $G = (V, E) \subseteq R(m, n)$  be an arbitrary grid graph with  $|E|$  edges and  $|V|$  vertices. Given the instance of the Hamiltonian path problem  $\text{HAMPATH}(G, s, t)$ , we construct the instance of DTD by first placing vertices and edges of  $G$  to the  $(2m - 1 \times 2n - 1)$ -rectangle grid in the way such that each vertex becomes a blank square and each edge becomes a blank square. Second, flag two squares that represent vertex  $s$  and vertex  $t$  as the starting point,  $s$ , and terminating point,  $t$ , respectively. The last step, fill the rest squares that do not flagged as a blank square with walls. For example, Figure 2 illustrates the transformation of an arbitrary grid graph  $G \subseteq R(4, 6)$  to  $(7 \times 11)$ -rectangle grid  $T$ .

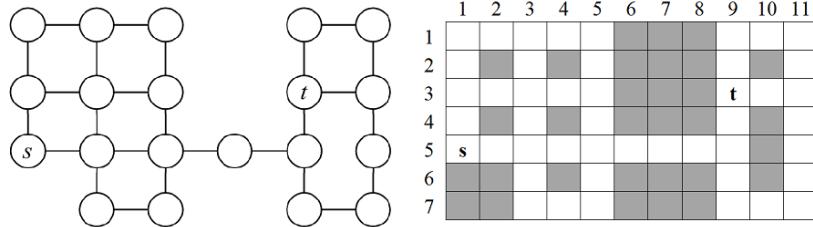


Fig. 2. Transform Grid Graph to DTD

To prove that the transform process is correct, it suffices to show that the original instance of  $\text{HAMPATH}(G, s, t)$  is a yes instance if and only if the transformed DDTD( $T, W, k, s, t, d$ ) instance is also a yes instance. The proof is shown in the following lemma.

**Lemma 1.**  $\text{HAMPATH}(G, s, t)$  has a solution if and only if  $\text{DDTD}(T, W, k, s, t, d)$ , where  $k = ||E| - (|V| - 1)|$  and  $d = 2|V| - 1$ , has a solution.

*Proof.* Suppose that  $\text{HAMPATH}(G, s, t)$  has a solution. There exists a path between vertex  $s$  and vertex  $t$  that visits each vertex in  $G$  exactly once. Note that, Hamiltonian path is also a longest path from  $s$  to  $t$ . This path has length exactly  $|V| - 1$  since it needs to connect  $|V|$  vertices without forming a cycle. This implies

that  $||E| - (|V| - 1)|$  edges are not included in the Hamiltonian path of the grid graph  $G$ . Each edge in  $G$  is represented by a blank square in  $T$ . This implies that if  $||E| - (|V| - 1)|$  blank squares that are not included in Hamiltonian path are filled with walls. The shortest path in  $T$  must follow the Hamiltonian path of graph  $G$ . The distance in the transformed grid can be calculated by subtracting all blank squares with the number of walls. The minimum distance from  $s$  to  $t$  is

$$\begin{aligned} d &= |E| + |V| - k \\ &= |E| + |V| - (|E| - (|V| - 1)) \\ &= 2|V| - 1, \end{aligned}$$

since the shortest path must take all available cells. Therefore, the shortest path in  $\text{DDTD}(T, W, k, s, t, d)$  is maximized by the longest path from  $s$  to  $t$  of  $G$ .

Suppose that  $\text{HAMPATH}(G, s, t)$  has no solution. Then, the grid graph  $G$  does not have a Hamiltonian path between vertex  $s$  and vertex  $t$ . Therefore, there does not exist a path of length at least  $|V| - 1$  that passes each node exactly once. This fact is also applied to the transformed grid  $T$  since all opponents in DTD always take the shortest path. They must not turn back to cells that have been passed. It follows that all paths from  $s$  to  $t$  in the transformed grid takes at most  $2|V| - 1$  available cells. Thus,  $\text{DDTD}(T, W, k, s, t, d)$  has no solution since it is impossible to get the minimum distance at least  $2|V| - 1$  after placing  $||E| - (|V| - 1)|$  walls.  $\square$

By Lemma 1,  $\text{HAMPATH}(G, s, t) \leq_P \text{DDTD}(T, W, k, s, t, d)$ , and the result follows.  $\square$

#### 4 Discussion

The proof in this research only considers the problem of maximizing the minimum distance. There are many problems that are embedded in the game.

In the real gameplay, there are many types of turrets. Each of them has its own firepower, ability, range, price, and cost of upgrading. The player should determine the number of turrets to buy so that the total price is less than or equal to the given gold. Moreover, the total firepower should be large enough for intercepting enemies. This problem can be classified as 0-1 Unbounded Multiple Constraint Knapsack Problem. Gens and Levner proved that this variant of Knapsack problem is  $\mathcal{NP}$ -complete [9].

Enemies in DTD also have distinct abilities, such as the resistance to certain types of turrets, the weakness against some types of turrets, the ability to spawn itself after getting the damage, and the ability to fly over the map. The player must plan the defense strategy for the given combination of enemies. The hardness of planning can be shown to be  $\mathcal{NP}$ -hard using the 3-SAT framework for proving Pushing Block puzzles [7, 6].

The problem of maximizing the shortest distance does not appear in the classic TD game. Enemies in the classic TD always walk on the predetermined

road. The player cannot place any obstructions on this road. Therefore, the main problem in the classic TD is to find where to place turrets such that their ranges cover the road as much as possible. This problem can be seen as the special case of the Art Gallery problem. The Art Gallery problem and its variations are also shown to be  $\mathcal{NP}$ -hard [15, 16, 1].

## 5 Conclusions

This research formally proves the hardness of maximizing the shortest path problem in the Desktop Tower Defense. The hardness of DTD follows from the  $\mathcal{NP}$ -hardness of the Hamiltonian path problem. The proof shows that the instance of  $\text{HAMPATH}(G, s, t)$  can be transformed to the instance of  $\text{DDTD}(T, W, k, \mathbf{s}, \mathbf{t}, d)$ , where the number of walls is  $||E| - (|V| - 1)|$  and the minimum distance is  $2|V| - 1$ .

There are open problems related to DTD and TD that have not been proven yet. The first problem is the resource allocation problem. It is easy to see that this problem is similar to the Knapsack problem. The second problem is the area coverage problem. This problem resembling the Art Gallery problem. The major difference between the Art Gallery problem and the area coverage in TD game is sentinels in the Art Gallery problem must cover all internal regions in the polygon. In contrast, turrets in TD only need to cover some limited area around the polygon edge.

## References

1. Aggarwal, A.: The Art Gallery Theorem: Its Variations, Applications and Algorithmic Aspects. Ph.D. thesis (1984)
2. Aloupis, G., Demaine, E.D., Guo, A., Viglietta, G.: Classic nintendo games are (computationally) hard. *Theor. Comput. Sci.* 586, 135–160 (2015), <http://dx.doi.org/10.1016/j.tcs.2015.02.037>
3. Avery, P., Togelius, J., Alistar, E., van Leeuwen, R.P.: Computational intelligence and tower defence games. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2011, New Orleans, LA, USA, 5-8 June, 2011. pp. 1084–1091. IEEE (2011), <http://dx.doi.org/10.1109/CEC.2011.5949738>
4. Breukelaar, R., Demaine, E.D., Hohenberger, S., Hoogeboom, H.J., Kosters, W.A., Liben-Nowell, D.: Tetris is hard, even to approximate. *International Journal of Computational Geometry and Applications* 14(1–2), 41–68 (2004)
5. Chen, S.D., Shen, H., Topor, R.W.: An efficient algorithm for constructing hamiltonian paths in meshes. *Parallel Computing* 28(9), 1293–1305 (2002)
6. Demaine, E.D., Demaine, M.L., Hoffmann, M., O'Rourke, J.: Pushing blocks is hard. *Comput. Geom.* 26(1), 21–36 (2003), [http://dx.doi.org/10.1016/S0925-7721\(02\)00170-0](http://dx.doi.org/10.1016/S0925-7721(02)00170-0)
7. Demaine, E.D., Demaine, M.L., O'Rourke, J.: Pushpush and push-1 are np-hard in 2d. In: Proceedings of the 12th Canadian Conference on Computational Geometry, Fredericton, New Brunswick, Canada, August 16-19, 2000 (2000), <http://www.cccg.ca/proceedings/2000/26.ps.gz>

8. Forišek, M.: Computational complexity of two-dimensional platform games. In: Proceedings of the 5th International Conference on Fun with Algorithms. pp. 214–227. FUN’10, Springer-Verlag, Berlin, Heidelberg (2010)
9. Gens, G., Levner, E.: Complexity of approximation algorithms for combinatorial problems: A survey. SIGACT News 12(3), 52–65 (Sep 1980), <http://doi.acm.org/10.1145/1008861.1008867>
10. Itai, A., Papadimitriou, C.H., Szwarcfiter, J.L.: Hamilton paths in grid graphs. SIAM J. Comput. 11(4), 676–686 (1982)
11. Johnson, R.W., Melich, M.E., Michalewicz, Z., Schmidt, M.: Coevolutionary optimization of fuzzy logic intelligence for strategic decision support. IEEE Trans. Evolutionary Computation 9(6), 682–694 (2005), <http://dx.doi.org/10.1109/TEVC.2005.856208>
12. Kasai, T., Adachi, A., Iwata, S.: Classes of pebble games and complete problems. In: Austing, R.H., Conti, D.M., Engel, G.L. (eds.) Proceedings 1978 ACM Annual Conference, Washington, DC, USA, December 4–6, 1978, Volume II. pp. 914–918. ACM (1978), <http://doi.acm.org/10.1145/800178.810161>
13. Kendall, G., Parkes, A.J., Spoerer, K.: A survey of np-complete puzzles. ICGA Journal 31(1), 13–34 (2008)
14. Keshavarz-Kohjerdi, F., Bagheri, A.: Hamiltonian paths in some classes of grid graphs. Journal of Applied Mathematics 2012 (2012)
15. Lee, D.T., Lin, A.K.: Computational complexity of art gallery problems. IEEE Trans. Information Theory 32(2), 276–282 (1986), <http://dx.doi.org/10.1109/TIT.1986.1057165>
16. O'Rourke, J.: Art Gallery Theorems and Algorithms. Oxford University Press, Inc., New York, NY, USA (1987)
17. Robson, J.M.: The complexity of go. In: IFIP Congress. pp. 413–417 (1983)
18. Robson, J.M.: N by N checkers is exptime complete. SIAM J. Comput. 13(2), 252–267 (1984), <http://dx.doi.org/10.1137/0213018>
19. Shannon, C.E.: Computer chess compendium. chap. Programming a Computer for Playing Chess, pp. 2–13. Springer-Verlag New York, Inc., New York, NY, USA (1988), <http://dl.acm.org/citation.cfm?id=61701.67002>
20. Sutoyo, R., Winata, D., Oliviani, K., Supriyadi, D.M.: Dynamic difficulty adjustment in tower defence. In: Procedia Computer Science. pp. 435–444 (2015)
21. Umans, C., Lenhart, W.: Hamiltonian cycles in solid grid graphs. In: FOCS. pp. 496–505. IEEE Computer Society (1997)
22. Viglietta, G.: Gaming is a hard job, but someone has to do it! Theory Comput. Syst. 54(4), 595–621 (2014), <http://dx.doi.org/10.1007/s00224-013-9497-5>
23. Zamfirescu, C., Zamfirescu, T.: Hamiltonian properties of grid graphs. SIAM J. Discrete Math. 5(4), 564–570 (1992)

## Learning Latent Word Representations for Enhanced Short Text Classification

Luepol Pipammaekaporn<sup>1</sup> and Suwatchai Kamolsantiroj<sup>2</sup>

Department of Computer and Information Science,  
King Mongkut's University of Technology North Bangkok,  
Bangkok, Thailand 10800

E-mail: {luepol<sup>1</sup>, suwatchaik<sup>2</sup>}@sci.kmutnb.ac.th

**Abstract.** Web short texts have been increasingly available in the past few years but conventional approaches to text classification are not suitable for short texts due to the data sparseness problem. In this work, we propose a novel representation learning method to tackle this challenge. Our key idea is to learn reliable low-dimensional dense representations for short text data based on latent word representations that capture semantics of words over corpus. To efficiently build the word representations, we first compute term similarity based on Word2vec, a deep learning tool that learns semantic vectors of words. We then learn a latent space of individual words from term similarity information using sparse autoencoder. By using the latent word space, we learn feature vectors for short documents based on error minimization. We conduct experiments on the two classification tasks: sentiment text classification and news title classification to evaluate the proposed method. Experimental results on two real-world datasets demonstrate that our proposed method produces more stable features that enhance short-text classification than state-of-the-art latent feature representations.

**Keywords:** short text classification, document representation, representation learning, latent features and sparse autoencoder.

### 1 Introduction

With the growing popularity of social networking web sites, such as Twitter, Micro-blogs and Facebook as well as e-commerce systems, short texts are becoming increasingly prevalent. This kind of text data poses challenges to traditional text classification methods that target at normal text. Compared to normal texts, short texts typically contain a few words in length as well as much noise (Tang et al. 2012). For example the longest tweet in twitter limited to 140 characters and Window Messenger of Microsoft allowing the longest message of 400 characters. Too limited words in short text always produce extremely sparse feature vectors and insufficient word co-occurrences to infer text similarity and the document's topic. Consequently, traditional text representation methods, such as TF-IDF, have limitations (Xu Z. et al. 2013) when directly applied to tasks in short text such as text classification (Mizzaro et al. 2014; Zhu et al. 2013) and text clustering (Yin and Wang 2014; Wei et al. 2003).

For many years, previous studies have attempted to address the extreme sparsity of short text data. Generally, these works can be categorized into two schemes. The first scheme targets on enriching the original text with additional information extracted from linguistic and collaborative repository. For example, the works studied by Mizzaro et al. (2014) and Wang et al. (2013) classifies the original text based on a predefined set of related categories in Wikipedia. In Pham et al. (2008) and Zhu et al. (2013), the authors try to learn hidden topics from large-scale Wikipedia's articles and then infer these topics for short documents. Some methods proposed to enrich the text representation with additional features extracted from linguistic knowledge-bases such as Hownet (Ning et al. 2014). However, these works are highly dependent on short text data.

In contrast with the above scheme, the other scheme explores internal semantic information, known as latent features, available in large amounts of short text data to address the limitations of the TF-IDF scheme. For example, bigram-based topic models (Yan et al. 2013) that learns topics over an entire corpus of short texts by modeling unordered word-pair co-occurrences. Some experiments reported that Latent Dirichlet Allocation (LDA) (Blei et al. 2003) has benefited for the classification of twitter posts (Hong et al. 2010). Matrix factorization-based methods such as Latent Semantic Analysis (LSA) (Pu and Yang 2006) and Non-Negative Matrix Factorization (NMF) (Xu W. et al. 2013) were applied to convert sparse vectors of texts to a low-dimensional dense space. Despite this, these methods often suffer from sparse word co-occurrences patterns when applied to short texts, leading to reduced generalization accuracy.

Motivated by these works, we propose a novel data representation method that learns compact and dense representations for short text classification. Our proposed method uses term similarity matrix that is less sparse but more stable data than sparse term-document matrix commonly used in conventional approaches. The term similarity matrix is also generated based on Word2Vec (Mikolov et al. 2013), an efficient tool that computes vector representations for words over corpus, capturing word contexts to represent similarity between words. Once obtained the term similarity matrix, we learn a compact latent space that shares semantics of the words using sparse autoencoder (Hinton and Ruslan 2006; Coates et al. 2013). By using the latent word space, we compute features for documents based on least square error method. We evaluate the proposed method in two classification tasks, i.e. sentiment text classification and news title classification. We conduct experiments on two real-world datasets, including Sentiment140 and 20newsgroups datasets. The experimental results demonstrated that our proposed method outperforms both LDA and LSA in both the tasks.

The main contributions of our work are summarized as follows:

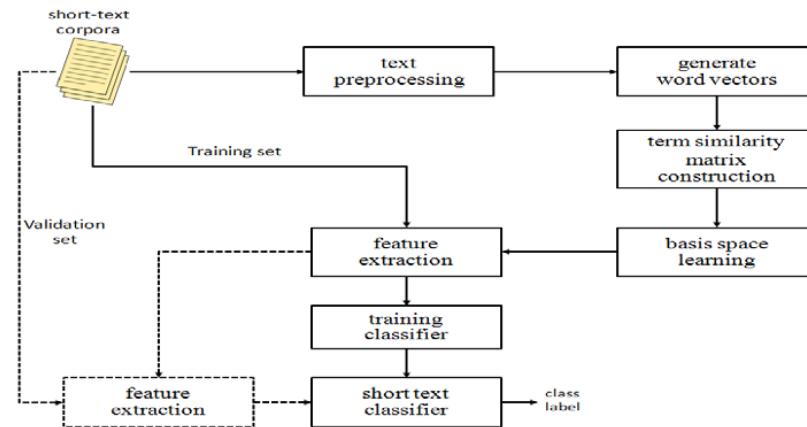
- We propose a novel representation learning method that overcomes the issue of over-sparsity in short text data by using term similarity data without any external data sources.
- Our method learns compact and dense representations in a low-dimensional space that better represents and classifies sparse short texts.
- We conduct experiments on two real-world datasets and demonstrate the effectiveness of the proposed method for short text classification compared to state-of-the-art text representation methods.

The rest of the paper is organized as follows. In Section 2, describes the overall framework of our proposed approach. In Section 3, we explain the details of our

methodology. The experimental settings and results are shown in Section 4. Finally, we make conclusions in Section 5 respectively.

## 2 Our proposed framework

Figure 1 illustrates our proposed framework for enhanced representations of short texts. We briefly explain the proposed framework as follows:



**Fig. 1.** Our proposed framework

**Step 1:** Construct term similarity matrix. In this step, each short text is preprocessed as the appropriate input of Word2vec. By using Word2vec, semantic vectors of each word are automatically learned from the training data. After that, the term similarity matrix  $S$  can be constructed using a vector similarity measure such as cosine similarity.

**Step 2:** Basis space learning stage. In this step, we learn a compact latent space  $Z$  that shares semantics of the words in the term similarity matrix  $S$ . In this work, we learn the latent basis matrix by formulating an objective function and minimizing this function using sparse autoencoder.

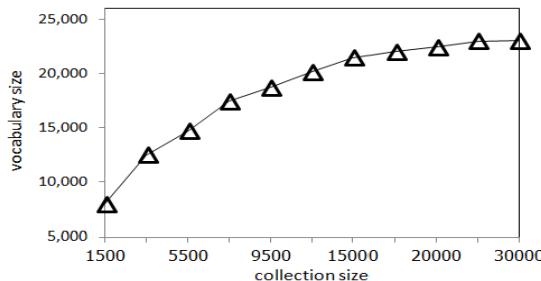
**Step 3:** Document representation stage. In this step, we compute features for each document based on the latent word space. We treat this problem as an optimization problem by minimizing an objective function based on least square fitting.

**Step 4:** Build a text classifier. In training process, we build a classifier based on the latent representation of short texts in the training set and then classify the texts in the validation set to obtain the best parameters.

### 3 Methodology

#### 3.1 Term Similarity Matrix

Most existing latent representation methods, such as LDA and LSA, usually learn a latent space from normal texts by decomposing the term-document matrix in several ways. However, when document length reduces, these methods that use document-level word co-occurrence statistics can suffer from data sparsity. To alleviate this problem, we need to learn the latent representation from data that is denser and more stable than term-document matrix. For this purpose, we utilize term similarity data for some reasons. First, the term similarity data is much denser than the term-document matrix. Second, this kind of data tends to be stable when the collection size of documents grows increasingly. Figure 2 illustrates the vocabulary size when the collection size of twitter posts in sentiment140 corpus becomes larger where terms whose document frequency is less than 4 are removed.



**Fig. 2.** The vocabulary size by varying the collection sizes of twitter posts in Sentiment140 dataset<sup>1</sup>

As seen in Figure 2, the vocabulary size of unique terms become more stable when the collection size of documents grows increasingly. To obtain the term similarity data, we employ Word2vec (Mikolov et al. 2013) an efficient deep learning tool for computing vector representations for words from Google. In Word2vec, feature vectors for individual words over corpus are learned using a two-layer neural network. The feature vectors basically capture word contexts within the contextual window, representing semantic similarity between the words.

Word2vec consists of two distinct models, namely skip-gram and CBOW (continuous Bag-of-Words). The skip-gram model learns the probability of context words given a word  $w_t$  by minimizing loss function:  $E = -\log p(w_{t-j}, \dots, w_{t+j} | w_t)$  where  $j$  is a window size. Conversely, CBOW learns to predict a word  $w_t$  given the context words. In this work, we choose the skip-gram model to build the term similarity matrix in a short-text collection because previous studies have shown suitable for sparse data and rare words compared to CBOW (Mikolov et al. 2013; Guthrie et al. 2006). Figure

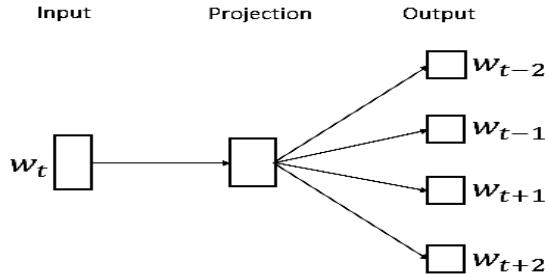
<sup>1</sup> <http://www.sentiment140.com>

3 illustrates the skip-gram architecture that learns to predict the context words  $w_{t-2}, w_{t-1}, w_{t+1}$  and  $w_{t+2}$  given the word  $w_t$  from text corpus.

We first preprocess each short text in a training set by eliminating stop words and word stemming. We then segment the preprocessed short texts into a collection of sentences as input for the skip-gram model. Since the implementation of this model is available in Word2vec tool<sup>1</sup>, we used it for inducing the word representations. Table 1 demonstrates parameter settings used for Word2vec. We used a minimum count of 4 occurrences for each word to reduce the vocabulary size. After learned the feature vectors of words, we compute the vector similarity between words  $w_i$  and  $w_j$  using cosine similarity.

$$\text{sim}(w_i, w_j) = \frac{C(w_i) \times C(w_j)}{|C(w_i)| \cdot |C(w_j)|} \quad (1)$$

where  $C(w_i)$  indicates the learned feature vector of word  $w_i$  and  $|C(w_i)|$  is the vector length of the word  $w_i$ . We finally construct term similarity matrix  $S$  from the vocabulary.



**Fig. 3.** Skip-gram achitecture

### 3.2 Latent space learning using sparse autoencoder

The next stage is to learn a compact latent space that shares semantics of words from term similarity matrix. We formulate this problem as an error minimization problem by using the following metric:

$$J(Z) = \|S - ZZ^T\|^2 \quad (2)$$

where  $S$  and  $Z$  indicate the term similarity matrix and latent basis matrix respectively. Each row of  $Z$  represents words in vocabulary whilst each column represents latent variables. Intuitively, the above metric can be considered as a distance function that measures the difference between term similarity matrix  $S$  and latent basis matrix  $Z$ .

---

<sup>1</sup><http://code.google.com/p/word2vec/>

To minimize the above metric, we apply autoencoder (Hilton and Ruslan 2006), a method to learn high-level feature representations of original data. Specifically, the autoencoder learns the output  $\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$  that reconstructs the input  $X = \{x_1, x_2, \dots, x_n\}$  with a hidden node layer  $Z = \{z_1, z_2, \dots, z_k\}$  where  $0 < k < n$ . The value of hidden layer node  $z_i$  is defined as

**Table 1** Parameter settings used in Word2vec

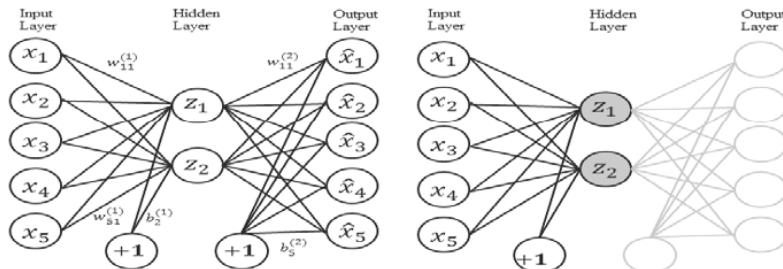
Parameter	Description	Value
-train	Name of input file	“word.txt”
-output	Name of output file	“vocab.out”
-cbow	training model 0: Skip-gram model 1: CBOW	0
-size	dimension of vectors	200
-window	window size	5
-negative	training method 0: hierarchical softmax 1: negative sampling	0
-sample	threshold of sampling	4
-threads	number of running threads	12
-mode	output mode 0: binary, 1:text	1

$$z_i = b_i^{(1)} + \sum_{j=1}^n w_{ij}^{(1)} x_j \quad (3)$$

where  $b_i^{(1)}$  is the bias of hidden node  $i$  and  $w_{ij}^{(1)}$  represents the weight between input node  $j$  and hidden node  $i$ . The output layer  $\hat{X}$  is built by using the activations of the hidden layer as input, bias  $b^{(2)}$  and weights  $W^{(2)} = \{w_{11}^{(2)}, w_{12}^{(2)}, \dots, w_{kn}^{(2)}\}$ :

$$\hat{x}_i = f \left( \sum_{j=1}^k w_{ij}^{(2)} a_j \right) \quad (4)$$

where  $f(z)$  is an activation function and  $a_j = f(z_j)$  is the output activation of hidden node  $j$ . Figure 4 illustrates the autoencoder network (left) and the output of the trained network (right).



**Fig. 4.** The autoencoder neural network (left) the network structure and (right) hidden node representation of input vector with trained network

In this work, we train an auto-encoder with  $k$  hidden nodes using backpropagation with an additional term of sparsity penalty that produces sparse representation in the hidden layer (Le 2013). We apply a linear activation function to the nodes comprising the neural network. Let  $S^{(i)} = \langle s_1^i, s_2^i, \dots, s_n^i \rangle$  be a feature vector of word  $i$  and  $\hat{S}^{(i)} = \langle \hat{s}_1^i, \hat{s}_2^i, \dots, \hat{s}_n^i \rangle$  be output vector for this word predicted by the neural network. The overall objective function is minimized by the autoencoder as follows:

$$J(W, B) = \underset{W, B}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \| \hat{S}^{(i)} - S^{(i)} \|^2 + \frac{\lambda}{2} \| W \|^2 + \beta \sum_{j=1}^k KL(\rho || \hat{\rho}_j) \right\} \quad (5)$$

where  $W, B$  indicate weights and biases of the training network.  $KL(\rho || \hat{\rho}_j) = \rho \cdot \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \cdot \log \frac{1 - \rho}{1 - \hat{\rho}_j}$  is the Kullback-Leibler (KL) divergence term. The value  $\rho$  is a sparsity parameter that indicates the desired frequency of activation of the hidden nodes. In the KL divergence term,  $\hat{\rho}_j$  is the average threshold activation of hidden node  $j$  over all training examples. In this work, we threshold the hidden node activations whose their output is greater than 0 become 1 and 0 otherwise. The values  $\lambda, \beta$  are hyperparameters that determine the relative importance of the weight decay regularization term and the sparseness term in the cost function. After trained the network, we extract  $k$  hidden node representation for each word vector  $S^{(i)} \in S$ .

$$Z^{(i)} = \langle z_1^{(i)}, z_2^{(i)}, \dots, z_k^{(i)} \rangle \quad (6)$$

We finally build the latent basis matrix  $Z$  from all the words in vocabulary and use this matrix to compute features for individual documents in the next stage.

### 3.3 Feature learning by least square error

The latent basis matrix  $Z$  basically captures the latent structure of individual words in a collection of documents. Once obtained, the next step is to compute latent features for documents representation. Most existing approaches compute the features for documents by projecting each document vector into the latent space that has much less dimensions but more semantics than the original matrix. However, this can be problematic for short texts because the extremely sparse and large term-document matrix  $X$  often obstructs estimating reliable features. To alleviate this problem, we formulate this problem as a least squares problem whose goal is to estimate matrix  $U$  that minimizes the error between matrix  $X$  and matrix  $ZU$  using the following function.

$$D(U) = \| X - ZU \|^2_F \quad (7)$$

where  $U$  indicates feature-document matrix. As seen in equation (8), the matrix  $U$  can be estimated by minimizing the Euclidean distance between the term-document matrix  $X$  and the product  $ZU$ , given matrix  $U$ . We then apply least square solution for estimating matrix  $U$  as the following equation.

$$\hat{U} = (Z^T Z)^{-1} Z^T X \quad (8)$$

---

**Algorithm 1** : The overall procedure of our proposed approach

---

**Input:** term-document matrix  $X$ , the latent feature number of  $k$   
**Output:** latent basis matrix  $Z$ , feature-document matrix  $\hat{U}$

- 1: Generate term similarity matrix  $S$  using skip-gram model
- 2: **Repeat:**
- 3: Minimize  $J(W, B)$  (5) by training sparse autoencoder
- 4: **Until** convergence;
- 5: Generate latent matrix  $Z$  from  $k$  hidden node activation vectors of each input  $S^{(t)}$  with the trained network
- 6:  $\hat{U} \leftarrow (Z^T Z)^{-1} Z^T X$
- 7:
- 8: **return**  $Z, \hat{U}$

---

Algorithm 1 describes the overall procedure of our proposed approach. The input data for the algorithm include a term-document matrix  $X$  and the number of  $k$  desired features. The rows in this matrix correspond to the documents and the columns to the terms. The outputs of the algorithm include the latent basis matrix  $Z$  and the feature-document matrix  $\hat{U}$ . After learning the matrix  $Z$  and  $\hat{U}$ , we can apply a standard machine learning technique, such as support vector machine (SVM) (Joachims 1999) to build a text classifier from the induced representations associated with their classes.

## 4 Experiments

In this section, we report empirical experiments to evaluate the proposed method. We test our method on two common tasks in short texts: 1) *sentiment text classification* and (2) *news title classification*. We also demonstrate the effectiveness of our approach on two real-world datasets and compare it with state-of-the-art feature representation methods for documents, including term frequency-inverse document frequency (TF-IDF), latent semantic indexing (LSI) and latent dirichlet allocation (LDA).

### 4.1 Datasets

We conducted this experiment on the two real-world datasets: 1) Sentiment140<sup>1</sup>, a tweet sentiment dataset and 2) 20 newsgroup<sup>2</sup> dataset. The twitter dataset contains 1,600,000 English twitter posts collected during April, 6, 2009 to June 25, 2009. Each tweet was hand-classified with either positive or negative sentiment. All special characters and emoticons are removed. We select only 40,000 tweets in each sentiment category for the evaluation of effectiveness. For news title classification, we choose the well-known 20 newsgroups (20NG) dataset, which contains over 20,000 English posts from 20 different newsgroups. We only experiment with six of these categories, in-

---

<sup>1</sup> <http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip>

<sup>2</sup> <http://www.ai.mit.edu/~jrennie/20Newsgroups>

cluding *comp.graphics*, *rec.autos*, *comp.sys.ibmpc.hardware*, *comp.sys.mac.hardware*, *sci.space*, and *rec.motocycles*. For the short text scenario, we only consider the subject field of each document as training and test data and discard other information, named “20NG” for short. Table 2 shows statistics of the two datasets. Both the datasets are preprocessed by using standard text processing tasks, including stop word removal and word stemming using Porter algorithm (Willet 2006), to reduce noises in text.

**Table 2** Statistical information of the two datasets

Dataset	Tweet	20NG Title
#documents	80,000	5,845
#unique words	2,514	1,860
average word per document	12.59	9.26
#class	2	6

#### 4.2 Experimental setup

We summarize the baselines as follows:

- **Term Frequency-Inverse Document Frequency (TF-IDF)**: we first compared the most basic feature representation for documents. Each document is represented as a word vector with weights that represent the importance of that word in the document and the text collection. For the short text scenario, rare words are typically brought. We do not perform any feature selection.
- **Latent Semantic Indexing (LSI)**: we also compared with LSI, the well-known latent space method that factorizes the term-document matrix into term-concept matrix and concept-document matrix. We split the training set into training and validation to tune the best parameter, which is the desired number of  $k$  concepts, by varying the different  $k$  values 5, 10, 20, 40 and 80. After that the LSI-based features are obtained by projecting the TF-IDF vectors into the  $k$  subspace.
- **Latent Dirichlet Allocation (LDA)**: we compared with LDA, a generative probabilistic model that discovers topics over corpus. We use the Gibb sampling-based LDA implementation in MATLAB<sup>1</sup>. Similar to LSI, we use the validation set from the training data to find the best model parameters, including the Dirichlet parameter  $\alpha$  and the desired number of K topics. In this feature representation, each document is represented as a distribution over topics.

We evaluate all the methods using a linear support vector machine classifier LIBLINEAR (Joachims 1999), a software library for large-scale linear classification. In each dataset, documents are randomly split into training and testing with 80% and 20% ratio respectively. For multi-class classification, we use the one-vs-the rest approach to train the SVM classifier.

---

<sup>1</sup> [http://psiexp.ss.uci.edu/research/programs\\_data/toolbox.htm](http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm)

Our proposed method includes three important parameters: the hyperparameters  $\lambda$ ,  $\beta$  and the desired number of  $k$  hidden nodes in the autoencoder network. In this work, we use sparse autoencoder implementation in Matlab code developed by Standford<sup>1</sup> for this experiment. The parameters  $\lambda$  and  $\beta$  can be automatically tuned using cross-validation on the validation set.

#### 4.3 Performance metrics

We evaluate the classification performance of all the methods using  $F_1$  measure. The  $F_1$  measure is a combination of both precision (P) and recall (R) performance metrics that indicate the extent to which a group of classified documents by system belonging to a particular class. The precision and recall for each of the classifiers  $c$  can be computed by the following formulas:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \quad (9)$$

where  $TP_c$  and  $FP_c$  are count of the documents correctly and incorrectly classified to the category  $c$  respectively.  $T_N$  and  $F_N$  are count of the documents correctly and incorrectly rejected from the category  $c$ . We then calculated  $F_1$  score that combines marco-averaged precision (P) and recall (R) as follows:

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

#### 4.4 Experimental results

Table 3 includes the average  $F_1$  score results for each of the representation methods on the two datasets. As can be seen from Table 3, SVM generated the best  $F_1$  scores with our method at all the  $k$  values where it reached the best performance on 20NG dataset at  $k = 80, \lambda = 0.07$  and at  $k = 80, \lambda = 0.04$  on the twitter dataset. SVMs generated by LDA and LSI were also achieved the best performance at between  $k = 40$  and  $k = 80$  for both the datasets. The most interesting findings revealed in Table 3 include:

- The big improvements achieved by all the methods (i.e. LDA, LSI and our method) over the TF-IDF method demonstrate the effective use of latent features to address the over-sparsity issue in short text data.
- It is clear that LDA-based features perform better than LSI in terms of  $F_1$  score. This result suggests that LDA can estimate word co-occurrence statistics over the corpus in an intuitive way, compared to the LSI features that solely come from linear combinations of the TF-IDF features into a low-dimensional subspace.
- Our proposed scheme clearly outperforms the LDA-based features and all the baseline methods for the classification tasks, especially twitter sentiment classification. Moreover, the encouraging improvement of our method is consistent to the news title data, which is highly sparse (the average document length is at 9.26). These results highlight our method that can learn robust features that better represent short texts. Furthermore, we find that the representations directly learned over term similarity data rather than term-document matrix are stable and sufficient in

---

<sup>1</sup> <http://ufldl.stanford.edu>

capturing meanings of words such as polysemy and synonymy within a very limited text.

**Table 3** Average  $F_1$  scores ( $\pm$  Standard deviation) on different methods

Data-set	Method	#Topics (K)				
		K=5	K=10	K=20	K=40	K=80
20NG	LDA	0.553( $\pm 0.019$ )	0.605( $\pm 0.009$ )	0.700( $\pm 0.007$ )	0.726( $\pm 0.015$ )	0.724( $\pm 0.003$ )
	LSI	0.501( $\pm 0.021$ )	0.564( $\pm 0.008$ )	0.629( $\pm 0.008$ )	0.642( $\pm 0.017$ )	0.667( $\pm 0.003$ )
	TF-IDF	0.349( $\pm 0.721$ )				
	Our method	<b>0.574</b> ( $\pm 0.022$ )	<b>0.664</b> ( $\pm 0.003$ )	<b>0.763</b> ( $\pm 0.010$ )	<b>0.782</b> ( $\pm 0.014$ )	<b>0.789</b> ( $\pm 0.005$ )
Tweet	LDA	0.449( $\pm 0.029$ )	0.510( $\pm 0.016$ )	0.587( $\pm 0.014$ )	0.632( $\pm 0.008$ )	0.633( $\pm 0.003$ )
	LSI	0.424( $\pm 0.035$ )	0.505( $\pm 0.023$ )	0.561( $\pm 0.017$ )	0.603( $\pm 0.012$ )	0.607( $\pm 0.002$ )
	TF-IDF	0.331( $\pm 0.442$ )				
	Our method	<b>0.601</b> ( $\pm 0.024$ )	<b>0.678</b> ( $\pm 0.019$ )	<b>0.756</b> ( $\pm 0.013$ )	<b>0.764</b> ( $\pm 0.018$ )	<b>0.770</b> ( $\pm 0.003$ )

## 5 Conclusion

We have proposed a novel representation learning method for short texts classification that still poses many challenges to conventional text representation methods. Our method focuses on alleviating the sparseness of short text data by utilizing term similarity matrix that is less sparse but more stable for feature learning than term-document matrix used in most existing latent representations. To achieve this goal, we first employ Word2vect over corpus to capture word contexts for representing semantic similarity between words in a very limited text. After that, we learn a compact latent space that shares semantics of words in the term similarity matrix using sparse autoencoder. By using the latent space, we learn features that represent each document by solving least square solution. We conducted experiments on the two classification tasks: tweet sentiment classification and news title classification. Experimental results on two real-world datasets demonstrated that our proposed method significantly improves the classification accuracy compared to state-of-the-art text representation methods.

## References

- Tang, J., Wang, X., Gao, H., Hu, X. and Liu, H.: Enriching short text representation in microblog for clustering. *Frontiers of Computer Science*, 6(1), 88-101 (2012).
- Mizzaro, S., Pavan, M., Scagnetto, I., and Valenti, M.: Short text categorization exploiting contextual enrichment and external knowledge. In Proc. of the first international workshop on Social media retrieval and analysis, 57-62. (2014).
- Phan, X.H., Nguyen, L.M. and Horiguchi, S.: Learning to classify short and sparse text & web

- with hidden topics from large-scale data collections. In Proc. of the 17th international conference on World Wide Web, 91-100 (2008).
- Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proc. of the 17th international conference on World Wide Web. (2008).
- Zhu, Y., Li, L. and Luo, L.: Learning to classify short text with topic model and external knowledge. In Knowledge Science, Engineering and Management, 493-503 (2013).
- Yin, J. and Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In Proc. of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining , 233-242 (2014).
- Genc, Y., Sakamoto, Y. and Nickerson, J.V.: Discovering context: classifying tweets through a semantic transform based on wikipedia. In Foundations of augmented cognition Directing the future of adaptive systems, 484-492 (2011).
- Wang, X., Chen, R., Jia, Y. and Zhou, B.: Short Text Classification using Wikipedia Concept based Document Representation. In Proc. of the IEEE International Conf. on Information Technology and Applications, 471-474 (2013).
- Yan, X., Guo, J., Lan, Y. and Cheng, X.: A biterm topic model for short texts. In Proc. of the 22nd international conference on World Wide Web, 1445-1456 (2013).
- Hong, L. and Davison, B.D.: Empirical study of topic modeling in twitter. In Proc. of the first workshop on social media analytics, 80-88 (2010).
- Pu, Q., & Yang, G. W.: Short-text classification based on ICA and LSA. In Advances in Neural Networks-ISNN, 265-270 (2006).
- Xu, W., Liu, X. and Gong, Y.: Document clustering based on non-negative matrix factorization. In Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 267-273 (2003).
- Ning, Y.H., Zhang, L., Ju, Y.R., Wang, W.J. and Li, S.Q.: Using Semantic Correlation of HowNet for Short Text Classification. In Applied Mechanics and Materials, Vol. 513, 1931-1934 (2014).
- Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space. In Proc. of International Conference on Learning Representation (2013).
- Coates, A., Ng, A.Y. and Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In International conference on artificial intelligence and statistics, 215-223 (2011).
- Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks.: A closer look at skip-gram modeling. In Proc. of the 5th International Conf. on Language Resources and Evaluation, 1-4 (2006).
- Le, Quoc V.: Building high-level features using large scale unsupervised learning. In Proc. of IEEE International Conf. on Acoustics, Speech and Signal Processing, 8595-8598 (2013).
- Salton, G. and McGill M.J.: Introduction to modern information retrieval. (1986).
- Hinton, Geoffrey E., and Ruslan R. Salakhutdinov.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504-507 (2006).
- Joachims, T.: Making large scale SVM learning practical. Universität Dortmund, (1999).
- Willett, Peter.: The Porter stemming algorithm: then and now. Program 40(3), 219-223 (2006).
- Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022 (2003).

## **Variant Annotation and Clinical Interpretation software for Cancer (VARCIN): Report Generating Software for Targeted Therapy Method**

Pitinat Asawasutsakorn<sup>1</sup>, BenjamardMeeboon<sup>1</sup>, Natini Jinawath<sup>2</sup>

Mingmanas Sivaraks<sup>1</sup>

<sup>1</sup>Image, Information and Intelligence Lab, Department of Computer Engineering, Faculty of Engineering, Mahidol University 25/25 Puttamonthon Nakhon Pathom 73170, Thailand  
pitinat@gmail.com, benjamard.mee@gmail.com,  
mingmanas.siv@mahidol.ac.th

<sup>2</sup>Faculty of Medicine Ramathibodi Hospital Mahidol University, 270 Rama VI Rd., Ratchatewi, Bangkok 10400, Thailand  
natini.jin@mahidol.ac.th

**Abstract.** Variant Annotation and Clinical Interpretation Software for Cancer (VARCIN) is a report generating software for producing an executive summary for a cancer patient. The report can assist physicians and researchers by using the result to find a specific or recommended medicine for each specific individual with cancer, known as “Targeted Therapy”, much more efficient. This software can shorten the process and provide an all-in-one solution for classifying and recommending treatments for cancer patients. Moreover, this software is developed as a web-based application with a responsive feature, so this software can be used anywhere and with any electronics device that has an access to a web browser with internet connection, improving ease-of-access thus more convenient than the traditional way.

**Keywords** Variant call format, cancer, report generating, cancer diagnostic software, targeted therapy, personalized medicine

### **1 Background and Significance**

Cancer is a genetic disease. A cause of cancer is when the cells in the body grow rapidly and uncontrollably, and subsequently kills normal cells. There are many types of cancer, each type of cancer affect on different area of the body. Cancer is a leading cause of death worldwide. There were an estimated 14.1 million cancer cases around the world in 2012, which 7.4 million cases were in men and 6.7 million in women. This number is expected to increase to 24 million by 2035. From the research show us how much the risk of cancer and why the cancer treatment nowadays is important (J. Ferlay, 2012).

In past decades, the most popular techniques for treating cancer patient are radiation therapy and chemotherapy, the purpose of both methods is to destroy cancer cells. However, these

methods have some disadvantages. They can destroy cancer cells, but they also inevitably kill normal cells. Especially normal body cells that have rapidly dividing behavior like cancer, for example skin cells, hair cells and epithelium cells. Although most of side effects will gradually subside after the treatment ends and the healthy cells have a chance to grow normally, but losing normal cells will have consequent side effect such as fatigue, pain, hair loss, blood clotting problems and radiation recall ((ASCO), 2015).

“Targeted therapy” uses a personalized medicine for cancer to select specific drugs designed to interfere with specific molecules necessary for individual tumor growth and progression. Traditional cytotoxic chemotherapies usually kill rapidly dividing cells in the body by interfering with cell division. A primary goal of targeted therapies is to fight cancer cells with more precision and potentially fewer side effects (Abramson, 2016). Although targeted therapy treatment is a good way for new cancer treatment, but a weak point of targeted therapy is a doctor or a researcher must analyze several data from various sources, then interprets it by themselves. Because of this reason, the doctor or the researcher wasted a lot of time for diagnosis each cancer patient.

Variant annotation and clinical interpretation software for cancer (VARCIN) is created for solving these problems. The main purpose of this software is for the most convenient use for a doctor or a researcher without the high cost, decreasing times in analyzing data and generates an interpretation report for cancer diagnosis with high efficiency, accuracy and compact. The software will generate a summarized report that includes advantageous data for diagnosing for each cancer patient, expediently and time-saving. This software is a web based application with a responsive platform, which is easy to access either by the PCs, tablets or smart phones.

## 2 Method

Variant Annotation and Clinical Interpretation Software for Cancer (VARCIN) has been developed using PHP as a core programming language and has been associated with ANNOVAR software (Wang K, 2016). ANNOVAR is free annotation software developed using a Perl script. ANNOVAR works by taking a data from the Variant Call Format (VCF) input ([Wikidata](#), Variant call format, 2016) which we obtained from tissue sample in laboratory by DNA sequencing compare with a preloaded database in the server. If the record does exist in any database, it will be shown in the output file, then VARCIN will analyze output file, filter low-quality data out and sort every record based on its score. Finally, VARCIN will generate a summary file as a pdf and full output file as a .csv file.

VARCIN uses some useful technical information from mycancergenome (Abramson, 2016) and pct.mdanderson (Pruthi. S., 2015) which provide us information such as clinically available drug for a specific gene, gene characteristic, implications for targeted therapeutics. The users can read more information by accessing the link provided.

VARCIN consists of 3 steps of work which are (1) Patient data record and annotation (2) Filtering and scoring data (3) Prepare and produce a summary report.

## 2.1 Patient data record and annotation

For every new patient, the user has to fill in the patient personal information. This information will be presented at the top of summary file and will not be revealed to the public or unauthorized users. A VCF file will be selected with some configurations. Any record in the VCF that do not meet requirements will be filtered out in this process to reduce the processing time of the program rather than using the whole VCF file which usually has the size more than 1 GB. This filtering process helps reducing some data that tend to be low accuracy and low quality out.

The file selected will be queued for processing in the program. The server will process one file at a time and take it as a first come first serve basis. After the previous files are finished, the file will be processed starting by the first stage called “annotation”. ANNOVAR will be used in this annotation step by comparing gene data to the database, if the data exist, it will be recorded as an output file.

## 2.2 Filtering and scoring data

After the annotation step, data will be filtered out with VARCIN by using the configuration defined by a user. Then the filtered data will be given different scores based on its confidence level. If the gene record is found to be a pathogenic gene, the system will give a positive score value, but if it is found to be a non-pathogenic gene, the record will be given a negative score value.

The databases are arranged in 3 categories depending on its confidence value for better evaluation, scores will be calculated separately from each category. Finally, each database group has been given different score types which will be described here.

- **Group I: Certified and verified database.**

This type of database includes certified data from national research center or organization, thus can be trusted with higher accuracy comparing to other databases. The Databases in this group are COSMIC (Wikidata, COSMIC cancer database)andClinVar (Landrum MJ, 2015).Table 1 shows an example result from ClinVar. The result from both databases will be either “Pathogenic” or “Non-Pathogenic”, VARCIN gives high score in favour for these databases which as 10 and -10 respectively. The score of gene in this group is shown in Table 2.

**Table 1.** Example result from ClinVar

Database	Sample Output
<b>Clinvar</b>	CLINSIG=untested; CLNDBN=not_specified; CLNREVSTAT=no_assertion_provided; CLNACC=RCV000121149.1; CLNDSDB=MedGen; CLNDSDBID=CN169374

**Table 2.** Score of gene in Certified and trustworthy database.

Database	Pathogenic	Non-Pathogenic
<b>COSMIC, Clinvar</b>	+10	-10

- **Group II: Allele frequency database.**

Since the data obtained from Group I are the certified data from organizations, therefore the information can not include the most up-to-date data thus fewer data will be found. More databases will be used to incorporate to support better decision making process.

The group II database uses the allele frequency database from the sample population group. It consists of frequency of pathogenic or non-pathogenic gene. In VARCIN, Databases in this group are 1000 Genomes (Landrum MJ, 2015), ExAc, ESP6500. The example of allele frequency database is shown in Table 3.

**Table 3.** Score of gene in allele frequency database.

Database	Sample allele frequency Output
<b>1000Gnome</b>	0.91
<b>ExAc</b>	0.8755
<b>ESP6500</b>	0.92

The allele frequency is the frequency of the gene to be found in a sample population group. If allele frequency is high in the databases of normal people, it means that the gene is more likely to be non-pathogenic. But if the allele frequency is found to be high in the cancer patient databases, the gene is in reverse more likely to be pathogenic. The databases collected in this group II are from both normal people and from cancer patients. The score given for this group is given as +1 and -1, as shown as Table 4 according to more likely and less likely to be pathogenic respectively. The threshold of the frequency can be selected by users; otherwise the default threshold value of the allele frequency is 0.05. The score is much smaller compared to the Group I because of it is not fully verified. However, if many databases in this group are supporting one another, the score will be become higher.

**Table 4.** Score of gene in Certified and trustworthy database.

Database	Frequency more than threshold value.	Frequency less than threshold value.
<b>Database that collect allele frequency from healthy people</b>	-1	+1
<b>Database that collect allele frequency from cancer patient</b>	+1	-1

- **Group III: Supporting databases**

This type of database includes supportive data of gene by using each individual database output to determine if the gene is pathogenic or non-pathogenic. The result from this group can be varied ,for example, SIFT database has 2 types of output which are ‘D’(Deleterious) or ‘T’(Tolerated) and Mutation Taster database has 4 output values which are ‘A’(Disease causing automatic) , ‘D’(Disease causing) , ‘N’ (Polymorphism) , ‘P’ (Polymorphism automatic). From every output in this group we determined a universal scoring standard for all databases. The example scores for this group is shown in Table 5. It can be seen that the score in this group is smaller than the score given by the first two groups. The score ranges between -0.1 to 0.1.

**Table 5.** Example of gene score in supportive database.

Database	Deleterious	Probably Deleterious	Probably Benign	Benign
<b>SIFT</b>	+0.1	-	-	-0.1
<b>Polyphen2_HVAR</b>	+0.1	-	-	-0.1
<b>MutationTaster</b>	+0.1	+0.05	-0.05	-0.1
<b>VEST3</b>	+0.1	-	-	-0.1

- **VARCIN final score evaluation.**

When VARCIN analysis is complete, it will combine all score from the previous 3 groups and produce a final score which can be interpreted by using Table 6. After finishing calculate every record, VARCIN will rearrange every record from highest score record to lowest score record.

### 2.3 Prepare and produce a summary report

Lastly, VARCIN will produce a final summary file in PDF format. Because PDF files has a wide range of supported device and it's easy to generate. The report will be divided into simplified summary and full summary. The simplified summary contains 3-4 pages of report to keep it compact. It aims to be understandable in a short period of time. The user should be able to

use the result without additional calculation. The full summary section contains detailed information about every gene that show in the simplified summary section. Such as gene position, technical data, COSMIC data and Clinvar data.

### 3 Results

Because of our software has three outputs from each section includes annotation section result, filter section result and report output results.

#### 3.1 Annotation result

VARCIN depends its annotation step on ANNOVAR program which restrict input file structure, some part of the database and partly execution time. The annotation result is shown in Fig.1.

The ANNOVAR output consists of three parts, which are start-end chromosomes, reference gene and results from various databases that related to cancer patient's VCF input file. The result provided is difficult to read and hard to summarise without further information on each database and moreover, many rows are low in quality and not necessary for diagnosis.

The diagram illustrates the structure of ANNOVAR output. It is divided into three main sections:

- Start-End chromosomes:** This section contains a table of genomic coordinates for chromosomes chr2 and chr3, including columns for Chr, Start, and End.
- Reference gene information:** This section contains a table of reference gene variants, including columns for Ref, Alt, Func.refGene, Gene.refGene, and various quality metrics.
- Results from various databases:** This section contains a table of results from databases like ExAC, AFRE, and ClinVar, including columns for dbSNP ID, allele frequency, and p-value.

**Fig. 1** Three parts of ANNOVAR result, there are start-end chromosomes, reference gene and results from various databases that related to cancer patient VCF file.

#### 3.2 Filtering annotation result

After annotation, we filter unnecessary results out in order to increase an efficiency of our program. There are three factors to filter out which are quality, depth and normal gene databases.

The data in QUAL, DT column outputs are used for filtering data with low quality and low precision. Normal genes are removed to ensure only probably disease gene is shown in our program report.

After Annotation section executing and filtration result finished, the results from various database can now be divided into three groups for ranking scores. Each group has a different algorithm to score any variants, according to the section 2. When finished group dividing, we define each variant score depend on groups and rank each variant referred to Table 6.

**Table 6.** Meaning of variant total score

<b>Score</b>	<b>Description</b>
<b>More than 20 scores</b>	This gene has very high possibility of being pathogenic gene and should be closely monitored. Because overall score more than 20 meaning that both COSMIC and Clinvar have state that this gene is pathogenic gene.
<b>Lower than 20 but more than 5</b>	This gene has possibility of being pathogenic gene.
<b>Lower than 5 but more than -5</b>	This gene cannot be concluded that this gene is pathogenic gene or non-pathogenic gene. Maybe because of conflict of data or insufficient data provided.
<b>Lower than -5</b>	This gene has possibility of being non-pathogenic gene.

### 3.3 VARCIN result

The result of VARCIN is an interpretation report for each cancer patient. The interpretive report includes 5 result sections.

- **Information section.**

In this section, personal information and clinical information for each patient are displayed for clinicians. More information defines here and consists of patient name, DOB, Accession ID, Family, Gender, Specimen, Referring Physician, Race, Received and Hospital. The example result of this section is shown in Fig. 2.

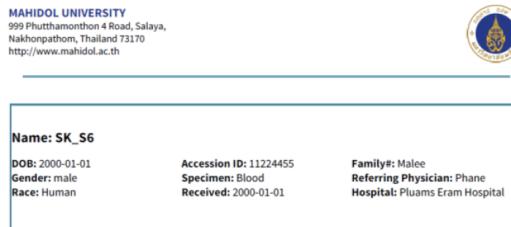


Fig 2 Example of Patient and Clinical information

- **Section A: Clinically actionable mutations section**

This section shows the information regarding variants is the certain cause of cancer, obtained from “mycancergenome.org” and “pct.mdanderson.org”. If our program detects the variant information under this condition, all of them must show in this section. The example result of clinically actionable section is shown in Fig.3. In this example the information of the gene ‘FLT3’ is found in both “mycancergenome.org” and “pct.mdanderson.org”, the results are shown with the detail referenced back to the websites. The useful information from “mycancergenome.org” consists of gene properties, frequency of gene mutation and clinical trials. The essential records of “mycancergenome.org” are shown in Table7. For “pct.mdanderson.org”, the useful information consists of gene properties, frequency of gene mutation, treatment affection, specific medicine and clinical trials. The essential records in pct.mdanderson.org are shown in Table 8.

**CLINICAL GENOME REPORT**

**RESULT SUMMARY**

**A. CLINICALLY ACTIONABLE MUTATIONS**

This table represent gene that already researched and declared in database. Description of gene and appropriate medicine to cure are shown below.

Gene: 9	Score: 20.5						
Chromosome	Chr13						
Start Position	2852542						
End Position	2852642						
RefGene	FLT3						
Alternate Gene	A						
refGene AAChange	FLT3 NM_004115 Exon20c_G2963T>G2963V						
Part 1: General Information							
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%;">FLT3</td> <td style="width: 50%;">Mycancergenome</td> </tr> <tr> <td colspan="2">Pathogenic</td> </tr> <tr> <td colspan="2">More information available. Please see detailed information at: <a href="https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/">https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/</a></td> </tr> </table>		FLT3	Mycancergenome	Pathogenic		More information available. Please see detailed information at: <a href="https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/">https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/</a>	
FLT3	Mycancergenome						
Pathogenic							
More information available. Please see detailed information at: <a href="https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/">https://www.mycancergenome.org/content/disease/acute-myeloid-leukemia/flt3/280/</a>							
Part 2: Gene Information							
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%;">MDanderson</td> <td style="width: 50%;">Gene Data: <a href="https://pct.mdanderson.org/#/home/FLT3?section=Overview">https://pct.mdanderson.org/#/home/FLT3?section=Overview</a></td> </tr> </table>		MDanderson	Gene Data: <a href="https://pct.mdanderson.org/#/home/FLT3?section=Overview">https://pct.mdanderson.org/#/home/FLT3?section=Overview</a>				
MDanderson	Gene Data: <a href="https://pct.mdanderson.org/#/home/FLT3?section=Overview">https://pct.mdanderson.org/#/home/FLT3?section=Overview</a>						

Fig. 3 Example of clinically actionable mutations result

**Table 7.**Essential records in “mycancergenome.org”

Record	Description
<b>1. Frequency of mutation variant</b>	Frequency of variant in a sample population.
<b>2. Implications for Targeted Therapeutics</b>	Useful information for targeted therapy treatment.
<b>3. Clinical Trials</b>	Clinical trials are used to determine whether new biomedical or behavioral interventions are safe, efficacious, and effective.

**Table 8.**Essential records in “pct.mdanderson.org”

Record	Description
<b>1. Treatment affection</b>	An unnatural form of behavior that is meant to effect other treatments.
<b>2. Specific medicine</b>	Specific medicine of each disease.
<b>3. Clinical Trials</b>	Clinical trials are used to determine whether new biomedical or behavioral interventions are safe, efficacious, and effective (illiam C. Shiel Jr., 2016)

- **Section B: High scores mutations section.**

This section represents genes with high circumstance become a mutation gene or a cause of cancer. Ten variants with the highest scores will show in this section. The example result of this section is shown in a Fig. 4. In this example, the variant ‘FLT3’ is again shown with the score 20.5 which can be confirmed that this gene is certainly pathogenic, supporting with the detailed found in section A.

#### B. HIGH SCORE MUTATIONS

This table represent genes that are reported with high score in every group. Genes in this group have a chance of being driver mutation or passenger mutation but still don't have a specific medicine to use with.

Reference Number	Gene Name	COSMIC	CLINVAR	Score
10	FLT3	208	Pathogenic	<b>20.5</b>
3	GATA2	1	-	<b>10.5</b>
27	BCORL1	11	-	<b>4</b>
14	ASXL1	3	-	<b>4</b>
15	ASXL1	1	-	<b>3</b>
24	ATRX	2	-	<b>2</b>
2	DNMT3A	-	-	<b>1.5</b>
11	IDH2	1	-	<b>1</b>
28	BCORL1	-	-	<b>0</b>
20	ASXL1	-	-	<b>0</b>

**Fig. 4** Example of high scores mutations section result

- **Section C: Possible new mutation section.**

This section represents gene that can a cause of cancer because it has not been confirmed by mycancergenome.org and pct.mdanderson.org but they have scores in the VARCIN program.

Ten variants with high scores are listed in this section. The example of possible new mutation section shows in Fig.5. The record found in this example are found to be ‘Deleterious’ in some databases but the score is not high enough and they have not been found to be non-pathogenic in any databases.

### C. POSSIBLE NEW MUTATIONS

This table represent gene that are reported with high score in 3rd group(supportive) database. The gene with high 3rd group score and not presented in B table will be shown here. Overall score must be at least 0 to be in this group.

Reference Number	Group 3: Deleterious	Group 3: Benign	Another Group Score	Overall Score
18	BCORL1	-	-	0
16	STAG2	-	-	0
12	ASXL1	-	-	0
11	ASXL1	-	-	0
5	IDH2	-	-	0
13	ASXL1	-	-	0

Fig. 5.Example of possible new mutations section result

- Gene reference information section.

This section contains essential data as references for three sections, includes of clinically actionable mutations section, high scores mutations section and possible new mutation section. The result is shown in Fig.6. In this example reference number 10 is shown in Fig. 4 as ‘FLT3’.The detail of ‘FLT3’ can be found in the fourth line in Fig.6 for the detail of that particular variant.

### GENE REFERENCE

This table contains essential data as a reference for table A, B and C

Reference Number	Chromosome Number	Start	End	Reference	Alternate
26	chrX	129156884	129156884	A	G
1	chr2	25457242	25457242	C	T
25	chrX	123184971	123184971	C	T
10	chr13	28592642	28592642	C	A
3	chr3	128202801	128202801	G	A
27	chrX	129190011	129190011	C	-
14	chr20	31022442	31022442	G	-
15	chr20	31022441	31022441	-	G
24	chrX	76920173	76920173	T	-
2	chr2	25469149	25469149	T	C
11	chr15	90631918	90631918	C	-
28	chrX	129190010	129190010	-	C

Fig. 6.Example of gene reference information section

- Full interpretation report.

This section contains all useful data for interpretation. Full report contains each variant information and score, dividing into three parts, which are general information and information found in the Cosmic and Clinvar databases. The example of full report showed in Fig.7. In this example Gene 10, with the most likely pathogenic is shown in detail in Fig.4 and the information obtained by each database.

Gene: 10	Score: 20.5
<b>Part 1: General Information</b>	
Chromosome	chr13
Start Position	28592642
End Position	28592642
Reference Gene	C
Alternate Gene	A
refGene AAChange	FLT3:NM_004119:exon20:c.G2503T;p.D835Y
<b>Part 2: COSMIC</b>	
ID 1	COSM783
Sample ID	998031
Primary Site	haematopoietic_and_lymphoid_tissue
Primary histology	haematopoietic_neoplasm
Subtype 1	acute_myeloid_leukaemia
Subtype 2	NS
Subtype 3	NS
Mutation CDS	c.2503G>T
Mutation AA	p.D835Y
Mutation Description	Substitution - Missense
<b>Part 3: Clinvar</b>	
Conclusion	Pathogenic
Disease Name	AML - Acute myeloid leukemia
Accession and version	RCV000017665.3
Database Name	GeneReviews;MedGen;OMIM;Orphanet;SNOMED_CT
Database ID	NBK47457;C0023467;601626;ORPHAS19:91861009

**Fig. 7** Example of full interpretation report

### 3.4 Evaluation of VARCIN software

In this validation result we use 20 different persons, 10 of which are cancer patients and the other 10 are those without cancer. The highest score variants in each patient are listed in Table 9. The patients with cancer tend to have a higher score than those without cancer.

Table 10 provides the result of using the score 10 as a threshold to differentiate between cancer and non-cancer patient, only 50% of the cancer can be obtained from the whole patients. No normal people are found in this group since to be categorized as a cancer patient using the score of 10, it has to be more certain the patient is having cancer.

To incorporate more supporting databases, the more flexible score is used in Table 11. The score 5, instead of 10, is used as a threshold for differentiate patients. The result gives 80% in accuracy, sensitivity and specificity. The results have shown that the supporting database can help find more cancer patient easily but with an expense of more normal patient being misdiagnosed. However, the fewer score the result gives, the less certain the diagnosis will be. More information will be needed to facilitate the decision making.

The report generated by VARCIN does not give only the scores to evaluate a patient, it also provide further information for clinicians. The software cannot use purely the score to classify two types of patients but rather than to assist and to guide for a fast diagnosing process toward patients in need.

Table 9: Result of the highest score variant from ten different people with and without cancer (Record 1-10 : with cancer , Record 11-20: without cancer)

Rec. 1	Rec. 2	Rec. 3	Rec. 4	Rec. 5	Rec. 6	Rec. 7	Rec. 8	Rec. 9	Rec.10
20.5	4	15	7.5	14.5	17.5	11.5	9.5	9.5	4
Rec.11	Rec.12	Rec.13	Rec.14	Rec.15	Rec.16	Rec.17	Rec.18	Rec.19	Rec.20
2	7.5	9.5	0	0	0	0	0	0.5	0.5

Table 10: The confusion Matrix from VARCIN with score 10 as a threshold

		Real Data	
		Disease	Non disease
Test Result	Certain Pathogenic (score >10)	50%	0%
	Not certain Pathogenic (Score<10)	50%	100%
	Total	50%	100%

Table 11: The confusion Matrix from VARCIN with score 5 as a threshold

		Real Data	
		Disease	Non disease
Test Result	Pathogenic (score >5)	80%	20%
	Non Pathogenic (Score<5)	20%	80%
	Total	80%	80%

It's possible there will be only Cosmic or Clinvar in prediction process. But the result score will be lower than variant that predicted with both Cosmic and Clinvar.

#### 4 Conclusion

In conclusion, VARCIN software is able to produce an innovative scoring system with a summary report in order to support clinicians' decision makings. The software is linked to many databases for better and more up-to-date references. It is also convenient to use, comparing to the traditional method, which gives scattered results in many different formats that are difficult to comprehend. However, the only type of the input file that the software can process must be in the format of VCF. The processing time also depends heavily on the size of the input file.

#### References

(ASCO), A. S. (2015, August). *Side Effects of Chemothe-rapy*. Retrieved 2016, from cancer.net: <http://www.cancer.net>

- Abramson, R. (2016). *Overview of Targeted Therapies for Cancer*. Retrieved 2016, from www.mycancergenome: <https://www.mycancergenome.org>
- EMBL-EBI. (2016). *IGSR: The International Genome Sample Resource*. Retrieved 2016, from 1000Gene: <http://www.1000genomes.org>
- Forbes, S. (2014). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, doi: 10.1093/nar/gku1075.
- William C. Shiel Jr., M. F. (2016, September). *Clinical Research and Clinical Trials*. Retrieved 2016, from medicinenet: [http://www.medicinenet.com/clinical\\_trials/article.htm](http://www.medicinenet.com/clinical_trials/article.htm)
- J. Ferlay, I. E. (2012). Cancer incidence and mortality worldwide: sources, methods and major patterns. *International Journal of Cancer*.
- Lalkhen, G. (2008). Clinical tests: sensitivity and specificity. *A. MB ChB FRCA*., 221-223.
- Landrum MJ, L. J.-S. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*, 862-868.
- Pruthi, S., e. a. (2015, June). *Personalized medicine and pharmacogenomics*. Retrieved 2016, from Mayo clinic: <http://www.mayoclinic.org/>
- Therapy, h. K. (2015). *Personalized Cancer Therapy*. Retrieved 2016, from EMBL-EBI. IGSR: The Interna-tional Genome Sample Resource.: <https://pct.mdanderson.org>.
- Wang K, L. M. (2016, February). *Annovar Documentation*. Retrieved from <http://doc-openbio.readthedocs.io>
- Wikidata. (n.d.). *COSMIC cancer database*. Retrieved May 2016, from [http://en.wikipedia.org/wiki/COSMIC\\_cancer\\_database](http://en.wikipedia.org/wiki/COSMIC_cancer_database)
- Wikidata. (2016). *Variant call format*. Retrieved April 2016, from [http://en.wikipedia.org/wiki/Variant\\_Call\\_Format](http://en.wikipedia.org/wiki/Variant_Call_Format)

## Medicine Recognition Using Intrinsic Geometric Property from Pill Image

Md. Zakir Hossan, Tanjina Piash Proma, M. Ashraful Amin

Computer Vision and Cybernetics Group, Department of Computer Science and Engineering,  
Independent University, Bangladesh, Bashundhara R/A, Dhaka, Bangladesh

[zshujon@gmail.com](mailto:zshujon@gmail.com), [tanjinaproma@gmail.com](mailto:tanjinaproma@gmail.com),  
[aminmmdashraful@iub.edu.bd](mailto:aminmmdashraful@iub.edu.bd)

**Abstract.** It is often the case that prescription pills do not come with blister or alu alu packaging where the identity of the pill is available, rather it comes in air-tight plastic bottles or labeled Ziploc bags. Problem with such bottle or pack is that, if the label is removed then it becomes difficult to tell what the pill is. Moreover, there is the issue of visually impaired people having difficulty identifying pills outside the pack. There are such many scenarios where it is good to have an automated pill recognition system. Due to the large variety of size, shape, color, texture it is a difficult task for human to tell about the identity of any individual medical pill. To localize a pill from a given dataset using computer vision techniques requires multiple steps. This paper will describe how to split a dataset according to the shape of pill. To find the shape information we used intrinsic geometric properties such as: eccentricity, extent and narrowness of pill which can be extracted from image using carefully selected image processing techniques. Reference values of discriminative parameters are determined using ‘RxIMAGE’, National Library of Medicine, USA database. The overall shape discrimination accuracy of the proposed system is 93.75%.

**Keywords:** Medical Imaging, Pill Image, Eccentricity, Extent, Narrowness

### 1 Introduction

Prescription pills are available in blister, alu-alu or container (bottle) pack and Ziploc pack. When a medical pill is out of its pack, it is almost an impossible task to recognize. It is also possible that in any way the label can be damaged. In many cases someone has to take two or more pill at a time, so it is hard to find the correct pill from several where the label is damaged or all of them are out of designated containers. Older people will face even more difficulty identifying pills out of its container.

For a visually impaired person this problem is even worse. As the printed labels are of no help to them. Though sometimes they can identify some pills by touching the embossed imprints on the surface of a pill. However, this works with some tablet form of pills, as in the capsule form of pill it is not possible to emboss any text. Moreover, sometimes the pills may not even have any imprint on them. However, sometimes here may be some imprint on the surface of a capsule but that is not sensible by touching. All of these issues may leads to wrong medication. Which may cause an unwanted serious health hazard.

Automated recognition of medicine is relatively a new concept though there exists a few work on this issue. Lee et al. [8] developed an application that is able to automatically identify illicit drugs. Hartl et al. [4], and Hartl [5] in their work tried to recognize medical pills using mobile device. However, their proposed method is limited to find shape and color of the pills.

In this paper we propose a method to provide an initial sorting of pill images based on proper measurement of their intrinsic geometric parameters values.

## 2 Proposed Method

The usual approaches to find features in image for object recognition are the use of, Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Pyramid Histogram of visual Words (PHOW) [6,7, 9-11].

Above mentioned feature descriptors perform excellent with objects that contains variations within the object under consideration. However, medical pill does not contain enough texture or corner features thus SIFT or SURF cannot extract enough discriminating features to separate pills from each other. The most dominant features for pills are size, shape and color. Hagedoorn [2], and Veltkamp and Hagedoorn [3] in their work provide a detailed survey about shape recognition techniques or algorithms. Such as tree pruning, Hough transform, Fourier descriptor, statistics, wavelength transform, deformable templets, curvature scale space, relaxation labeling, and neural network.

Because only features available for discrimination are shape, size and color; the proposed method intends to extract values of intrinsic geometric property of medical pills from image by using image processing techniques for a set of training image and then determine a threshold of those parameters and use it on differentiate pills.

### 2.1 The NLM Dataset

It is common practice to use standard datasets [1, 12] for developing and testing a new idea. On January 2016, NLM open a challenge under a federal notice “Pill Image Recognition Challenge [1]”. NLM’s purpose of this Challenge is to find a set of algorithm and software that can rank an input pill according to the similarity to images of unknown prescription pills to known prescription pill images in the NLM RxIMAGE dataset. There is two image-set one for consumer quality image and other for reference image and a ground truth table.

The RxIMAGE dataset contains 5000 images of 1000 different pills in the consumer quality image-set. To mimic the quality of image in consumer level the photos were taken with digital cameras built in to mobile devices. Fig. 1 shows some sample images from the database.



**Fig. 1.** Sample images from the reference set of the NLM RxIMAGE dataset

## 2.2 Image Processing Steps

Once the images are collected next task is to measure the geometric properties of the main object (pill) in an image. However, before we can do so, series of image processing steps are required to acquire the shape of the pill in the image. Consider the image in Fig 2(a). It is not possible to measure the intrinsic parameters values directly.



**Fig. 2.** (a) 24-bit Color pill image from the reference database, (b) 8-bit Gray pill image for the image from color image, (c) Detected Edges of the pill image for the Gray image, (d) Edge image after application of closing morphological operator, (e) Image after application of fill morphological operation, (f) Segmented pill region from pill image.

**Converting RGB image to Grayscale image:** Every reference image in the database is a 1600 X 2400 X 3 matrix. For further processing the color image is converted into a 1600 X 2400 8-bit grayscale image (Fig. 2(b)).

**Edge Detection:** Once the images are converted into an 8-bit gray image Canny edge detection algorithm is applied on it to acquire edge in the image (Fig. 2(c)).

**Closing Morphological Operation:** To segment out the pill from image it is necessary to find the boundary of pill. Morphological closing operation is used with disk shape structuring element to connect any edge pixel that is missed in the Canny edge detector during edge filtering process (Fig. 2(d)).

**Fill Morphological Operation:** Note that the image in Fig. 2(d) contains many smaller components, however we are interested in a single object representing the pill. Thus we apply fill morphological operator to finally acquire a single object (Fig. 2(e)).

**Detecting the Pill in the Image:** Now apply boundary enclose on the whole image the pill will be detected in the image (Fig. 2(f)).

## 2.3 Metric for Shape Measurement

Now that we have the pill image in more suitable form for measuring geometric parameter values. Depending upon the types of measurements used in image processing or analysis it is possible to split and categorize any dataset in various way. One can categorize the measurement types based on scale. Roundness, which can express the radius of curvature of the object corners in the next smaller scale measurement. The types of measurement can also be categorized based on the assumption level and the degree to which the results are calculated. In this case linear results such as area, perimeter can be calculated from the pixel map of the image. By these other results such as spherical equivalent volume and the circular equivalent diameter are calculated.

Finally, using several relations and ratios of the mentioned factors, metrics such as the aspect ratio, circularity, convexity, solidity, spherical equivalent volume, eccentricity, extent, elongation, convex hull area etc. may be calculated.

To categorize the RxIMAGE dataset we used Eccentricity, Extent and Narrowness (using aspect ratio) to separate the dataset into four categories. These are, Circular, Oval, Oblong, and Special (Fig 1).

**Eccentricity:** The ratio of the distance between two focal of the ellipse and its major axis length is known as eccentricity. Fig. 3(left) illustrates, how eccentricity can be a measure of shape. Its value is between 0 and 1. The shape with 0 (zero) eccentricity is actually a circle and eccentricity 1 (means) it's a line or line segment. It is a rotation and scale invariant property.

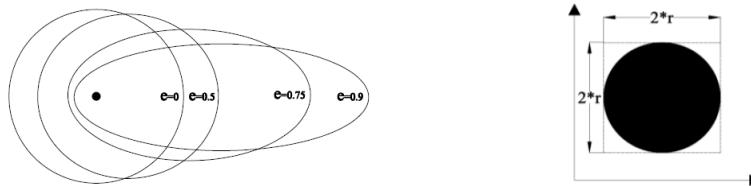


Fig. 3. (left) A perfect circle has an eccentricity of 0, while an oval or ellipse has 1, (right) Dimension of bounding box for a circle

**Extent:** It is the ratio of the area of the region to the total area of its bounding box (bounding box: the smallest rectangle containing the region) Fig. 3(right) illustrate and Eq. 1 gives the formula to measure extent of an image object. Though it scale invariant, however not rotation invariant. Thus we have to rotate the image according to its major axis to measure the correct value of extent. In theory if the extent is 0.7854 it's a circular shaped object.

$$\text{Extent} = \frac{\text{Object Area}}{\text{Bounding Box Area}} \quad (1)$$

**Narrowness:** It is defined as the absolute difference of aspect ratio and inverse of aspect ratio. The ratio between major axis and minor axis length is the measure aspect ratio (Eq. 2). If the narrowness is 0 (zero) the shape is circular. A higher value of narrowness (Eq. 3) will means the object is narrower.

$$\text{aspectRatio} = \frac{\text{majorAxisLength}}{\text{minorAxisLength}} \quad (2)$$

$$\text{narrowness} = \text{abs}\left(\text{aspectRatio} - \frac{1}{\text{aspectRatio}}\right) \quad (3)$$

### 3 Experimental Results and Discussion

We did all our experiments using the reference image-set from the RxIMAGE dataset. Table I provides distribution of data according to class.

TABLE I. Distribution of training and testing image

Type	Training Set	Testing Set
Circular	814	90
Oval	188	20
Oblong	722	80
Special	78	8
Sub-total	1802	198
Total		2000

For all the images of the train dataset we had calculated the values for Eccentricity, Extent and Narrowness and empirically measured the minimum and maximum values for all three features such that we can differentiate four shapes from each other. The maximum and minimum values for three features is given in Table II. Here note that, it is possible to find in theory one precise value for each of the features, however in reality the value is rather a range not a single floating point number. We believe this is happening because, during the image processing steps, information loss occurs and shapes does not appear as strict and structured they should be. Moreover, the shapes of the pills are most of the time circle like, oval like, oblong like, triangle like, trapezoid like not exactly the geometric shape they are supposed to be.

**TABLE II.** Shapes vs minimum and maximum value of the features

		Circular	Oval	Oblong	Special
<b>Eccentricity</b>	Min	0.0107	0.7212	0.7637	0.0298
	Max	0.2104	0.8938	0.9496	0.8522
<b>Extent</b>	Min	0.7713	0.7598	0.8003	0.6050
	Max	0.7935	0.8013	0.9732	0.9439
<b>Narrowness</b>	Min	0.00011471	0.7509	0.9034	0.0008874
	Max	0.0453	1.7818	2.8779	1.3683

Based on the minimum and maximum values of all three feature values of Table II test dataset was tested. Value of Table II can group the test dataset into four shape category with 93.75% accuracy. Detailed result of this testing is given in Table III.

**TABLE III.** Detailed test results for all shapes of the test dataset

	In Test Image Set	Correctly Identified	Accuracy
<b>Circular</b>	90	90	100%
<b>Oval</b>	20	20	100%
<b>Oblong</b>	80	80	100%
<b>Special</b>	8	6	75%
<b>Total</b>	196	196	99%
<b>Average Accuracy</b>			93.75%



**Fig. 4** Two pill images that were classified as oval shaped image instead of being classified as special type

Here note that, except the special shaped pills all others are correctly grouped according to their shapes. The two pills that are not correctly identified are classified as Ovals. In Fig 4 we are providing the pill images that are not correctly identified. It can be said that this particular morph is from an oval shaped object roughly we may call it oval. Intrinsic value suggest that it is an oval however visually they are not.

#### 4 Conclusion

In this paper we have proposed an intrinsic geometric feature based approach to discriminate pill images into four categories: Circular shaped, oval shaped, oblong shaped and special shaped. The proposed method can discriminate them with 93.75% accuracy. However, it might appear in mind that size and color should be more dominant feature to discriminate prescription pills from their images. However, taking size as a feature to discriminate pills from images is a difficult one, as size is not scale invariant and photo is taken from two different height, so pills will have different size in the image. The goal of this work is to provide an initial screening of the pill images into smaller groups and then next we are working on how to farther split the dataset according to the color of the pill. Moreover, we intend to incorporate OCR techniques to farther refine the search. Such that we can determine the text or symbol inside the interior of a pill, and that will help us more to accurately identify what kind of medicinal pill is that.

#### References

1. Pill Image Recognition Challenge by National Library of Medicine (NLM) <http://pir.nlm.nih.gov/challenge/>.
2. M. Hagedoorn, “Pattern Matching Using Similarity Measures”, PhD thesis, Universiteit Utrecht, 2000.
3. R. C. Veltkamp, and M. Hagedoorn, “State of the Art in Shape Matching”, Technical Report, Utrecht, 1999.
4. A. Hartl, C. Arth, and D. Schmalstieg “Instant Medical Pill Recognition on Mobile Phone”, IASTED International Conference on Computer Vision, 2011.
5. A. Hartl “Computer-Vision based Pharmaceutical Pill Recognition on Mobile Phones”, CESCG, 2010.
6. D. G. Lowe. Distinctive Image Features from Scale- Invariant Keypoints. IJCV, 2004.
7. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). CVIU, 2008.
8. Y. B. Lee, U. Park, and A. K. Jain, “PILL-ID: Matching and Retrieval of Drug Pill Imprint Images”, ICPR, 2010.
9. S. Saha, A. Mahmud, A. A. Ali, M. A. Amin, “Classifying Digital X-Ray Images into Different Human Body Parts”, ICIEV, 2016.
10. M. M. Rahman, B. Poon, M. A. Amin, H. Yan, “Recognizing Bangladeshi Currency for Visually Impaired”, ICMLC, 2014.
11. E. Hossain, S. M. S. Alam, A. A. Ali, M. A. Amin, “Fish Activity Tracking and Species Identification in Underwater Video” ICIEV, 2016.
12. M. A. Amin, and M. K. Mohammed, “Overview of the ImageCLEF 2015 medical clustering task”, ImageCLEF2015, 2015.

## A Regression-based SVD Parallelization using Overlapping Folds for Textual Data

Uraiwan Buatoom<sup>1</sup>, Thanaruk Theeramunkong<sup>1</sup>, and Waree Kongprewechnon<sup>1</sup>

<sup>1</sup> School of Information, Communication and Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand  
 uraiwan.buatoom@student.siit.tu.ac.th, uraiwanb31@gmail.com  
 thanaruk@siit.tu.ac.th, waree@siit.tu.ac.th

**Abstract.** One of the most difficult issues in text mining is high dimensionality caused by a large number of features (keywords). While various multivariate analyses, such as PCA and SVD (in information retrieval, called LSI), are developed to solve this curse of high dimensionality, they are computationally costly. This paper investigates a regression-based reconstruction method that enables parallelization of PCA/SVD by decomposing a document-term matrix into a set of sub-matrices with consideration of overlapped terms, and then to re-assemble using regression technique. To evaluate our method, we utilize two text datasets in the UCI Machine Learning Repository, called “Bag of Words” and “Reuter 50 50”. To measure the closeness between two documents, cosine similarity is applied while the accuracy is measured in the form of rank order mismatch. Finally, the result shows that, the matrices decomposition and re-assembly can preserve the quality of relation/representation.

**Keywords:** Decomposition, Re-assembly, Sub-matrix, SVD, Regression, LSI

### 1 Introduction

One of the most important challenges in text mining is how to handle large-scale textual datasets that usually hold characteristics of sparseness and high dimensionality with irrelevant, redundant, and noisy features, resulting in low performance but high computational cost. Towards the solution, recent works have proposed a number of data reduction [1] and data transformation methods, including feature selection [2], feature extraction [3], and dimensionality reduction [4].

Even now, it is well-known that Singular Value Decomposition (SVD) is a popular tool for feature extraction and dimensionality reduction, as for an analysis of multivariate data. SVD can be applied to a lot of fields that share the purpose of reducing the number of features in the dataset by selecting a few singular values that may preserve the original spectrum of the values. In [5], Wall et al. showed SVD can detect and extract small signals from noisy data. However, as a side effect, it is well-known that data reduction triggers a lower performance. Although in the past there have been a large number of works on how to reduce the dimension, most of them still suffered

from high computational cost [6]. In [7], Gao, J. and J. Zhang reported that SVD becomes inefficient when the dataset is large.

With a specific task, such as PCA on image data where spatial dependency are essential factors in correlation calculation, Segmented-PCA (S-PCA) [11] and Folded-PCA (F-PCA) [12] were introduced and were shown to improve time complexity in feature extraction and data reduction by segmenting a data record into equal-size partition by choosing and accumulating portions in the main diagonal of the original covariance matrix. However, S-PCA and F-PCA support for image datasets [8, 9], which every feature is not related all together like text datasets.

As an alternative to dimensionality reduction, partitioning the complete feature set into a number of overlapping subgroups (proxy matrices) and performing regression on each subgroup, can help us improve the accuracy. As an example of this approach, Janire Carlos et al. [10] proposed a method to solve the problem of segmentation and to combine a complete dataset from several sub-group data. They used the regression model to combine the subset data by using the remain point in the original data to predict remain output for making the new large amount of life cycle assessment data and then reduce the time consumption.

In this paper, we propose a method that is a combination of dimensionality reduction and overlapping subgroup regression. The method firstly decomposes a large matrix into a number of smaller overlapping matrices (i.e., sub-group data), and then applies SVD to reduce the dimensions before performing regression to combine the results of such sub-group data with the consideration of the correlation between words from SVD process. While performing SVD on the large matrix triggers high computational cost, whereas applying SVD on sub-group data can be done much faster. The complexity of the proposed method is proved to be lower than the original SVD.

To evaluate our method, the cosine similarities calculated from the original matrix and those calculated from the overlapping subgroup regression are compared. Besides the absolute values, we also consider the rank order mismatches between these two types of similarities. The result showed that our method has low similarity difference and rank order mismatches. Moreover, the complexity of the proposed method is analyzed and compared with the original SVD.

The remainder of the paper is organized as follows. Section 2 describes the mathematical formulation for motivation, including singular value decompose (SVD), linear regression (LR), text segmentation, and document-term matrix. In Section 3, the proposed method and simulation model is illustrated. Section 4 presents experiment settings and performance measures, i.e., cosine similarity and rank order mismatch. In Section 5, the experimental result and error analysis are discussed. Finally, a conclusion is given in Section 6.

## 2 Motivation

This section starts singular value decomposition, regression and document representation in the form of document-term matrix.

## 2.1 Singular Value Decomposition and Complexity Analysis

For data compression method, the Singular Value Decomposition (SVD) is a factorization of a real or complex matrix. It is the generalization of the Eigen-decomposition of a positive semi-definite normal matrix to any  $m \times n$  matrix via an extension of polar decomposition. Let  $X$  denote an  $m \times n$  matrix of real-valued data ( $m$  is the number of documents and  $n$  is the number of keywords), where  $m \leq n$  and  $r$  is the best initial good rank then  $1 \leq r \leq m$ . The equation of SVD of  $X$  can be represented as follows.

$$X = U \times \Sigma \times V^T \quad (1)$$

Here,  $U$  is a column-orthonormal  $m \times r$  matrix,  $\Sigma$  is a diagonal  $r \times r$  matrix with the

singular values. Each element at the diagonal  $r \times r$  matrix represents the eigenvalues  $\lambda_i$  are sorted in descending order.  $V$  is a column-orthonormal  $r \times n$  matrix with  $r$  is the rank of the matrix  $X$  [11] expressed as:

$$r \in [1, \min(m, n)] \quad (2)$$

The left-singular vectors of  $X$  represents the eigenvectors of  $XX^T$ , The right-singular vectors of  $X$  also represents the eigenvectors of  $X^TX$ , and the non-zero singular values of  $X$  represents the square roots of the non-zero eigenvalues of both  $X^TX$  and  $XX^T$  [12].

In a study of SVD, the results depict that the SVD can detect and remain the characterize structure of a new matrix from decompose, although the alternate position of columns. Finally, the performed result of this experiment is shown below as:

**Table 1.** The comparation of a new matrix from decompose with alternate column position

$X$	$\Sigma$	$V^T$
$X1 =$	$\Sigma 1 =$	$V^T 1$
1 2 3 5 7	21.9414 0 0	-0.3540 -0.4305 -0.5069 -0.3822 -0.5351
4 5 6 5 7	0 5.0571 0	-0.4703 -0.3847 -0.2992 0.4276 0.5987
7 8 9 5 7	0 0 0.0000	0.7632 -0.5953 -0.1679 0.1868 -0.0004
$X2 =$	$\Sigma 2 =$	$V^T 2$
5 7 1 2 3	21.9414 0 0	-0.3822 -0.5351 -0.3540 -0.4305 -0.5069
5 7 4 5 6	0 5.0571 0	0.4276 0.5987 -0.4703 -0.3847 -0.2992
5 7 7 8 9	0 0 0.0000	-0.8099 0.5745 0.0339 -0.0953 0.0615
$X3 =$	$\Sigma 3 =$	$V^T 3$
7 1 2 3 5	21.9414 0 0	-0.5351 -0.3540 -0.4305 -0.5069 -0.3822
7 4 5 6 5	0 5.0571 0	0.5987 -0.4703 -0.3847 -0.2992 0.4276
7 7 8 9 5	0 0 0.0000	-0.5793 -0.2022 -0.1306 0.3327 0.7041

From the real datasets, the number of real datasets in the documents ( $m$ ) is smaller than the number of keywords ( $n$ ). In-fact, for this research in SVD is focused only on

the test matrix  $V$  which represents the relation between keywords and same like that in a study of synonymy and polysemy [5]. The SVD control methodology is to detect the type of words.  $V$  is denoted as Eigen value of keyword. The synonymy of word is defined as a strongly correlated with the same Eigen-key term

## 2.2 Linear Regression, Complexity Analysis and Error Estimation

The numeric values (continuous values) can be applied to the mathematics linear regression to analyze weights for combining attributes. The prediction uses the attributes, which have a relation together to transform the data from one side to another side by calculating the new values from weights [10]. The mathematic expression in this methodology represents the variable  $y$  is a model which related with independent variable  $x$  so we can define the linear regression as [13]:

$$y = w_0 + w_1 x \quad (3)$$

Here, we assume that  $y$  is a constant, and  $w_0$  and  $w_1$  are regression coefficients (weights). The regression coefficients can be expressed as:

$$\vec{w} = (A^T A)^{-1} A^T \vec{y} \quad (4)$$

To evaluate the result, it is possible to apply regressions for matching. The measurement of error data between original and predictive data can be expressed as:

$$\text{Percent of Error} = \frac{|\text{Measured Value}-\text{Actual Value}|}{\text{Actual Value}} \times 100\% \quad (5)$$

## 2.3 Document-Term Matrix Construction

To process textual data is necessary to translate a document into a computable form. The common form is a document-term matrix, where each column is a unigram keyword feature and each row corresponds to a document, represented by a vector of terms in the document, called document vector. While there are several possibilities in encoding (weighting) terms in the document vector, some popular weighting methods are TF, TFIDF [14], and binary weighting [9], such representations are needed in text classification, text clustering, and text summarization. As the most popular weighting applied in the field of information retrieval and text mining, term frequency-inverse document frequency (TFIDF) is used for representing the importance of a word in this work.

$$\text{TFIDF}(t) = \text{TF}(t, d) \times \text{IDF}(t) \quad (6)$$

The term frequency (for short, TF) shows how many times the term  $t$  appears in the document  $d$ . The higher term frequency the term has, the more the term contributes to the document in terms of semantics. The inverse document frequency (for short, IDF) is added into the weighting (Eq. 1) since there is a trend that a term appearing in many documents has less contribution to the semantics than a term that occurs in a few doc-

uments. Even there are several alternative forms of IDF, one of the mos equations is as follows.

$$IDF(t) = \log\left(1 + \frac{N}{n_t}\right) \quad (7)$$

where  $N$  represents the total number of documents and  $n_t$  is the number of documents that include the term  $t$ .

### 3 Proposed Method and Simulation Model

The design proposed methodology and simulation model is classified as follows:

#### 3.1 Overlapping Matrix Decomposition

To reduce the dimensionality, firstly we decompose a document-term matrix into a set of sub-matrices. The model represents  $M$  rows and  $N$  columns. Here,  $o$  represents the number of columns which is overlap between sub-matrices by  $o > M$ , as shown in Fig.1. and  $a$  represent the non-overlapping columns. Let  $A \in \mathbb{R}^{mxn}$  is a real matrix, then  $\text{Split}(A) = [A_1, A_2, A_3, \dots, A_P]$  where  $A_1 \cap A_2, A_2 \cap A_3, \dots, A_{P-1} \cap A_P \neq \emptyset$

The function for calculating the amount of segments is as follows:

$$\begin{aligned} N &= (2o + a) + (P - 1)(a + o) \\ &= 2o + a + Pa - a + Po - N = o + P(a + o) \\ P &= \frac{N-o}{a+o} \end{aligned} \quad (8)$$

From Fig1.  $a = n - 2o$  then  $o = \frac{(n-a)}{2}$  So  $P = \frac{N-o}{n-2o+o} = \frac{N-o}{n-o}$   
If represent  $r$  is percentage for overlap area between 2 sub-group, then

$$P = \frac{N-(r*n)}{n-(r*n)} = \frac{N-(r*n)}{n*(1-r)} \quad (9)$$

The width of range for segments equal as:

$$n = 2o + a \quad (10)$$

where  $P$  is the number of partitions of sub-matrices.

$o$  is the number of columns that is an overlapping region of sub-matrix.

$a$  is the number of columns that is a non-overlapping region of sub-matrix.

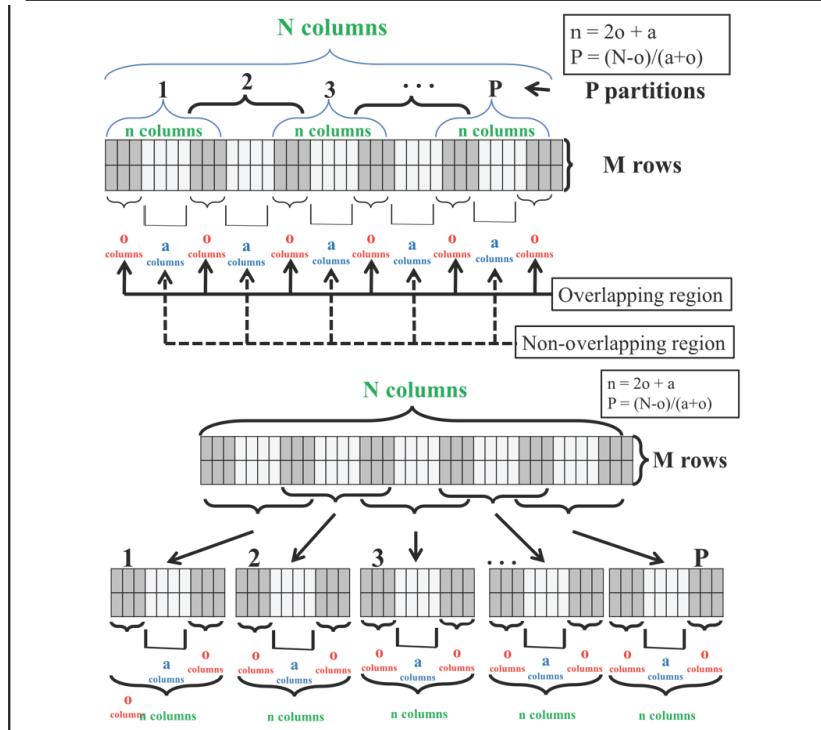
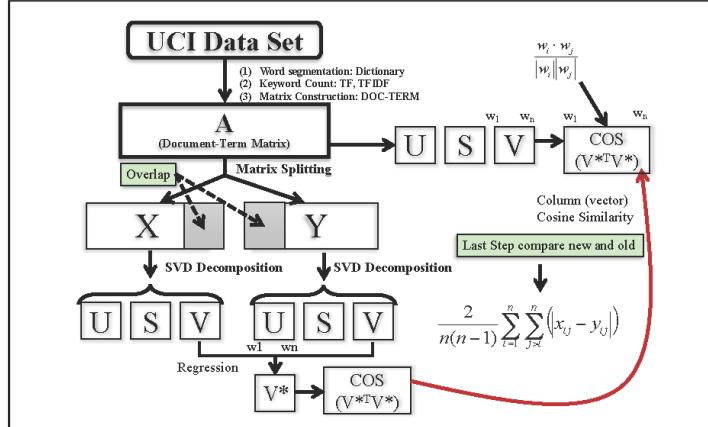


Fig. 1. The Initial Range for Sub-Matrices Model

### 3.2 Splitting matrix and Individual dimensionality reduction

In this work, we propose a method to reduce computational cost of SVD calculation on a large-scale matrix by first splitting the matrix into a number of overlapping subgroup matrices (proxy matrices), performing SVD on each subgroup matrix and then combining the result by regression techniques. This step we propose to find the optimize bound of small subgroup matrices by testing different percentage of overlapping region under control to cut rank of choosing the small diagonal entries of  $\Sigma$ , which consists of a descending order (mentioned in section 2.1). The difference cosine similarity is given for quality measurement of result between original methodology and re-assemble methodology. The complete model of modified feature extraction as shown in Fig. 2.



**Fig. 2.** The modified feature extraction implementation

#### 4 Experiment setting and performance measurement

The experiment setting and performance measurement consist of dataset and stimulate feature measurement, which is expressed as:

##### 4.1 Data Set

To evaluate our method, we utilize two text datasets in the UCI Machine Learning Repository[15, 16], namely “Bag of Words” (later, BOW) and “Reuter 50 50” (later, C50). Due to the sake of computational complexity, for the former dataset, we select the “NIPS full papers” subgroup, which is composed of 1,500 documents with 12,419 distinct words, and approximately 19,000,000 words in total. The latter dataset, i.e., C50, contains 2,500 documents. In the experiments, we select 1,500 from 2,500 documents due to the computational reason (14,284 distinct keywords, nearly 21,500,000 words in total.). For document-term representation, the TF-IDF weighing is used.

##### 4.2 Stimulate Feature Measurement Similarity between word

The new matrices V from SVD, focus on the relation of word view. The results from 2 groups estimate the efficiency by measuring the similarity in 2 ways and it is discussed as follows:

###### 4.2.1 Cosine Similarity

The Cosine Similarity is one of the popular measurement, which is used for measuring the similarity of vectors. This measurement proposed to measure the same composition of vectors by cosine of the angle. Garcia [17] has shown that the matrix of the

new vector coordinates, which was reduced dimensional space from SVD process for the closest of matrix, which has higher score of cosine similarity than other vectors. The estimated equation is follow as

$$\cos(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \cdot \|w_2\|} \frac{\sum_{t_i=t_j} w_i \cdot w_j}{\sqrt{\sum w_i^2 \sum w_j^2}} \quad (11)$$

#### 4.2.2 Rank Order Mismatch

The Rank order mismatch is the method to measure the order of ranking data. Pavan et al. shown to use this tool to measure the quality of order sequence of text structure and they described when data is most quality ranking then the value of rank will nearly be one [18]. The proposed methodology highlights the comparison of difference of order ranking between similarity keywords. The resulted cosine similarity measurement method is used as an input of this method. The quality of order ranking expresses rank equation as:

$$\begin{aligned} \text{Rank}(w_1, w_2) &= \frac{\text{Match}(w_1, w_2)}{\text{Match}(w_1, w_2) + \text{MisMatch}(w_1, w_2)} \\ &= \frac{\sum_{i=1}^n \sum_{j=i-1}^m \text{Match}(w_i, w_j)}{\sum_{i=1}^n \sum_{j=i-1}^m \text{Match}(w_i, w_j) + \sum_{i=1}^n \sum_{j=i-1}^m \text{MisMatch}(w_i, w_j)} \end{aligned} \quad (12)$$

Where,  $\text{MisMatch}(w_i, w_j)$  is the number of mismatched term in batch of  $w_1, w_2$ .  $\text{Match}(w_i, w_j)$  also represents the number of matching rank of both  $w_i$  and  $w_j$ . Moreover, Comparison loop shows the loop assigned remark score by considering value of between  $w_i$  and  $w_j$ . In case of  $w_i$ , it is greater than  $w_j$  then remark score is +0.5. In case of  $w_i$ , it is lesser than  $w_j$  then remark score is -0.5 and when the value is equal to assign remark the score is 0. Finally, the rank order mismatch model is shown below as

	W1	W2	W3	W4	W5
W1	1	0.2	0.4	0.3	0.5
W2	0.2	1	0.7	0.1	0.3
W3	0.4	0.7	1	0.4	0.5
W4	0.3	0.1	0.4	1	0.6
W5	0.5	0.3	0.5	0.6	1

+

	W1	W2	W3	W4	W5
W1	1	0.4	0.3	0.2	0.5
W2	0.4	1	0.7	0.5	0.2
W3	0.3	0.7	1	0.4	0.3
W4	0.2	0.5	0.4	1	0.5
W5	0.5	0.2	0.3	0.5	1

**Fig. 3a.** The  $v^*$  matrix from method 1

**Fig. 3b.** The  $v$  matrix from method 2

	Method 1	Method 2	Result of mismatch
W2w3	-0.5	+0.5	1
W2w4	-0.5	+0.5	1
W2w5	-0.5	-0.5	0
W3w4	+0.5	+0.5	0
W3w5	-0.5	-0.5	0
W4w5	-0.5	-0.5	0

**Fig. 3c.** The desired result of rank mismatch

## 5 Results and discussion

The proposed methodology is applied to regression to combine the sub-matrices. Comparing the results by different percentage of overlapping of dataset. We found that from table2 (row 1<sup>st</sup> and 2<sup>nd</sup>) that the error of predicted for mapping data is always less than 1% by using the both datasets. Fig. 4. Shows the Comparison performance of prediction for mapping data. Most of the results shown the trend error of data stability is in lowest range between 30% and 35% for overlapping region columns.

The methodology applied statistics to measure quality of the result two main topics. Which is about similarity of relation word between original matrices and new data from re-assembly matrices and compare reduce cost of time from stimulate.

### 5.1 Measuring Similarity of Relation of Words

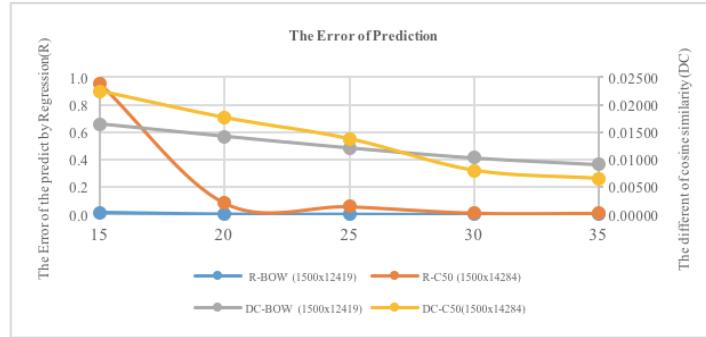
Table 1, the row 3<sup>rd</sup> and 4<sup>th</sup> elaborates the two groups of dataset for comparing the different percentage of overlapping data. These values are statistics of similarity cosine. We can infer that the performance of several datasets follows the degree of percentage of overlapping data. However, we could also infer that the different similarity of words between the lowest and highest range is not much different. So, the most of data in re-assemble matrix is closest to the original data matrix.

**Table 2.** The Error of Predict Regression and Different Similarity Cosine Results (no. of rows = 1500)

Measurement Method	The percentage of overlapping	15	20	25	30	35
Error Regression	R-BOW	0.01370	0.00380	0.00270	0.00083	0.00079
	R-C50	0.94980	0.08430	0.05790	0.01190	0.01010
Different Similarity Cosine	DC-BOW	0.01640	0.01420	0.01210	0.01030	0.00910
	DC-C50	0.02240	0.01760	0.01370	0.00800	0.00650

**Table 3.** The Result of Rank Order Mismatch

Dataset	Match( $w_i, w_j$ )	MisMatch( $w_i, w_j$ )	rank
BOW(overlapping 35%)	907,483,622,718	50,140,139,531	0.94764
C50(overlapping 35%)	1,384,715,601,633	72,383,611,191	0.95032



**Fig. 4.** The Error of Predict Regression and Different Similarity Cosine Results

Moreover, to get the best results, we choose the best result of different similarity cosine from both mentioned two groups to test the rank order mismatch by checking every column. Table 3 uses the same dataset of Table 2. Table 3 shows the performance in the view of how much performance for ranking order mismatch of data from similarity cosine between words. The best rank is C50 data set at 0.95032. The result approach 1 shows that the meaning order rank of data between two groups is nearly the same ranking. Finally, the results show that the rank is nearly 1.

## 5.2 Measuring Reduce Cost of Simulation

Table 4 highlights the three group type of data which is reformed by the different sub-matrices data. Table 4. also compare the result of the average SVD runtimes from five times. That shows the 3 sub-matrices group have the lowest time. At fig 5 also shows the trend of data from three groups have low running time followed by the amount of sub-groups data. Therefore, we can conclude that, if we reform data to many sub-groups then it ensures that the running time will decrease and the results will follow to calculate the BIG-O table.

Although, Fig. 5 depicts the graph in 3 groups (P), which is segmented in original matrix to 3 sub-matrices by overlapping 10% region, has sharply dropped of running time follow the size of data. Moreover, Table5 compares the performance of BIG-O. This also shows the performance of Covariance matrix and Eigen Problem follow the higher number of partitions and the less of percentage of overlapping. However, in mapping matrices by linear-regression had overhead, but it still has the lower cost than old-SVD.

**Table 4.** The performance of running time

Time running SVD (secs)	The size of Data (MxM) Segmented Sub-group(P) overlapping 10%					
	100	1,000	5,000	10,000	15,000	20,000
1 groups (P)	$4.89 \times 10^{-3}$	$7.32 \times 10^{-1}$	$7.35 \times 10$	$5.29 \times 10^2$	$1.95 \times 10^3$	$4.73 \times 10^3$
2 groups (P)	$3.37 \times 10^{-3}$	$2.74 \times 10^{-1}$	$2.71 \times 10$	$1.87 \times 10^2$	$6.37 \times 10^2$	$1.47 \times 10^3$
3 groups (P)	$2.56 \times 10^{-3}$	$1.46 \times 10^{-1}$	$1.23 \times 10$	$8.92 \times 10$	$1.87 \times 10^2$	$6.54 \times 10^2$

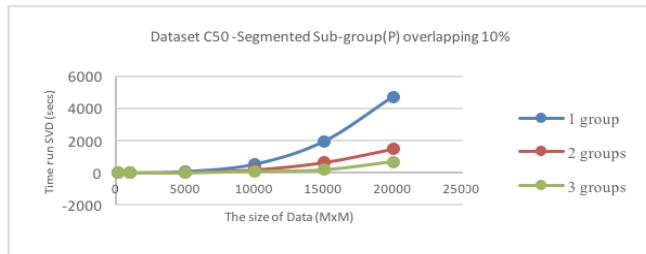


Fig. 5. Comparing Performance of SVD by Running Time

The proposed BIG-O between 2 methods is as follows:

**Table 5.** The BIG-O of method between SVD and Modified sub-matrices + SVD

Process	SVD (A)	Sub-matrices + SVD (B)	Ratio (A/B)
Covariance matrix size	$O(MN^2)$	$O(Pmn^2)$	$r^2P$
Eigen problem	$O(N^3)$	$O(Pn^3)$	$\left(\frac{r^2}{1-r}\right)P$
Mapping by linear regression	-	$O((P-1)(r^*n))$	-

where,  $M, m$  : The number of samples  
 $N, n$  : The dimensions of constructed matrix for computation.  
 $P$  : The number of times to mapping data  
 $o$  : The number of overlapping region (from section 3.1  $o = \frac{(n-a)}{2}$ )  
 $r$  : The percentage of overlapping region

## 6 Conclusion

Effect of data segment by using dimension reduction methodology, we can reduce the cost of time, then we can also decrease the size of matrix by re-assemble to sub-matrix. The proposed methodology, the segment sub-matrix, under the condition shows that, the set of column size of example data is bigger than the row data. The column size data represents the number of set example data for finding the weights of coefficient of linear regression. The linear regression can be used as a tool for combining matrix back to the same size of original matrix. The experiment results also claim that with the measurement of the different similarity cosine of word correlation, which is gained from SVD, between original matrix and sub-matrix. To check the better performance, the rank order mismatch is used, which has high value matching i.e. nearly equal to one.

This paper shows that, we still have a chance to re-build a new matrix which has less overlapping region under less timing computation and still preserve the quality of matrix. However, to reform the back size regression in this methodology still has the high BIG-O, so, we can also focus on the cost of predictive data for combining the less than linear regression. Finally, we can improve a number of group to segment, which could be studied and experiment in the future.

**Acknowledgement.** This work is financially funded and supported by Sirindhorn International Institute of Technology, Thammasat University and Burapha University.

## References

1. Chen, Y.H. and L. Ting-Chia. *Dimension reduction techniques for accessing Chinese readability*. in *Machine Learning and Cybernetics (ICMLC), 2014 International Conference on*. 2014.
2. Ketui, N. and T. Theeramunkong, *Effect of Weighting Factors and Unit-Selection Factors on Text Summarization*, in *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, D.-N. Pham and S.-B. Park, Editors. 2014, Springer International Publishing: Cham. p. 891-897.
3. He, Q. and X. Ding, *Sparse representation based on local time-frequency template matching for bearing transient fault feature extraction*. Journal of Sound and Vibration, 2016. **370**: p. 424-443.
4. Bharti, K.K. and P.K. Singh, *A three-stage unsupervised dimension reduction method for text clustering*. Journal of Computational Science, 2014. **5**(2): p. 156-169.
5. Wall, M.E., A. Rechtsteiner, and L.M. Rocha, *Singular Value Decomposition and Principal Component Analysis*, in *A Practical Approach to Microarray Data Analysis*, D.P. Berrar, W. Dubitzky, and M. Granzow, Editors. 2003, Springer US: Boston, MA. p. 91-109.
6. Jun, S., S.-S. Park, and D.-S. Jang, *Document clustering method using dimension reduction and support vector clustering to overcome sparseness*. Expert Systems with Applications, 2014. **41**(7): p. 3204-3212.
7. Gao, J. and J. Zhang, *Clustered SVD strategies in latent semantic indexing*. Information Processing & Management, 2005. **41**(5): p. 1051-1063.
8. Zabalza, J., et al., *Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing*. ISPRS Journal of Photogrammetry and Remote Sensing, 2014. **93**: p. 112-122.
9. Xiuping, J. and J.A. Richards, *Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification*. IEEE Transactions on Geoscience and Remote Sensing, 1999. **37**(1): p. 538-542.
10. Pascual-González, J., et al., *Combined use of MILP and multi-linear regression to simplify LCA studies*. Computers & Chemical Engineering, 2015. **82**: p. 34-43.
11. Qiao, H., *New SVD based initialization strategy for non-negative matrix factorization*. Pattern Recognition Letters, 2015. **63**: p. 71-77.
12. Shlens, J., *A tutorial on principal component analysis*. 2003.
13. Theeramunkong, T., *Introduction to concepts and techniques in data mining and application to text mining*. 2012.
14. Kittiphanthanabawon, N., T. Theeramunkong, and E. Nantajeewarawat, *News Relation Discovery Based on Association Rule Mining with Combining Factors*. IEICE Transactions, 2011. **94-D**: p. 404-415.
15. Lichman, M., *{UCI} Machine Learning Repository*. <http://archive.ics.uci.edu/ml>. 2013.
16. ZhiLiu, *{UCI} Machine Learning Repository*. [https://archive.ics.uci.edu/ml/datasets/Reuter\\_50\\_50](https://archive.ics.uci.edu/ml/datasets/Reuter_50_50). 2011.
17. Garcia, D.E., *Latent Semantic Indexing (LSI) A Fast Track Tutorial*. 2006.
18. Pavan Kumar, P., A. Agarwal, and C. Bhagvati, *A Structure Based Approach for Mathematical Expression Retrieval*, in *Multi-disciplinary Trends in Artificial Intelligence: 6th International Workshop, MIWAI 2012, Ho Chi Minh City, Vietnam, December 26-28, 2012. Proceedings*, C. Sombaththeera, et al., Editors. 2012, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 23-34.

## Virtual Reality System with Smartphone Application for Height Exposure

Suppanut Nateeraitaiwa<sup>1</sup>, Narit Hnoohom<sup>2</sup>

<sup>1,2</sup>Image, Information and Intelligence Laboratory, Department of Computer Engineering,  
Faculty of Engineering, Mahidol University, Nakorn Pathom, Thailand

<sup>1</sup>Suppanut.n@hotmail.com, <sup>2</sup>narit.hno@mahidol.ac.th

**Abstract.** One of the treatment methods for phobias is behavioral therapy by creating patient's fear environment for the patient to confront that situation. Virtual reality is one of the most interesting technologies for creating three-dimensional virtual environments. The virtual reality technology makes users feel like they are immersed in a virtual world. To accomplish creating fear environment, this paper presented a virtual reality system with a smartphone application for height exposure. The virtual reality system is simple, which consists of software and hardware. Users can use this system easily in their own home. With an evaluation on 20 participants, the impact of user involvement on realism and fear of heights was explored. Paired samples t-test showed that the user's influence on the height of the building was significant. Moreover, the user's influence on the realism and fear of heights when the sound is activated were significant.

**Keywords:** Height exposure, Virtual reality, Smartphone application, Virtual reality glasses, Remote controller.

### 1 Introduction

Virtual reality technology is creating three-dimensional virtual environments by computers. The technology makes users feel like they are immersed in a virtual world and that they can control their avatar by action in the real world [1]. Since the year 1990, virtual reality technology has become popular. It has been developed and used in various fields, particularly in entertainment, as well as a tool to help treat some diseases [2].

A phobia is a type of anxiety disorder, which has an impact on a patients' quality of life. Patients with phobias have excessive and irrational fear of a situation or a particular object. The fear happens often and patients cannot stop their fear. Patients will avoid behavior to confront that situation, making their life miserable [3]. The examples of phobias include acrophobia (fear of heights) [4], cynophobia (fear of dogs) [5], arachnophobia (fear of spiders) [6], and claustrophobia (fear of confined spaces) [7].

One of the treatment methods for phobias is behavioral therapy by creating a patient's fear environment for the patient to confront that situation. Virtual reality technology can create a virtual environment that makes users feel like they immersed in a virtual world. We can simulate a virtual environment that displays a patient's fear situation.

Virtual reality exposure therapy is provided to patients immersed in a computer generated virtual environment, either through the use of a head-mounted display (HMD) device or entry into computer-assisted virtual environments (CAVEs) where images are present all around. Computer will create and control that virtual environment to show the situation that patient fears. You can see that using virtual reality to create patients' fear situation is help patients can directly confront patients' fear situation without risks of doing the same in real life [8].

Hodges et al. [9] described a pilot study that uses a virtual reality technology for treating acrophobia. The result showed subject had to act in accordance with the situation show in the virtual world that subject facing and, subject has difference level of anxiety in difference situations. In terms of emotional processing theory, the fear structure of subjects showed clearly that the subjects' responses to facing situation and, the subjects' responses clearly that anxiety of subjects is reduced. The result support that the fear structure are changed when treated.

Haworth et al. [10] were proposed a virtual reality system for treating phobias: Acrophobia and Arachnophobia. This virtual reality exposure therapy (VRET) system is a low-cost, and readily available software and hardware components. The system using Kinect to track patient's body. However, the system must always be connected to internet for communicating between the clinician and the patients.

Schafer et al. [11] have presented a virtual reality technology to create virtual environment and player avatar for exploring the effective treatment of acrophobia. The selected 42 subjects divided into two groups. First group uses the system with avatar and second group uses the system only. The result showed significant differences in score of two groups.

This paper is organized as follows: first, a general introduction to the appearance of acrophobia and treatment guidelines using virtual reality technologies are briefly described. The hardware and software requirements for this research are given in Section 2. Next, Section 3 considers the method to develop the smartphone application. The experimental settings and results are shown in Section 4. Finally, this research is concluded in Section 5.

## 2 System

The system is composed of hardware and software components that work together to simulate a virtual environment. This system is a head-mounted display set up with a

smartphone to display a virtual environment. We simulate it and use a wireless remote controller for control of the avatar in the virtual world.

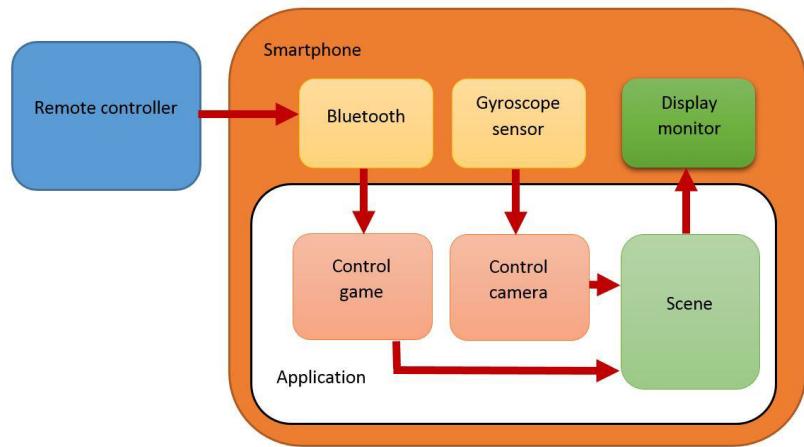


Fig. 1. Block diagram of system process

## 2.1 Hardware selection

The related hardware of the research is illustrated in Figure 1, contains main 4 devices: display monitor, gyroscope sensor, virtual reality glasses, and remote controller. The following sections present a brief summary of each device.

**Display monitor:** We use a smartphone which runs an android operating system for display. These days a smartphone is an important device. Almost everyone has their own smartphone and android is a popular smartphone operating system. On September 29, 2015, Google announced that there are currently 1.4 billion active android devices worldwide [12]. For developers, the Android software development kit (SDK) is an open source software. Android has the cardboard SDK for our integrated development environment (IDE) to easily adapt an existing 3D application for virtual reality (VR).

**Gyroscope sensor:** We use a gyroscope sensor in the smartphone. The gyroscope measures the rate of rotation in radian per second around a device's x, y, and z axes. Rotation is positive in the counter-clockwise direction; that is, an observer looking from some positive location on the x, y or z axes at a device positioned on the origin would report positive rotation if the device appeared to be rotating counter-clockwise. This is the standard mathematical definition of positive rotation and is not the same as the definition for roll that is used by the orientation sensor. This research uses the value from the gyroscope as input to control the camera view in the application.

**Virtual reality glasses:** A pair of virtual reality glasses is a simple head-mounted display that can allow a smartphone to display two images, one for the left eye and one for

the right. Virtual reality glasses contain two polarized lenses for adjusting the focus of the eyes. The Virtual reality glasses used in this research are a pair of 3D Shinecon Glasses, as shown in Figure 2.



**Fig. 2.** 3D VR Shinecon glasses

**Remote controller:** The basic controller devices are well compatible with an android smartphone. Communication via Bluetooth. Device is also simple and easy to use. Figure 3 shows a universal wireless remote controller for controlling the movement of an avatar in the virtual world.



**Fig. 3.** Universal wireless remote controller

## 2.2 Software selection

Unity 5.3.2 is used as an engine for developing our application. Unity can import Cardboard SDK for support to develop a virtual reality application and can build the application for use with many platforms of smartphones. This SDK makes it easy to:

- Build an existing 3D application for mobile application by using a display of the mobile application that does not deviate from the Unity 3D application.
- Adapt an existing Unity 3D application to a virtual reality application.

- We can control the camera viewing to be compatible with head tracking.

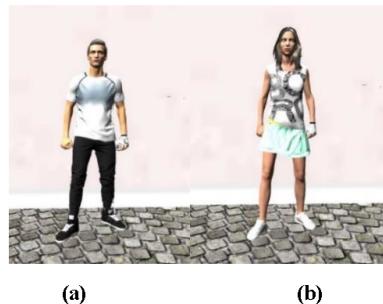
Unity is a game engine for developing virtual reality game applications that we can download from Play Store, for example RollerCoaster, Crazy Swing, and Zombie Run developed by FiBRUM, and VR Roller Coaster, VR Cave Flythrough, and VR Volcano Flythrough developed by frag. And there are some papers that use Unity for developing applications as well, for example a serious game with virtual reality for travel training for those with Autism Spectrum Disorder [13], design of a video game for rehabilitation using motion capture, EMG analysis and virtual reality [14], and Immersive VR for natural interaction with a haptic interface for Shape Rendering [15].

### 3 Propose Methods

#### 3.1 Model selection

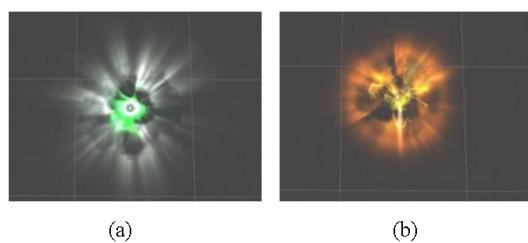
This section discusses the features of the models, which are used in the scene for making our scene more realistic. We select the following models.

- **User avatar.** First Person Lover published by ISBIT GAME [16, 17], shown in Figure 4.



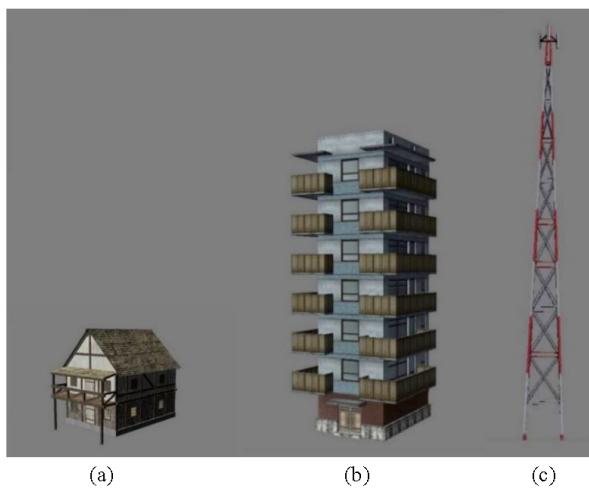
**Fig. 4.** (a) Male avatar model, (b) Female avatar model

- **Teleport Effect.** KY Magic Effect published by Kakky [18], shown in Figure 5.



**Fig. 5.** (a) Green teleport effect, (b) Orange teleport effect

- **Buildings for height exposure.** Medieval Buildings published by 7th Dimension [19], Block Building Pack published by CGY (Yemelyan K.) [20], Radio Tower - Low Poly published by VR [21], shown in Figure 6.

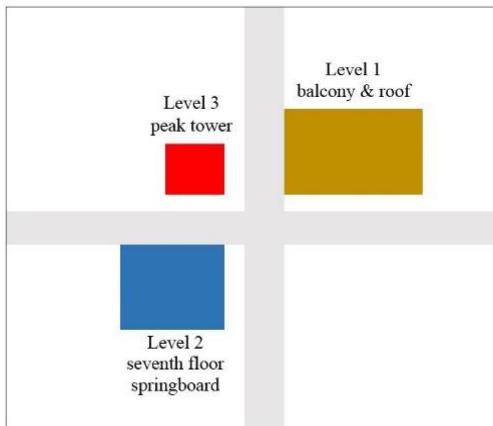


**Fig. 6.** (a) Building Level 1 balcony & roof, (b) Building Level 2 seventh floor springboard, (c) Building Level 3 peak tower

### 3.2 Map design

The objective of this research is to use a virtual reality application to simulate a virtual environment for users to confront several levels of height situations. The scene in the application simulates the user avatar in a virtual city. Users can control his/her avatar to move in a designated area. In the designated area, there are 3 buildings that users can move upward to the top of in order to confront different levels of heights. Users can move up a building by teleporting which shows a scene like a green effect. When users move into the green effect, the user's avatar will teleport to the top of this building at a fixed position. And users can teleport down by moving into an orange effect, then users will teleport to the start position. The three buildings are different structures as follows:

- Level 1 balcony & roof: This building is a two-story house that has a balcony, as shown in Figure 6(a).
- Level 2 seventh floor springboard: Figure 6(b) shows a high building, which has 7 floors. On top of the building there is an open terrace.
- Level 3 peak tower: This building is a radio tower that is very high, as shown in Figure 6(c).
- Other objects in the scene are intended to make the scene more realistic, for example other buildings, roads, trees, etc. Figure 7 shows a top view of the map.



**Fig. 7.** Top view of map

### 3.3 Avatar design

User avatars are created by the object character controller from Unity. The Avatar has a capsule shape that height of 6 Unity units (unit of Unity) and a radius of 1.5 Unity units. On top of the capsule there is a main camera for display of output. The radius of the capsule uses this value to always keep the main camera on a capsule collider. The character controller has a script for the control of movement of the user's avatar, so jumping, animation and rotation follow the camera view, illustrated in Figure 4.

### 3.4 Teleport design

To teleport is to go from one place to another place. If we want to translate the avatar position to somewhere that the user cannot get to using basic movement, such as to the top of the tower, we use teleport to solve this problem. This application has 2 types of teleport. First is the green teleport and the second is the orange teleport. Users can move into the green teleport in the scene, shown in Figure 5 (a), then the user's avatar will translate the position to the top of that building. If the user wants to bring the avatar down, the user can teleport down by moving into the orange teleport, shown in Figure 5 (b), and then the user avatar will come back to the start position. The script of the teleport has 3 inputs, which are x, y, and z for translating the user's avatar by combination with input x, y and z.

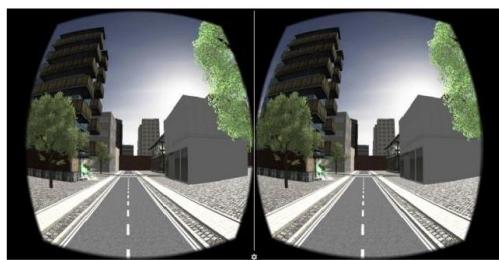
## 4 Results and Discussions

### 4.1 Results

We obtained a virtual reality system for height exposure. The system is a head-mounted display that put on a user's head. Users can control his/her character via the wireless remote controller, as seen in Figure 8. The application is a virtual reality application that has 2 paths of display, one for each eye. When the application starts, the user's character will appear at the start point, illustrated in Figure 9, and the menu will appear on the display for training the control of the character. Users can control character movement in the map, and users can use the teleport for moving up the buildings to start the height exposure.



**Fig. 8.** The system



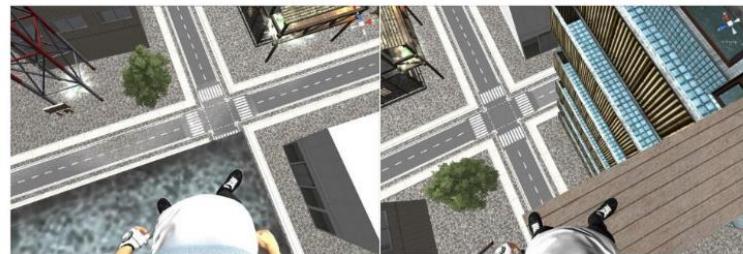
**Fig. 9.** Display output in virtual reality model

Level 1 balcony & roof: When user teleport upward, shown in Figure 10, their user avatar will stay on the balcony of this house. That balcony has a railing and is not so high and then users can continue to teleport upward to the house's roof from the balcony. That roof does not have a railing like the balcony, but is still not so high.



**Fig. 10.** Scene of balcony and roof

Level 2 seventh floor springboard: When a user teleport upward, the avatar will stay at the fringe of the terrace, as shown in Figure 11. The terrace has a platform outstretched from the terrace like a springboard. The platform is on the 7th floor with a small area, without a railing. When the user moves upward to the top of this building, the sound effect will be changed to a wind sound.



**Fig. 11.** Scene of seventh floor springboard

Level 3 insular: When a user teleport upward, the avatar will stay at the top of the radio tower with a very small area and the height is very high, as shown in Figure 12. When users move upward to the top of the radio tower, the sound effect will be changed to a wind sound.



**Fig. 12.** Scene of top of tower

#### 4.2 Discussion

In order to evaluate the performance of the proposed system, we conducted a questionnaire survey on 20 participants. The questionnaire consists of two parts. The first part contains questions relating to demographic data (gender, age, virtual reality experience, fear of heights). In the second part, included 16 questions which the view of participants reflected on the factors that might affecting the realism of the virtual environment and fear of heights.

The total number of participants is 20 persons, 13 men and 7 women, aged 15 - 40. Among the 20 persons, 5 persons have virtual reality experience and 2 persons have a fear of heights as displayed in the Table 1.

**Table 1.** Participants

Characteristics	Frequency	Percentage
Gender		
Male	13	65
Female	7	35
Age		
15 - 20	3	15
20 - 25	12	60
25-30	4	20
35-40	1	5
Virtual reality experience		
Always	1	5
Sometimes	4	20
Never	15	75
Fear of heights		
Yes	2	10
No	18	90

Data for the analysis are the scores received from the questionnaires. The scores were set from 1 – 4 as follows: 1 = few, 2 = average, 3 = much and 4 = very much. After the analysis by paired samples t-test with reliability at 95% (alpha = 0.05) to compare the scores of every question, the results show that when users work in the high building, the fear of heights becomes higher depending on the height of the building with significance (comparing building level 1 with level 2,  $t = 4.682$ ,  $p = 0.000$ ), (comparing building level 2 with level 3,  $t = 2.333$ ,  $p = 0.031$ ). In comparison between the score of realism and fear of heights when the sound is activated and deactivated, the score when the sound is activated is higher than with the deactivated sound with significance (when realism  $t = -9.200$ ,  $p = 0.000$ ), (fear of height  $t = 6.097$ ,  $p = 0.000$ ). When users can see their avatar model, the realism score is more than when they are unable to see the avatar with significance ( $t = 2.449$ ,  $p = 0.024$ ). Additionally, when the avatar model suitable with user to represent gender identity (male or female), the users will

have an increased score of fear with significance ( $t = 2.990$ ,  $p = 0.008$ ), while the scores of fear when users are able to see their avatar or unable to see it do not differ from each other ( $p = 0.385$ ). Also, there is no difference in the realism of the virtual environment scores when the avatar model and the user are suitable or unsuitable ( $p = -.104$ ).

**Table 2.** Descriptives

Score of (1 - 4)	Frequency	Mean	SD
Realism of virtual environment	20	2.30	0.733
Fear of heights at building level 1	20	1.20	0.523
Fear of heights at building level 2	20	1.95	0.686
Fear of heights at building level 3	20	2.30	0.865
Realism when sound activated	20	2.65	0.745
Realism when sound deactivated	20	1.60	0.681
Fear of heights when sound activated	20	2.60	0.754
Fear of heights when sound deactivated	20	1.85	0.745
Realism when user's avatar activated	20	2.65	0.875
Realism when user's avatar deactivated	20	2.05	0.826
Fear of heights when user's avatar activated	20	2.40	0.883
Fear of heights when user's avatar deactivated	20	2.20	0.768
Realism when using avatar model suitable with user	20	2.15	0.933
Realism when using avatar model unsuitable with user	20	1.95	0.826
Fear of heights when using avatar model suitable with user	20	2.25	0.851
Fear of heights when using avatar model unsuitable with user	20	1.85	0.875

**Table 3.** Paired samples t-test

Pairs	t	df	p
Fear of heights at building level 1 - Fear of heights at building level 2	-4.682	19	0.000
Fear of heights at building level 2 - Fear of heights at building level 3	-2.333	19	0.031
Realism when sound activated - Realism when sound deactivated	9.200	19	0.000
Fear of heights when sound activated - Fear of heights when sound deactivated	6.097	19	0.000
Realism when user's avatar activated - Realism when user's avatar deactivated	2.449	19	0.024
Fear of heights when user's avatar activated - Fear of heights when user's avatar deactivated	0.890	19	0.385
Realism when using avatar model suitable with user - Realism when using avatar model unsuitable with user	1.710	19	0.104
Fear of heights when using avatar model suitable with user - Fear of heights when using avatar model unsuitable with user	2.990	19	0.008

In conclusion, the height of the building, the sounds and the use of the avatar model which matches the user affects the fear of heights score rate the same as the use of sound and the user being able to see their avatar model affects the score of realism of the virtual environment. While playing, the moment that users can see the avatar model or cannot see it does not impact the fear of heights the same as choosing the model of their avatar to suitable with the user does not affect the realism of the virtual environment score. The means of each topic are displayed in the Table 2 and comparison by paired samples t-test is displayed in the Table 3.

Comparing the score of realism of the virtual environment of the application to 2 at 90% of reliability ( $\alpha = 0.1$ ), the received score is more than 2 with significance ( $p = 0.083$ ). To conclude, the score of realism of the virtual environment of the application is ranked highly and at the highest level as shown in the Table 4.

**Table 4.** One sample t-test

Score of (1-4)	t	df	p
Realism of virtual environment	1.831	19	0.083

## 5 Conclusion

We have created an application to be used for exposure to heights that runs on smartphones, as currently all types of smartphones are already widely used. Users just buy a few extra accessories to help create a virtual reality world. These accessories are virtual reality glasses and a wireless remote controller. The price is affordable. Users can use this application for confrontation with height situations at his/her own home, or anywhere. The hardware system is a standalone head-mounted display that does not connect with any wires. The software system is used to simulate and display a virtual environment via smartphones. In a virtual environment, users can confront 3 levels of height situations. Each level has a different height. Users can start with a low level of height and increase the level of height when users can pass the previous level. Application scenes are developed by Unity engine. The design is deployment object models in scenes that are easy-to-use and provide realism. In the future work, we are planning to use a new controller that can be used to perform more interactions with the virtual environment, and to make the application capable of connecting to the internet in order to contact psychiatrists to monitor users when using the application and give suggestions from the psychiatrist in the application scenes.

**Acknowledgements.** This work is supported by the Department of Computer Engineering, Faculty of Engineering, Mahidol University. The authors would like to thank all lecturers and members of the Image, Information and Intelligence Laboratory (I3 Lab).

## References

1. What is virtual reality? - virtual reality, <http://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html>.
2. Virtual reality, technology of the future. Episode 1, <https://blog.eduzones.com/darkfairytale/35>.
3. Ruangtrakool, S.: Textbook of psychiatry. Ruenkaew Printing, Bangkok (1999).
4. Costa, J.P., Robb, J., Nacke, L.E.: Physiological acrophobia evaluation through in vivo exposure in a VR CAVE. 2014 IEEE Games Media Entertainment. pp. 1–4. IEEE, Toronto (2014).
5. Benavides, C.: Virtual reality in the treatment of cynophobia. 2015 10th Computing Colombian Conference (10CCC). pp. 499 – 503. IEEE, Bogota (2015).
6. Cavrag, M., Lariviere, G., Cretu, A.-M., Bouchard, S.: Interaction with virtual spiders for eliciting disgust in the treatment of phobias. 2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings. pp. 29–34. IEEE, Richardson (2014).
7. Bruce, M., Regenbrecht, H.: A virtual reality claustrophobia therapy system - implementation and test. 2009 IEEE Virtual Reality Conference. pp. 179 – 182. IEEE, Lafayette (2009).
8. Tull, M.: How virtual reality exposure therapy (VRET) treats PTSD, <https://www.verywell.com/virtual-reality-exposure-therapy-vret-2797340>. [18 April 2016]
9. Hodges, L.F., Kooper, R., Meyer, T.C., Rothbaum, B.O., Opydyke, D., de Graaff, J.J., Williford, J.S., North, M.M.: Virtual environments for treating the fear of heights. Computer. 28, 27–34 (1995).
10. Haworth, M.B., Baljko, M., Faloutsos, P.: PhoVR: a virtual reality system to treat phobias. Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry - VRCAI '12. pp. 171–174. ACM (2012).
11. Schafer, P., Koller, M., Diemer, J., Meixner, G.: Development and evaluation of a virtual reality-system with integrated tracking of extremities under the aspect of Acrophobia. 2015 SAI Intelligent Systems Conference (IntelliSys). pp. 408 – 417. IEEE, London (2015).
12. Callaham, J.: Google says there are now 1.4 billion active Android devices worldwide, <http://www.androidcentral.com/google-says-there-are-now-14-billion-active-android-devices-worldwide>.
13. Bernardes, M., Barros, F., Simoes, M., Castelo-Branco, M.: A serious game with virtual reality for travel training with autism spectrum disorder. 2015 International Conference on Virtual Rehabilitation (ICVR). pp. 127 – 128. IEEE, Valencia (2015).
14. Rincon, A.L., Yamasaki, H., Shimoda, S.: Design of a video game for rehabilitation using motion capture, EMG analysis and virtual reality. 2016 International Conference on Electronics, Communications and Computers (CONIELECOMP). pp. 198 – 204. IEEE, Cholula (2016).
15. Covarrubias, M., Bordegoni, M.: Immersive VR for natural interaction with a haptic interface for shape rendering. 2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI). pp. 82 – 89. IEEE, Turin (2015).
16. ISBIT GAMES: First Person Lover - Male Character, <https://www.assetstore.unity3d.com/en/#!/content/40848>.
17. ISBIT GAMES: First Person Lover - Female Character, <https://www.assetstore.unity3d.com/en/#!/content/41056>.
18. Kakky: KY Magic Effects Free, <https://www.assetstore.unity3d.com/en/#!/content/21927>.

19. 7thDimension: Medieval Buildings, <https://www.assetstore.unity3d.com/en/#!/content/34770>.
20. CGY (Yemelyan K.): Block Building Pack, <https://www.assetstore.unity3d.com/en/#!/content/13925>.
21. VR: Radio Tower - Low Poly, <https://www.assetstore.unity3d.com/en/#!/content/2299>.

## CLASSIFICATION OF DIABETIC RETINOPATHY STAGES USING IMAGE SEGMENTATION AND AN ARTIFICIAL NEURAL NETWORK

Narit Hnoohom<sup>1</sup> and Ratikanlaya Tanthuwapathom<sup>2</sup>

Image, Information and Intelligence Laboratory, Department of Computer Engineering,  
Faculty of Engineering, Mahidol University, NakornPathom, Thailand  
<sup>1</sup>narit.hno@mahidol.ac.th, <sup>2</sup>t.ratikanlaya@gmail.com

**Abstract.** Diabetic retinopathy, which can lead to blindness, has been found in 22 percent of diabetic patients in the latest survey. Therefore, diabetic patients should have an eye examination at least once a year. However, it has been found that currently there is a problematic lack of specialists in ophthalmology. Detection and treatment of diabetic retinopathy are thus delayed. The idea to create a classification system of diabetic retinopathy stages to facilitate the making of preliminary decisions by ophthalmologists is introduced. This paper presents the classification of diabetic retinopathy stages using image segmentation and an artificial neural network. This proposed method applies local thresholding to separate the foreground region from the background region so that the optic disc and exudate regions are able to be identified more clearly. The experiment was carried out with 100 fundus images from the Institute of Medical Research and Technology Assessment database. The prediction model had an accuracy rate of up to 96 percent.

**Keywords:** Diabetic retinopathy, Exudates, Fundus image, Artificial neural network.

### 1 Introduction

Due to a genetic disease characteristic, 371 million people around the world currently have diabetes mellitus, and 500 million diabetic patients are expected by 2030. In Thailand, the number amounts to approximately 3.5 million patients with diabetes mellitus. In addition, a recent survey found that 22 percent of people with diabetes mellitus have diabetic retinopathy, which can lead to blindness. Furthermore, the risk of loss of eyesight among people with diabetes mellitus is at least 20 percent higher than ordinary people. Although diabetic patients may experience the same pathologies as ordinary people, e.g. cataracts, glaucoma and optic nerve terminal inflammation, these pathologies are encountered at younger ages with greater frequency, severity and treatment complexity than in other people, even though all of these symptoms can be

treated with the same methods used for non-diabetics [2, 13]. However, there is one critical condition that is only found in diabetic patients, and that is diabetic retinopathy, which usually occurs in patients who have had diabetes mellitus for a long time. According to previous findings, patients who have had diabetes mellitus for less than 10 years are at 7 percent greater risk for diabetic retinopathy. However, the risk increases to 63 percent for patients with diabetes mellitus for more than 15 years and patients with good glucose control can still have diabetic retinopathy in the long term or at an older age.

The treatment of patients with eye problems requires an ophthalmologist with advanced instruments. Although there are currently around 700 ophthalmologists in Thailand, this number is insufficient for the increasing number of diabetic patients. Thus, diabetic retinopathy screening is considered an important process in helping patients. In order to treat this group of patients in time, all patients with diabetes mellitus should have their eyes examined at least once every year to thoroughly check for symptoms of diabetic retinopathy.

This article focuses on the development of detection and classification methods that can be utilized for diabetic retinopathy screening. In this area, several methods have been developed and the area is still being explored. These methods can be divided into two categories: blood vessel detection and exudate detection.

The blood vessel detection method is used to detect and classify the amount of blood vessels in the retinal image. K. Verma et al [1] presented the classification of moderate and non-proliferative phases of diabetic retinopathy (NPDR) by finding the amount of blood vessels and blood in the retinal images. Blood vessels in the retinal image can be used to classify the differences between the vascular area and background. The principle means of the segmentation in this research was thresholding and the use of the dot-blot hemorrhage microaneurysm (MA) as the exudate indicator. This method employs the use of a retinal image database with a total of 65 fundus images consisting of 30 images showing normal stages, 23 images showing moderate stages and 12 images showing severe stages from a structured analysis of the retina. The result of the system test was a correct classification rate of 93 percent.

Exudate detection is used to detect and classify the amount of exudates in the retinal image. Z. Ahmad et al [2] presented the development of this detection and classification system with segmentation. The DR classification process depends on the amount of exudates. In the retinal images, the system is able to help ophthalmologists perform early screening of patients with diabetes mellitus. There are two processes of exudates detection; rough and fine exudates segmentation. Rough segmentation is performed by using the morphology operation and column-wise neighborhoods operation, while a good classification should be done by using the morphological reconstruction. This method employs the use of a retinal image database with a total of 239 fundus images consisting of 110 images showing normal stages, 63 images showing mild stages, 36 images showing moderate stages and 30 images showing severe stages from Sungai Buloh Hospital in Malaysia together with more appropriate retinal feature extractions. This classification method can also provide a simple and fast basis for analysis. The result of the system test was 60 percent correct classification.

S. Sreng et al [3] proposed early detection of diabetic retinopathy using the green channel and red channel, exudate extraction with the image binarization, region of interest (ROI) based segmentation and morphological reconstruction. This method employs the use of a retinal image database with a total of 100 fundus images from Bhumibol Adulyadej Hospital. The result of the system test was 91 percent correct classification.

M. Eman Shahin et al [4] presented the method for detection of exudates in retina images using a morphological and candy method. Following microaneurysms detection using Histogram equalization, morphological and candy edge detector, this method employs the use of a retinal image database with a total of 340 fundus images consisting of 89 images showing normal stages and 251 images showing abnormal. The result of the system test was 92 percent correct classification by ANN.

Du Ning et al [5] presented the method of using computers for monitoring the diabetic retinopathy stage using a color fundus images process. The techniques of morphological processing and method of texture analysis are applied to the retinal images for the feature examination, for example, vascular areas, exudates, sharpness of the identical areas, and features that will be added to the support vector machine (SVM). This method employs the use of a retinal image database with a total of 52 fundus images consisting of 10 images showing normal stages, 35 images showing NPDR stages and 7 images showing proliferative phases of diabetic retinopathy (PDR) stages from the standard diabetic retinopathy database. This research showed the correct classification of diabetic retinopathy at 93 percent and detected hard exudates in color fundus images of the human retina.

C. Jayakumari et al [6] presented the method of exudates detection using the following process: histogram equalization then segmentation of the image using a contextual clustering algorithm. The selected sets of features are the standard deviation of the intensity, mean, intensity, area, edge strength and learning system by the echo stage neural network (ESNN). The images are classified class into two categories, which are normal and abnormal. A total of 50 images are used to find the exudates, out of which 35 images are used to train the ESSN and the remaining 15 images are used to test the neural network. This article confirms 93 percent sensitivity and 100 percent specificity in terms of exudates based classification.

This paper presents a classification method for diabetic retinopathy screening by using image segmentation and an artificial neural network to classify retina images of diabetic retinopathy at the normal, moderate, and severe stages. The proposed method applies a local thresholding to separate the foreground region exudates from the background region. This can identify the optic disc and exudate regions in fundus images more clearly. Detection and extraction of the optic disc are obtained using a morphological operation. In order to demonstrate the reliability of the method, the proposed method is compared with the results of the examination by ophthalmologists. The proposed method also applies related principles in order to accelerate the treatment of patients.

The remainder of the paper is organised as follows. First, a general introduction to the detection method of diabetic retinopathy is briefly described. In Section 2, a description of the proposed method with the processes of classification of diabetic retinopathy

stages is addressed. Experiment results and discussions are included in Section 3. Finally, the conclusions are presented in Section 4.

## 2 Proposed method

The main objective of the proposed method is to classify the diabetic retinopathy stages. The proposed method is shown in Figure 1, and contains the following main five processes: (1) Image acquisition; (2) Pre-processing; (3) Exudate extraction; (4) Features extraction; and (5) the Artificial neural network. The following sections present a brief summary of each step.



**Fig. 1.** Overview of the proposed method

### 2.1 Image Acquisition

The first phase of the classification of diabetic retinopathy stages is the image acquisition. The fundus images were collected from the Institute of Medical Research and Technology Assessment (MRTA) database, which contains 100 images. The database consists of both normal retinal images and abnormal retinal images. The set of normal retinal images contains 18 images and the set of abnormal retinal images contains 82 images, in which 47 images are moderate retinal images and 35 images are severe retinal images. The original fundus images, which are a size of  $3872 \times 2592$  pixels in the Joint Photographic Experts Group (JPEG) file format, were captured by the fundus photographing optical system.

### 2.2 Pre-processing

During the pre-processing stage, the fundus image data is prepared for use in the classification of diabetic retinopathy stages.

#### 2.2.1) Image resize

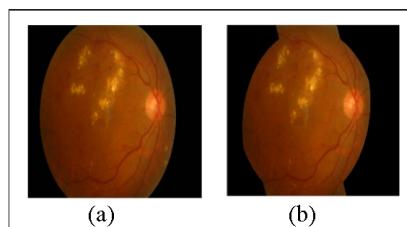
The standard size of the input fundus images used in this proposed method is defined as  $512 \times 512$  pixels. This can reduce computation time necessary for the detection of the exudate regions. Figure 2 shows the result image after resizing.



**Fig 2.** Image after resizing

#### 2.2.2) Edge removal

To remove the edge of the resized image, we delete the edges to trim the brightness at the edges of the image. The results of reducing the brightness of the image area is not a problem for segmentation in the next step. The result after removing edge is shown in Figure 3.

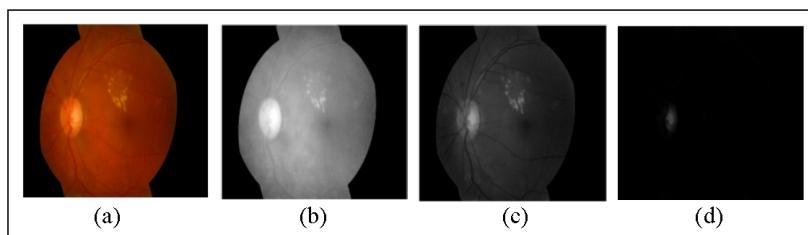


**Fig. 3.** (a) Resized image, (b) Removed edge image

#### 2.2.3) Color model selection

In this step, we split the input image into three parts: red channel, green channel and blue channel. When the channels are separated, the findings were as follows.

In Figure 4, it can be seen that the red channel is the channel with the most visibly intense light, but the differences of intensity in each component is relatively low making it hard to distinguish the composition. The green channel is the channel that can best show the differences in light intensity of each component of the images with the optic disc and exudates being the most intense. The blue channel is found to have the lowest color intensity and cannot be used to help in the feature differences analysis.



**Fig. 4.** (a) Original image, (b) Red channel, (c) Green channel, (d) Blue channel

### 2.3 Exudate Extraction

Exudate extraction is composed of two stages: the region of interest extraction and the optic disc removal.

#### 2.3.1) Region of interest extraction

The region of interest extraction consists of the optic disc and exudate regions in the green channel of fundus images. In order to extract the region of interest, this proposed method applies a local thresholding to separate the foreground region from the background region so as to be able to identify the optic disc and exudate regions more clearly. Figure 5 shows the green channel of a fundus image and the region of interest extraction in the fundus image.

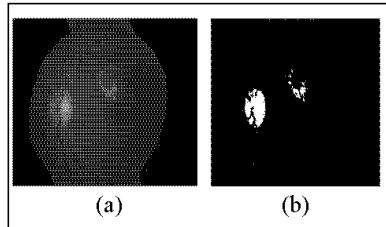
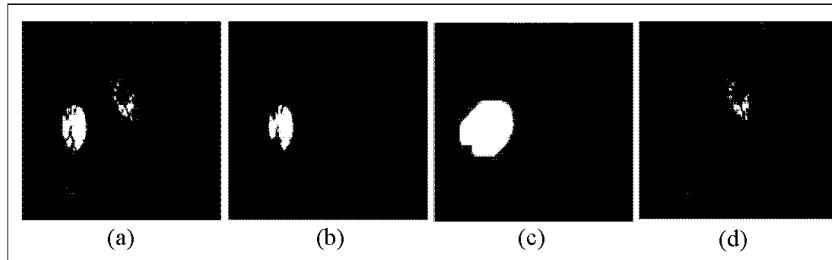


Fig. 5. (a) Green channel image, (b) Optic disc and exudate regions extracted using local thresholding

#### 2.3.2) Optic disc removal

This section consists of two steps: the optic disc extraction and the optic disc removal. The optic disc region is obtained by removing the exudate regions in the result of the region of interest extraction using a morphological opening operation, which includes erosion and dilation methods in processing [7, 8, 9].

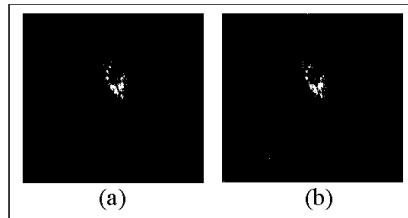
In the optic disc removal, the completed optic disc region is created by the morphological dilation operation in order to expand the shape of optic disc contained in the region of interest extraction. After that, the region of interest extraction is separated from the complete optic disc region in order to remove the optic disc region, and the exudate regions are obtained. Figure 6 shows the various stages of the optic disc removal process.



**Fig. 6.** (a) Region of interest extraction, (b) Optic disc extraction, (c) Completed optic disc obtained using dilation, (d) Exudate region extracted using morphological operation

### 2.3.3) Connected area analysis

Finding the exudates by using adjoining regions is a very good method. Therefore, this paper adopts this method for the study. The researchers were able to classify the exudates by considering neighboring pixels to help decide whether or not the exudates were present. On the other hand, if the neighboring pixels do not meet the specified criteria, the image will be considered to be contaminated by the noise within it, as shown in Figure 7.



**Fig. 7.** (a) Output of exudate extraction with green channel, (b) Output of exudate extraction with gray scale

### 2.4 Features extraction

Feature extraction is used to find the special features of the exudate extraction in retinal images. This paper used eleven features for classification of diabetic retinopathy stages:

- Standard deviation (SD) of the intensity, which is white pixels, is used to quantify the amount of variation of exudate regions in the fundus image. The SD formula is defined as follows:

$$SD_{intensity} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N |A_i - \mu|^2} \quad (1)$$

where  $SD_{intensity}$  is the standard deviation of intensity,  
 $A$  is the area of exudate,  
 $N$  is the number of exudates and  
 $\mu$  is the mean of intensity in the fundus image.

- Mean of intensity, is the average value of the gray scale image. Mean intensity can be determined by equation (2).

$$\mu_{intensity} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N A_i} \quad (2)$$

where  $\mu_{intensity}$  is the mean intensity,

$A$  is the area of exudate and

$N$  is the number of exudates.

- Sum of intensity, is the total of intensity, which can be determined by equation (3).

$$S_{intensity} = \sum_{i=1}^N A_i \quad (3)$$

where  $S_{intensity}$  is the sum intensity,

$i$  is number of exudate,

$A$  is the area of exudate and

$N$  is the number of exudates.

- Edge strength, is measured as the average of the edge values in the perimeter of the region. The edge values were obtained after the application of a Prewitt operator. The edge strength formula is defined as follows:

$$ES = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N E_{ij} \quad (4)$$

where  $ES$  is the edge strength,

$MN$  is the total pixel of exudate,

$E$  is the edge value of exudate and

$i, j$  is the position of pixel.

- Compactness, is another measure of circularity where  $A$  and  $P$  are the area and perimeter of the candidate region. Compactness can be determined by equation (5).

$$C = \frac{P^2}{4\pi A} \quad (5)$$

where  $C$  is exudate compactness,

$A$  is the area of exudate and

$P$  is the exudate perimeter.

- Area, is the area of exudates. Area can be determined by equation (6).

$$A = \sum_{i=1}^N \sum_{j=1}^N B[i, j] \quad (6)$$

where  $A$  is the exudate area,  
 $B$  is the exudate and  
 $i, j$  is the position of pixel.

- SD of hue, saturation, and value (HSV), is applied to quantify the amount of variation of the exudate regions in the H, S and V.
- SD of the green channel, which is white pixels, is used to quantify the amount of variation of exudate regions in the green channel.
- Perimeter is the distance around the exudate regions.
- Major axis length is the longest diameter of the exudate regions.
- Minor axis length is the shortest diameter of the exudate regions.

## 2.5 Classification

This paper describes exudate classification by an artificial neural network. Training is a process of using samples to develop the artificial neural network utilizing the types of input with the correct answers [12]. In this process, the group of samples with known output is repeatedly sent to the network to train the network system. In the proposed method, the training process was carried out until there were differences between input and output, and the pattern for the training set value was acceptable. Several methods are available for the network training. The back-propagation method is commonly used and is capable of working successfully in two steps. In the first step, input is sent forward through the network to produce output. In the second step, the difference between real output and expected output generates an error signal that returns the value over the network to improve the weight of the input.

Research on diabetic retinopathy has been given a lot of attention and there are numerous articles that have proposed many methods. Thus, it is difficult to compare our algorithm with the other reported work in the literature. The paper from Jayakumari [6] was chosen for comparison because there were findings on the exudates detection for classification of diabetic retinopathy stages and some features that were used identically.

## 3 Results and Discussions

This paper aimed to evaluate the performance by accuracy, precision and recall. The algorithm used in the classification of diabetes retinopathy was employed for experimentation with WEKA 3.7.1 for comparison. System learning with a series of 100 retinal images, training and testing was conducted with the WEKA system by a multilayer perceptron classification. In this paper, the system learning value was set at

a learning rate of 0.3, a momentum of 0.2, a training time of 500, and a 10-fold cross-validation for the two sets to be tested. The Jayakumari features set comprised the SD of intensity, mean intensity, sum of intensity, edge strength, compactness and area. The proposed features set comprised the SD of HSV, SD of green channel, perimeter, major axis length and minor axis length.

The performance series features of the two sets of algorithms in the multilayer perceptron test compared the efficiency of the two features by measuring the effectiveness of information classification, precision and recall, by which accuracy was calculated by Equations (7), (8) and (9):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

where recall is the recollection of classification categories for each group.  
 Precision is the accuracy of the classification categories for each group.  
 True Positive ( $TP$ ) is the prediction of  $p$ , and the actual value is  $p$ .  
 False Positive ( $FP$ ) is the prediction of  $p$ , but the actual value is  $n$ .  
 True Negative ( $TN$ ) is the prediction of  $n$ , and the actual value is  $n$ .  
 False Negative ( $FN$ ) is the prediction of  $n$ , but the actual value is  $p$ .  
 $p$  is the correct prediction for the model.  
 $n$  is the incorrect prediction for the model.

**Table 1.** Results of qualifying features set for the work

Feature sets	Accuracy	Precision	Recall
Jayakumari features set			
SD of intensity			
Mean of intensity			
Sum of intensity	94%	94.1%	94%
Edge strength			
Compactness			
Area			
Proposed features set			
SD of HSV			
SD of green channel	96%	96.2%	96%
Perimeter			
Major axis length			
Minor axis length			

From Table 1, the Jayakumari features set had a result for accuracy of 94 percent, for precision of 94.1 percent and for recall of 94 percent. The proposed features set had a result for accuracy of 96 percent, for precision of 96.2 percent and for recall of 96 percent. As a result, the proposed features set yielded a higher accuracy than the Jayakumari features set.

From the experiments, we found that the SD of HSV feature and the SD of green channel feature were better than the SD of intensity feature. That brings good results because in the process of segmentation in the green channel there is more detail than when using the gray scale. The perimeter feature is more suitable for this paper than the area feature, because the shape of the exudates has a variety of forms. Therefore, the perimeter feature had better results than the area feature. When we know the major axis length and the minor axis length, we can know the shape of exudates.

Furthermore, it was found that the proposed exudates screening will work well when the input image is clear with the optic disc and exudates clearly seen. The presentation of this research can be compared to the research conducted by Sreng [3] with the elimination of the optic disc even if the optic disc is not circular.

#### 4 Conclusions

This paper presented the classification of diabetic retinopathy stages using image segmentation and an artificial neural network. The data sets are separated into three groups that are the set of normal retinal images and the set of abnormal retinal images, in which are moderate retinal images and severe retinal images. The groups were classified by exudates detection, and if detected, the exudate in the retinal image is categorized as a moderate or severe class. Nevertheless, the proposed method has a limitation in the optic disc extraction, which is that the optic disc image must be clearly seen. In future research work, this can be extended in order to classify the mild stages of diabetic retinopathy.

**Acknowledgment.** This project is supported by Department of Computer Engineering, Faculty of Engineering, Mahidol University. The authors would like to thank all lecturers and members of Image, Information and Intelligence Laboratory (I3 Lab). We also would like to thank Institute of Medical Research and Technology Assessment for the database.

#### References

1. Verma, K., Deep, P., Ramakrishnan, A.G.: Detection and classification of diabetic retinopathy using retinal images, In: India Conference (INDICON), pp.1-6. IEEE Press, India (2011)
2. Ahmad Zikri, R., Hadzli, H., Syed, F.: A Proposed Diabetic Retinopathy Classification Algorithm with Statistical Inference of Exudates Detection. In: 2013 International Conference, Electrical, Electronics and System Engineering (ICEESE), ,pp.80-95.IEEE Press, Malaysia (2013)

3. Sreng, S., Maneerat, N., Isarakorn, D., Pasaya, B., Takada, J., Panjaphongse, R., Varakulsiripunth, R.: Automatic Exudate Extraction for Early Detection of Diabetic Retinopathy. International Conference. In: Information Technology and Electrical Engineering (ICITEE), pp.31-35.2013. IEEE Press, Thailand (2013)
4. Shahin, E.M., Taha, T.E., Al-Nuaimy, W., El Rabaie, S., Zahran, O.F., El-Samie, F.E.A.: Automated Detection of Diabetic Retinopathy in Blurred Digital Fundus Images. In: 8th International Computer Engineering Conference (ICENCO), pp20-25, U.K (2013)
5. Du Ning, Li Yafen: Automated identification of diabetic retinopathy stages using support vector machine. In: 32nd Chinese on Control Conference (CCC), pp. 3882 - 3886. IEEE Press, China (2012)
6. Jayakumari C., Maruthi R: Detection of Hard Exudates in Color Fundus Images of the Human Retina. Procedia Engineering, Volume 30, 2012, pp. 297-302 (2012)
7. Saravanan, V., Venkatalakshmi, B., Farhana, S.M.N.: Design and development of pervasive classifier for diabetic retinopathy. In: 2013 IEEE Conference on Information & Communication Technologies (ICT), pp.231-235. IEEE Press, India (2013)
8. Zeljkovic, V., Bojic, M., Tameze, C., Valev, V.: Classification algorithm of retina images of diabetic patients based on exudates detection. In: 2012 International Conference on High Performance Computing and Simulation (HPCS), pp.167-173. IEEE Press, New York (2012)
9. Dhiravidachelyi, E., Rajamani, V.: Computerized detection of optic disc in diabetic retinal images using background subtraction model. In: 2014 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp.1217-1222. IEEE Press, India (2014)
10. Usman Akram M., Shehzad K., Shoab Khan A.: Identification and classification of microaneurysms for early detection of diabetic retinopathy. Pattern Recognition, Volume 46, Issue 1, pp.107-116 (2013)
11. María G., Clara Sánchez I., María López I., Daniel A., Roberto H.: Neural network based detection of hard exudates in retinal images. Computer Methods and Programs in Biomedicine, Volume 93, Issue 1, pp. 9-19 (2009)
12. Anitha, J., Selvathi, D., Hemanth, D.J.: Neural Computing Based Abnormality Detection in Retinal Optical Images. In IACC 2009. IEEE International on Advance Computing Conference, pp.630-635. IEEE Press, India (2009)
13. J. R. Jensen. Calculate Confusion Matrices. Retrieved December 20, 2015, Web site: <http://www.exelisvis.com/docs/CalculatingConfusionMatrices.html>.

## Examination timetabling using prey predator algorithm

Surafel Luleseged Tilahun and Jean Medard T Ngnotchouye

School of Mathematics, Statistics and Computer Science,  
University of KwaZulu-Natal, 3209, Pietermaritzburg, South Africa,  
surafelaau@yahoo.com

**Abstract.** Education timetabling problem is a problem of arranging exams so that there will not be a clash with the objective of using a minimum number of time slots. It is a combinatorial optimization problem which has been studied by different authors. For combinatorial problems in general, and educational timetabling problem in particular, different metaheuristic algorithms have been proposed and used. Some of these algorithms are originally proposed for continuous problems and later extended for combinatorial optimization problems and are found to have a promising performance. Based on the well-known 'no free lunch theorem', there is no superior algorithm for all problem domains. This implies that if an algorithm performs better than another algorithm in some problems or problem domain then there exist another set of problems or problem domains in which it will be out performed. Hence, having different approaches to deal with these problems is advantageous, while how to choose an algorithm for a problem at hand remains to be a research to be studied further. Hence, in this paper, prey predator algorithm will be modified to suit combinatorial optimization problems in general, and educational exam timetabling in particular, with adapted local search mechanism which guides the search over the constrained set of solutions. In order to test the approach five problem instances are chosen from Carter uncapacitated exam timetabling benchmark problem. These problem are well known benchmark problems for high dimensional combinatorial optimization problems. Simulation results show that the proposed approach is promising and comparable with previous approaches.

**Keywords:** Educational timetabling, prey predator algorithm (PPA), metaheuristic

### 1 Introduction

An optimization problem is a problem of finding the best action which leads to the optimum measure of an objective function. These problems appear in different areas ranging from our day to day activity to a complex business planning and policy making. Due to the limitation of resources, these problems are constrained. Furthermore, the decision variables can be assigned with only integers. A combinatorial optimization problem is an optimization problem which has a

finite and discrete, usually large, solution space [1]. Different real scenarios which are formulated as combinatorial optimization problem can be mentioned including exam timetabling problem [2], travel salesman problem [3] and bin packing problem [4].

Different exact optimization methods are introduced to solve optimization problems based on their properties, for instance if the problem is linear and continuous then one can use simplex algorithm. However, these exact methods become shorthanded when dealing with difficult real problems which do not fall under the categories for which exact methods are introduced for. In such cases, metaheuristic algorithm will become an ideal choice. Even though metaheuristic algorithms don't guarantee optimality, they are tested and found to give a reasonable solution under appropriate implementations. Since the introduction of evolutionary algorithms different metaheuristic algorithms are introduced. Prey predator algorithm is one of the recently introduced metaheuristic algorithms, which mimics the scenario of a predator which runs after its prey. Adjusting the degree of exploration and exploitation is one of the challenging issues and has been in the forefront of research issue in the field [5]. Prey predator algorithm gives a clear way of controlling between exploration and exploitation. It also has been studied and also tested in different problems and is found to be effective [6, 7, 8, 9, 10, 11, 12]. In addition, it is a generalized swarm based algorithm where some well-known algorithms fall as a special case of this algorithm, including particle swarm algorithm and modified firefly algorithm [13].

Along with continuous optimization problems, metaheuristics algorithms are used for combinatorial optimization problems. Many of these algorithms are modified so that it can be used for constrained optimization problems. The 'No free lunch theorem' proves that there is no superior metaheuristic algorithm for combinatorial optimization problems. Therefore, modifying the newly introduced prey predator algorithm for the use of solving constrained combinatorial optimization problem is one contribution to the solution methods of these problems. Perhaps, it will perform better for a good number of problem domains because one can easily control the degree of exploration and exploitation as needed in prey predator algorithm. Furthermore, it will add to the set of possible solution algorithms for combinatorial optimization problems. Hence, this paper presents the modification of prey predator algorithm for high dimensional constrained combinatorial optimization problems, based on a new modified local search technique to be used for combinatorial optimization problems in general and educational timetabling problems in particular. The approach will be tested on uncapacitated timetabling problem. It will be shown that on the selected benchmark problems the proposed approach indeed performs well.

In the next section basic concepts will be discussed on the constrained combinatorial optimization problem along with a brief explanation of prey predator algorithm. In section 3 the proposed approach will be given followed by demonstration on selected problems in section 4. Finally, the conclusion will be presented in section 5.

## 2 Basic concepts

### 2.1 Combinatorial optimization problems

A combinatorial optimization problem is an optimization problem with discrete and finite solution space [14]. It has an objective function,  $f(x)$ , along with the feasible region,  $S$ . In a typical combinatorial optimization, if the values that can be assigned to the decision variables are from set  $E$  then the feasible region will be a subset of power set of  $E$ , as given in equation 1.

$$\begin{aligned} & \min f(x) \\ \text{s.t. } & x \in S \subseteq 2^E \end{aligned} \tag{1}$$

Note that a solution  $x$  is a vector of  $n$  dimension with entries from  $E$ , for example  $E$  can be integers between 1 and 100. In constrained optimization problems, some of the components are restricted not to take some values from set  $E$ . Hence, a typical constrained optimization problem can be given by:

$$\begin{aligned} & \min f(x) \\ \text{s.t. } & x \in S \subset 2^E \end{aligned} \tag{2}$$

A particular example will be exam timetabling problem. For each exam you can assign a time slot which is arranged as 1, 2 and so on. However, there should not be a clash between exams, which means exams with common students should not be arranged in the same slot. Hence, you cannot assign a slot randomly for all the exams, as it will result clash and that results infeasibility. In addition to exam clashes other constraints like exam room capacity and availability of invigilators can be used.

### 2.2 Prey predator algorithm (PPA)

Prey predator algorithm (PPA) is one of the recently introduced metaheuristic algorithms [13, 15]. It mimics the interaction between a predator (which runs after its prey) and its prey in the natural environment. In the algorithm a randomly generated set of feasible solutions will be put in three categories, a predator, ordinary prey and a best prey, based on their performance in the objective function which needs to be optimized. In each iteration of the algorithm the predator tends to explore the solution space with bigger step length and also run after the weak prey which is the prey with worst performance in the objective function. However, the best prey will only do a local search and totally focuses on the exploitation of the neighborhood. On the other hand, the ordinary prey will run away from the predator or follow better prey, prey with better performance in the objective function, based on the algorithm parameter called probability of follow-up. The prey predator algorithm is introduced and used for continuous problems.

Probability of follow-up is one of the algorithm parameter which controls the movement of the ordinary prey. High probability follow-up means 'following better prey' will have higher chance than 'randomly run away from the predator'. If a randomly generated number is under the probability of follow-up then it will follow better prey and do a local search as well, but if the randomly generated number is greater than the probability of follow-up then the prey will randomly run away from the predator. The other algorithm parameter is step length. There are two step lengths,  $\lambda_{max}$  and  $\lambda_{min}$ .  $\lambda_{max}$  is a step length for exploration whereas  $\lambda_{min}$ , which is shorter than  $\lambda_{max}$ , is a step length for exploitation.  $m$  is another algorithm parameter, which determines the number of random unit directions used for exploitation or local search purpose. The algorithm is summarized in table 1.

Step 1	Generate random and feasible $N$ solutions
Step 2	Sort the solutions based on their performance in the objective function from worst to best and Assign $x_1$ to be the predator, $x_N$ to be the best prey and the rest ordinary prey
Step 3	Move the predator randomly and also towards $x_2$ i.e. $x_1 := x_1 + (\lambda_{max} \cdot rand)u + (\lambda_{min} \cdot rand)u_{1,2}$ where $rand$ is a random number from uniform distribution between 0 and 1, $u$ is a random unit vector and $u_{1,2} = \begin{cases} \frac{x_2 - x_1}{\ x_2 - x_1\ }, & \text{if } x_1 \neq x_2 \\ 0, & \text{otherwise} \end{cases}$
Step 4	Move the best prey using a direction chosen from randomly generated $m$ unit directions and zero vector i.e. $x_N := x_N + (\lambda_{min} \cdot rand)u$ where $u = \min_{u_i \in u_1, u_2, \dots, u_m, 0} f(x_N + (\lambda_{min} \cdot rand)u_i)$
Step 5	Update the ordinary prey either by following better performing prey if probability of follow-up is met or randomly away from the predator. i.e. if $rand \leq$ Probability of follow-up $x_i := x_i + (\lambda_{max} \cdot rand)u_i + (\lambda_{min} \cdot rand)u$ where $u_i = \frac{\sum_{j=1}^{j=i+1} (x_j - x_i)}{\ \sum_{j=1}^{j=i+1} (x_j - x_i)\ }$ and $u$ is similar with the one in step 4 else (if $rand >$ Probability of follow-up) $x_i := x_i + (\lambda_{max} \cdot rand)u_{rand}$ where $u_{rand} = \begin{cases} u & \text{if } \ x_1 - (x_i + u)\  \geq \ x_1 - (x_i - u)\  \\ -u & \text{otherwise} \end{cases}$ for a random unit direction $u$
Step 6	If a termination criterion is met stop else go back to step 2.

Table 1. Prey predator algorithm

### 3 The proposed approach

As presented in section 2, combinatorial optimization problem is an optimization problem with discrete values for the decision variables. When there are cases which needs to be fulfilled, the problem become challenging and finding a feasible solution by itself will be another challenging task because a random vector within the boundary doesn't necessary be feasible. In order to construct a feasible solution to proceed with PPA, different problem specific methods can be used. Even after generating initial feasible solutions, the other biggest challenge in applying PPA for these types of problems is the updating process. Unlike the continuous case, when we update a solution, using the equations in table 1 it will become infeasible or will become slow due to the rounding of the solutions to

be feasible. Hence the updating process should be modified while mimicking the same scenario. In this paper we propose a version of PPA for constrained combinatorial optimization problems with a new updating and local search mechanism. Consider the following problem:

We want to minimize  $f(x)$  by choosing  $n$  dimensional vector value for  $x$  in such a way that  $x$  fulfill conditions given as a set in  $S$ , feasibility. Without loss of generality we can also assume that each component or entries of  $x$  can be assigned with an integer between 1 and a natural number  $D$ , for  $D > 1$ .

First using different methods based on the problem, like saturation degree for exam timetabling problem [16], a random set of feasible solutions will be generated, say  $x_1, x_2, \dots, x_N$ . The number of solutions should be more than two as we have three categories of solutions in PPA. Among this solutions based on their performance in the objective function,  $f(x)$ , they will be put in three categories; the one with least performance will be called predator represented by  $x_p$ , the one with best performance is called best prey represented by  $x_b$  and the rest as ordinary prey. Unlike in the continuous case, the local search will be done by changing only a number of entries of the solution based on a new algorithm parameter called probability of change ( $p_c$ ). In this context a local search is when solutions try to look for another better solution by changing only a small number of its components. Hence for each component if a randomly generated number is less than  $p_c$ , then that entry will be changed as long as the solution remains feasible. However, if the change results infeasibility it will be rejected. Hence, small number of  $p_c$  means exploitation whereas larger number of  $p_c$  leads to exploration. As discussed in the previous section, the other algorithm parameter in the continuous optimization case is  $m$ , the number of random directions for the local search of the best prey. Similarly we can do  $m$  local updates and choose the one with best performance; we add the original best prey in this comparison. The updating process of the best prey is summarized in table 2.

The updating process of an ordinary prey  $x_i$  depends on the probability of follow-up ( $p_f$ ). If a randomly generated number is at most equal to the probability of follow-up then for each component, based on the probability of change ( $p_c$ ), here  $p_c$  should be higher than the one used for the local search, will be updated by moving the towards the components of better prey unlike in the local search using random arrangement of natural numbers ( $randperm$ ), provided that the solution will remain to be feasible, using equation 3.

$$x_i(k) := round[x_i(k) + (2rand)sign(\sum_j (x_j(k) - x_i(k)))] \quad (3)$$

for each better prey  $j$  and components  $k$ ,  $rand$  is a random number between 0 and 1.

Similarly, if the probability of follow-up is less than a randomly generated number then the prey will randomly run away from the predator; that will be done again component wise as given in equation 4.

$$x_i(k) := round[x_i(k) + (2rand)(x_p(k) - x_i(k))] \quad (4)$$

---

```

Input  $p_c, m, D$ 
for  $k = 1 : m$ 
     $y_k = x_b$ 
    for  $i = 1 : n$ 
        if  $rand \geq p_c$ 
             $T = randperm(D)$ 
            for  $j = 1 : D$ 
                if  $y_k(i) \neq T(j)$ 
                    if making  $y_k(i) = T(j)$  doesn't make  $y_k$  infeasible
                        then  $y_k(i) = T(j)$ 
                        break the  $j$  loop (or update  $j = k$ )
                    end if
                end if
            end for
        end if
    end for
end for
Compare  $y_1, y_2, \dots, y_m, x_b$  and take the best to update  $x_b$ 

```

---

**Table 2.** updating process for the best prey ( $randperm(D)$  is a random arrangement of  $D$  natural numbers)

for the predator  $x_p$  and components  $k$ ,  $rand$  is a random number between 0 and 1.

In both of the equation above we used a magnifying value 2, which sometimes allow the updating process to explore beyond the point of interest. The updating process of an ordinary prey is summarized as follows in table 3.

The predator will do a total exploration, hence the updating will be done similar with the one used with the best prey but with higher  $p_c$ .

Note that the step lengths,  $\lambda_{min}$  and  $\lambda_{max}$ , in the standard prey predator algorithm is replaced by  $p_c$ .

#### 4 Simulation results and discussion

The proposed approach is tested using uncapacitated exam timetabling problems, which is a constrained combinatorial optimization problem and will be discussed in the section 4.1.

##### 4.1 Educational exam timetabling problem

Examination timetabling problem is a problem of assigning a given number of exams to available time slots in such a way that exams with common students should be arranged in different slots. There are two types of exam timetabling problems, capacitated and uncapacitated. In capacitated exam time tabling problem the capacity of the exam room will be taken into consideration whereas

---

```

Input  $p_f, p_c, D$ 
if  $rand \leq p_f$ 
    for  $i = 1 : n$ 
        if  $rand \leq p_c$ 
            if updating  $x_i(k)$  using equation 3 doesn't make  $x_i$  infeasible
                then update  $x_i(k)$  using equation 3
            end if
        end if
    end for
else
    for  $i = 1 : n$ 
        if  $rand \leq p_c$ 
            if updating  $x_i(k)$  using equation 4 doesn't make  $x_i$  infeasible
                then update  $x_i(k)$  using equation 4
            end if
        end if
    end for
end if

```

---

**Table 3.** updating of an ordinary prey  $x_i$ 

in the uncapacitated case the room capacity is not considered. With the constraint being there is no clash between exams with common students, the objective will be to spread the exam in such a way that students will have good gap of slots between exams. i.e. The objective is to minimize a penalty for two exams with common students scheduled with smaller number of slots in between and is given by equation 5.

$$f(x) = \frac{1}{S} \sum (floor(2^{5-|x_i-x_j|})) N_{ij} \quad (5)$$

where  $S$  is total number of students,  $N_{ij}$  is number of students taking exam  $i$  and exam  $j$ , and the  $floor$  function rounds up the result to the smallest and nearest integer.

Carter benchmark problems are well known uncapacitated exam timetabling benchmark problems [17]. Five benchmark problems from Carter uncapacitated timetabling problems found in [18] are selected. These five benchmark problems are STA83, UTE92, TRE92, CAR92 and UTA92. The properties of these data sets or problems are given in table 4.

#### 4.2 Simulation results

The simulation is done using *MATLAB R2011b* on *Intel Core™ i5-2400 CPU @ 3.10GHz* with 32 bit operating system desktop machine. The algorithm parameters are tuned based on recommendations on [13] and [15]. The total number of initial solutions was set to be 20, with probability of follow-up, probability of change for local search and probability of change for exploration being 0.75, 0.2

Problems	Exams Number	Student Number	Enrolment	Density	Time slots
STA83	139	611	5751	0.14	13
UTE92	184	2750	11793	0.08	10
TRE92	261	4360	14901	0.18	23
CAR92	543	18419	55522	0.14	32
UTA92	622	21266	58979	0.13	35

**Table 4.** Properties of the benchmark problems [18]

and 0.6, respectively. Furthermore, the number of random directions for local search,  $m$ , is set to be 100. The maximum number of iteration, which was set to be 50, was used as a termination criterion. In addition the number of best prey was increased and set to be 8, to increase the local search behavior of the algorithm.

In order to generate feasible initial solutions, saturation degree was used with minimum penalty. The simulation is done 30 times in order to compute the mean and standard deviation of the results and is reported in table 5.

Problems	STA83	UTE92	TRE92	CAR92	UTA92
Best result	151.1554	24.0767	9.6676	5.9065	5.0671
Mean	154.3231	24.3879	9.7568	6.0203	5.1418
Standard deviation	2.0369	0.1692	0.0496	0.0824	0.0631
Average CPU time	69.5811	203.5500	1331.1000	1145.0000	44448.0906

**Table 5.** Simulation results of PPA on the benchmark problems

Exam timetabling has been one of the application of combinatorial optimization problems and different studies has been conducted especially using Carter benchmark problems. Table 6 summarize the solution ranges of the selected five problems in literature.

Problems	Best results using PPA	Best results from literature		Worst results from literature	
		Results	References	Results	References
STA83	151.1554	157.3	[19]	168.3	[20]
UTE92	24.0767	24.4	[21]	29	[22]
TRE92	7.9679	7.87	[23]	10	[24]
CAR92	5.9065	3.9	[25]	6.2	[17]
UTA92	3.5671	3.1	[25]	4.2	[24]

**Table 6.** Solutions for the five benchmark problems from the literature

The simulation results shows that, PPA outperforms existing results in two of the problem instances, in STA83 and UTE92, whereas it performs in between the best and worst results in TRE92 and CAR92. However, the results in UTA92 is worst compared to results in literature. Furthermore, the results achieved using PPA is stable as they have less variability with smaller standard deviation except for STA83. Hence, it is possible to say that prey predator algorithm has performed in a promising way whereas its behavior for different parametric value can be put as possible future work along with a detailed study on the strength and the weakness of the algorithm.

## 5 Conclusion

This paper extends the recently introduced metaheuristic algorithm, prey predator algorithm, to be used for constrained combinatorial optimization problem. The algorithm is inspired by the interaction between a predator which runs after its prey and how the prey tries to survive in this situation. It is a swarm based algorithm, in which updating a solution is done based on a given direction, which may result moving the solution out of the feasible region for constrained problems. Hence for the case of constrained combinatorial optimization problems the local search on the updating process was modified based on an algorithm parameter called probability of change,  $p_c$ , which replaced two of the step length parameters of the standard prey predator algorithm. The updating is done componentwise for each entries of the solution vector. To test the approach five problem instances were selected from the uncapacitated exam timetabling. Uncapacitated exam timetabling problem is one of the well known problem in the domain of combinatorial optimization problem since its introduction in mid 1990's. Hence, many research output has been published and it is an ideal problem domain to test the approach and compare with the wide set of results in literature. From the simulation results the proposed approach outperforms previous methods in two of the problem instances whereas it is comparable in the other two and it produces the worst solution in one of the problem instance. Based on *no free lunch theorem*, there is no superior algorithm for combinatorial optimization problems but on average the performance of one algorithm is the same as any other algorithm over the set of all problems, hence, having different solution approach is a good idea. Parameter tuning, constraint handling and dealing with even higher dimensions can be possible future works with testing the approach in problems from different domain and a detailed comparison with other methods.

## Bibliography

- [1] A. Schrijver. On the history of combinatorial optimization (till 1960). In K. Aardal; G. L. Nemhauser and R. Weismantel, editors, *pp. 1 - 68, Discrete Optimization, volume 12 of Handbooks in Operations Research and Management Science*. Amsterdam, The Netherlands: North - Holland, 2005.
- [2] M. Ayob, A. Malik, S. Abdullah, A.R. Hamdan, G. Kendall, and R. Qu. Solving a practical examination timetabling problem: a case study. In O. Gervasi and M. Gavrilova, editors, *pp. 611 - 624, ICCSA 2007, Part III, LNCS, vol. 4707*. Springer, Heidelberg, 2007.
- [3] D. C. Little John, Katta G. Murty, Dura W. Sweeney, and Caroline Karel. An algorithm for the traveling salesman problem. *Operations Research*, 11:972 - 989, 1963.
- [4] R. E. Korf. A new algorithm for optimal bin packing. In *In Eighteenth national conference on Artificial intelligence*, pages 731–736, Edmonton, Alberta, Canada, 2002.
- [5] X.-S. Yang. *Nature-Inspired Metaheuristic Algorithms*. Luniver Press, University of Cambridge, United Kingdom, second edition, 2010.
- [6] B. Bahmani-Firouzi, S. Sharifnia, R. Azizipanah-Abarghooee, and T. Niknam. Scenario-based optimal bidding strategies of gencos in the incomplete information electricity market using a new improved prey-predator optimization algorithm. *IEEE SYSTEMS JOURNAL*, pages 1 – 11, 2015.
- [7] W. Dai, Q. Liu, and T. Chai. Particle size estimate of grinding processes using random vector functional link networks with improved robustness. *Neurocomputing*, 169:361 – 372, 2015.
- [8] S. L. Tilahun and H. C. Ong. Fuzzy graph representation of bus timetabling problem and its solution method using prey-predator algorithm. 2012.
- [9] N. Hamadneh, S. L. Tilahun, S. Sathasivam, and H. C. Ong. Prey-predator algorithm as a new optimization technique using in radial basis function neural networks. *Research Journal of Applied Sciences*, 8:383–387, 2013.
- [10] S. L. Tilahun and H. C. Ong. Comparison between genetic algorithm and prey-predator algorithm. *Malaysian Journal of Fundamental and Applied Sciences*, 9:167–170, 2013.
- [11] S. L. Tilahun and J. M. T. Ngnotchouye. Prey predator algorithm with adaptive step length. *International Journal of Bio-Inspired Computing [in press]*.
- [12] S. L. Tilahun, H. C. Ong, and J. M. T. Ngnotchouye. Extended prey-predator algorithm with a group hunting scenario. *Advances in Operations Research*, vol. 2016:14 pages, 2016.
- [13] S. L. Tilahun. *Prey predator algorithm: A new metaheuristic optimization approach*. PhD thesis, School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia, April 2013.
- [14] B. Korte and J. Vigen. *Combinatorial Optimization: Theory and algorithm*. Springer-Verlag, Germany, second edition, 2002.

- [15] S. L. Tilahun and H. C. Ong. Prey predator algorithm: A new metaheuristic optimization algorithm. *International Journal of Information technology & Decision Making*, 14:1331 – 1352, 2015.
- [16] P. Ross, E. Hart, and D Corne. Some observations about ga-based exam timetabling. In E.K.Burke and M. Carter, editors, *LNCS 1408, Practice and Theory of Automated Timetabling II: Second International Conference, PATAT*, 1997.
- [17] M. W. Carter, G. Laporte, and S. Y. Lee. Examination timetabling algorithmic strategies and applications. *Journal of the Operational Research Society*, 47:373 – 383, 1996.
- [18] M. W. Carter, G. Laporte, and S. Y. Lee. Benchmark data sets in exam timetabling.
- [19] L. G. Merlot, N. Boland, B. Hughes, and P. Stuckey. A hybrid algorithm for the examination timetabling problem. In E. Burke and P. Causmaecker, editors, *Practice and Theory of Automated Timetabling IV, 207-31*. Springer, Heidelberg, 2003.
- [20] E. K. Burke and J. P. Newall. Enhancing timetable solutions with local search methods. In E. K. Burke and P. De Causmaecker, editors, *Practice and theory of automated timetabling: Selected papers from the fourth international conference*, pages 195 – 206. , Springer-Verlag, 2003.
- [21] M. Caramia, P. DellOlmo, and G. F. Italiano. New algorithms for examination timetabling. In S. Naher and D. Wagner, editors, *WAE 2000 - LNCS*. Springer, Heidelberg, 2001.
- [22] G. M. White and B. S. Xie. Examination timetables and tabu search with longer-term memory. In E.K. Burke and W. Erben, editors, *PATAT 2000, LNCS Vol. 2079, 85 - 103*. Springer, Heidelberg, 2001.
- [23] F. C. Weng and H. B. Asmuni. An automated approach based on bee swarm in tackling university examination timetabling problem. *International Journal of Engineering and Computer Science (IJECS)*, 13:8 – 23, 2013.
- [24] L. D. Gaspero and A. Schaerf. Tabu search techniques for examination timetabling. In E.K. Burke and W. Erben, editors, *Practice and Theory of Automated Timetabling III Third International Conference, PATAT 2000, LNCS 2079, 104 - 117*. Springer, Heidelberg, 2001.
- [25] N. R. Sabar, M. Ayob, and G. Kendall. Solving examination timetabling problems using honeybee mating optimization (etp-hbmo). In *Multidisciplinary International Conference on Scheduling : Theory and Applications (MISTA 2009)*, 2009.

## K-Mean Algorithm for Finding Students' Proficiency with a Framework's Item Examination

Nongnuch Ketui<sup>1</sup>, Kanitha Homjun<sup>2</sup>, and Prasert Luegkhong<sup>3</sup>

<sup>1</sup>Computer Science Program, Faculty of Sciences and Agricultural Technology  
Rajamangala University of Technology Lanna, Nan, Thailand

<sup>2</sup>Information Technology Program, Faculty of Sciences and Agricultural Technology  
Rajamangala University of Technology Lanna, Nan, Thailand

<sup>3</sup>College of Integrated Science and Technology,  
Rajamangala University of Technology Lanna, Chiang Mai, Thailand  
 {nongnuchketui, kanithaasc, prasert}@rmut1.ac.th

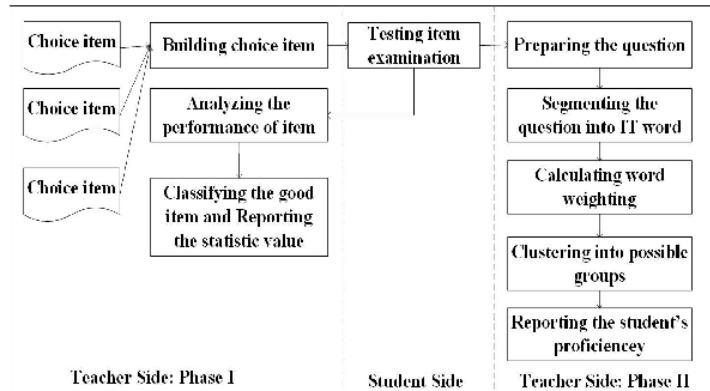
**Abstract.** The effective system of item examination is used to assess the achievement students while the proficiency of students can be classify theirs real aptitude. In this paper, we introduce a framework of item examination which having two phases; (1) building and analyzing item examination and (2) finding the student's proficiency. To analyze the online item, a number of experiments are conducted using 1,000 items of Information Technology (IT) which having four choices in each item and evaluated by two groups of students. The questions that having the right answer are segmented into a small unit (IT word), assigned with the frequency weighting, and clustered into six groups of IT aptitudes. Based on measures of KR-20 called the reliability value and evaluated the performance of item examination three factors; difficulty index, discrimination values, and distracter efficiency. While the student's proficiency is explored by assigning IT keywords with the standard weighting (TF/IDF) in order to clustering with K-Mean clustering algorithm. The experimental results show that the reliability is good level (0.98) and the performance of students are quite good in information technology.

**Keywords:** item examination, student proficiency, K-Mean algorithm

### 1 Introduction

Online exam bank is useful for the purposes of assessing the achievement of students and evaluating educational innovations [2]. While the proficiency of student can directly measure with the testing score, a GPA is the normal way to cluster or rank the efficiency of student. A variety of data mining techniques for analyzing how students interact with Intelligent tutoring systems (ITSs), including methods for handling hidden state variables (HMM), and for testing hypotheses were surveyed in [1]. A Naive Bayes classifier was used to extract patterns using the Data Mining Weka tool [9]. A K-Mean clustering algorithm was applied to the level of students' performance [3, 8]

K-Mean for Finding Students' Proficiency with a Framework's Item Exam

**Fig. 1.** A Framework of Item Examination

The purpose of this work, we introduce the framework of item examination in Section 2. Experimental settings are shown in Section 3. Experimental results and discussions are described in Section 4. Finally, conclusion and future work are given in Section 5.

## 2 A Framework of Item Examination

This section presents our framework of item examination which is composed of two main phases. The performance of choice items and student's proficiency are analyzed in teacher side while the student's examination works on the student side.

### 2.1 Phase I: Building and Analyzing Item Examination

- Build the item examination; the multiple choice items are collected from published websites and builded the item exam.
- Test the item examination by IT students; we got the score as the input is used to calculate the statistic values.
- Analyze the performance of item examination; the items are measured by considering importance level of three statistic values.
- Report the statistics; the good items are classified and the summary is reported.

### 2.2 Phase II: Finding the Students' Proficiency

- Prepare the questions; we use the correct questions (From Phase I) as the input which each student having the right answer.

### K-Mean for Finding Students' Proficiency with a Framework's Item Exam

- Segment the running question and classify into IT keywords; a Thai running question is segmented into a sequence of tractable units and selected keywords.
- Calculate the Term Frequency/Inverse Document Frequency: TF/IDF [7] as the weighting factor and assign to each word.
- Cluster the IT words into groups.
- Report the number of student's proficiency.

## 3 Experimental Settings

### 3.1 The Performance of Item Examination

This work utilizes Information Technology (IT) online examination in Thai which comprises 1,000 items with 4 choices in many categories such as teacher assistance examination<sup>1</sup> <sup>2</sup>, government examination<sup>3</sup>, and company examination<sup>4</sup>. To examine the performance of item exam, we divided the computer science students into two groups; (1) twenty students are the clever student which having GPA over than 2.50 and (2) twenty students are the poor students which having GPA less than 2.50.

The experiment aims to find out the performance of IT item exam. We use KR-20 [6] to investigate the reliability of IT item exam. The quality of a multiple choice item are considered with three values; (1) the difficulty index, (2) the discrimination values, and (3) the distracter efficiency. Here, the examination quality measurement standards describes the meaning of difficulty index ranges and the discrimination value levels [4]. Difficulty Index is a measure of the proportion of students or examinees who answered the item correctly. The discrimination value measures the quality of an item that is able to distinguish between examinees who are knowledgeable and those who are not. The distracter efficiency is used to measure the performance of the incorrect response options. Here, the difficulty index is between 0.20 and 0.80 while the discrimination value is higher than 0.20. However, the distracter efficiency should be greater than 0.05.

### 3.2 Exploring the Proficiency of Student

In this experiment, we selected the questions that classify into seven groups; introduction of computer (M1), database (M2), communication and network (M3), technology internet (M4), application (M5), programming (M6), information technology (M7). Given each running question, the question was segmented by Thai E-Class [10] and then assigned the weighting factor of word with TF/IDF [7]. Then, we clustered the IT vocabularies into the possible groups by a standard method as K-Mean algorithm [5]. We iterated 10 times for clustering.

<sup>1</sup> [www.kruwandee.com](http://www.kruwandee.com), [www.krupuchoi.com](http://www.krupuchoi.com), [www.trueplookpanya.com](http://www.trueplookpanya.com)

<sup>2</sup> [www.kruchiangrai.net](http://www.kruchiangrai.net), [www.krootoey.com](http://www.krootoey.com)

<sup>3</sup> [www.goosiam.com](http://www.goosiam.com), [www.konsorb.com](http://www.konsorb.com), [www.thailocalmeet.com](http://www.thailocalmeet.com)

<sup>4</sup> <http://tutor.msu.ac.th>, <http://lumtian.blogspot.com/2008/07/40.html>

K-Mean for Finding Students' Proficiency with a Framework's Item Exam

**Table 1.** The difficulty index of the number of online item exam.

Difficulty index	Meaning	#Item
0.80 - 1.00	Very easy	412
0.60 - 0.79	Easy	279
0.40 - 0.59	Fair	139
0.20 - 0.39	Hard	88
0.00 - 0.19	Very hard	82

**Table 2.** The discrimination value of the number of online item exam.

Discrimination value	Meaning	#Item
0.60 - 1.00	Very good	76
0.40 - 0.59	Good	161
0.20 - 0.39	Fair	269
0.10 - 0.19	Low (need to improve)	173
0.00 - 0.09	Very low (need to discard)	160
Less than 0.00	NOT classify (need to discard)	161

This experiment aims to find the student's proficiency by considering the occurring words in the questions. Normally, we use the maximum score of IT groups as the aptitude of students. Sometime, we do not know the real aptitude since each question may be composed of many IT keywords. So that, we should consider the occurring frequency word as word weighting and cluster IT words into the possible group. Top-20 students which having the maximum score is used to set in this experiment.

## 4 Result and Discussion

### 4.1 The Performance of Item Examination

This section investigated the reliability of IT online item examination and measured by three statistic values. The result of KR-20 calculation, the average of the student's score equals to 661.60 (full score is 1,000) and the variance of the entire online item exam is 11,565.44. The reliability of IT online item exam KR-20 is good level (0.98) since the reliability value is normally over than 0.70. The difficulty index of online item examination are shown in Table 1. The very easy level of item has the maximum items (412 of 1,000 items) while the range of difficulty index between 0.00 to 0.19 and 0.20 to 0.39 get the values as a same (82 and 88, respectively). The number of easy item is greater than the fair item. The average of the difficulty index equals to 0.66 so that the overall of online item is easy level.

Table 2 displays the discrimination values of the number of online item exam. We found that almost items classify the examinees to 269 items in the fair level (0.20 - 0.39) while the best classification equals to 76 items. The number of discarded items is 321 items (160+161) since the discrimination value is less

## K-Mean for Finding Students' Proficiency with a Framework's Item Exam

**Table 3.** Comparison of testing score and K-Mean clustering algorithm.

Top-20 Students	Methods	IT Categories							Total
		M1	M2	M3	M4	M5	M6	M7	
1	Testing Score	323	29	142	81	184	19	17	795
	K-Mean	-	1	4	1	1	4	21	32
2	Testing Score	321	29	138	74	186	18	17	783
	K-Mean	-	1	4	1	10	1	15	32
3	Testing Score	314	29	137	73	183	19	16	771
	K-Mean	-	1	4	5	1	1	20	32
4	Testing Score	318	20	140	86	169	19	17	769
	K-Mean	-	4	4	1	1	1	21	32
5	Testing Score	301	29	142	82	178	19	17	768
	K-Mean	-	1	4	1	1	4	21	32
6	Testing Score	326	21	126	72	187	19	9	760
	K-Mean	-	9	1	4	1	2	15	32
7	Testing Score	309	22	135	74	180	18	16	754
	K-Mean	-	1	1	4	4	21	1	32
8	Testing Score	318	27	110	69	187	17	4	732
	K-Mean	-	1	1	2	1	10	17	32
9	Testing Score	292	20	119	67	174	16	11	699
	K-Mean	-	1	10	2	1	3	15	32
10	Testing Score	276	14	109	85	159	15	12	670
	K-Mean	-	1	8	1	1	1	20	32
11	Testing Score	285	12	95	49	183	15	5	644
	K-Mean	-	2	8	1	1	19	1	32
12	Testing Score	253	26	120	51	164	12	13	639
	K-Mean	-	1	4	1	1	4	21	32
13	Testing Score	276	16	100	47	168	18	6	631
	K-Mean	-	2	1	1	7	12	9	32
14	Testing Score	269	17	108	62	150	16	9	631
	K-Mean	-	1	4	8	2	1	16	32
15	Testing Score	250	25	114	49	153	15	13	619
	K-Mean	-	4	1	1	4	1	21	32
16	Testing Score	265	13	84	53	149	11	7	582
	K-Mean	-	2	8	1	1	19	1	32
17	Testing Score	248	14	65	43	131	11	2	514
	K-Mean	-	1	12	1	6	10	2	32
18	Testing Score	229	16	72	39	142	15	1	514
	K-Mean	-	6	2	1	1	7	15	32
19	Testing Score	240	14	75	39	123	11	2	504
	K-Mean	-	2	21	1	1	1	6	32
20	Testing Score	179	13	59	43	105	9	4	412
	K-Mean	-	2	12	4	5	1	8	32

than 0.09. 173 items need to be improved for online item exam. To consider the average of the discrimination value, we need to improve IT online item since the

### K-Mean for Finding Students' Proficiency with a Framework's Item Exam

discrimination value equals to 0.17. The good distracters get 450 items while 410 items are the good online items. Since IT online item is the example or guideline of examination and the students are computer science program, the items are very easy. The other reason, we found that the used IT items are not up-to-date so that many items need to be improved.

#### 4.2 Exploring the Students' Proficiency

Table 3 shows the comparison of the testing score (count the frequency of corrected items) and K-Mean clustering algorithm. The number questions of M1-M7 are 400, 40, 165, 96, 256, 26, 17, respectively. In this case, we use thirty-two IT keywords in all questions. The value of testing score is the number of correct items while each row of the K-Mean clustering algorithm displays the cluster size of IT keywords. For introduction of computer (M1), since we can not classify this group into the others, we should consider all words that occurring in the questions (M2-M7) only.

For the testing score, we considered the average of corrected answer scores in each groups. The result shown that the database (M2) has the corrected items 20.30 (0.52%), the communication and network (M3) equals to 109.50 scores (0.66%), the technology internet (M4) gets 61.90 (0.64%), the application (M5) is 162.75 (0.63%), the programming (M6) has 15.60 (0.74%), and the information system (M7) achieves 9.90 (0.58%). Almost students get the highest score in the application (M5) while the database group has the lowest score. While the K-Mean clustering algorithm, it can be find the most student's proficiency is to be clever in information system (M7) (13 persons), the programming (M6) (4 persons), and communication and network (M3) (3 persons).

To compare the result of testing score and K-Mean clustering algorithm, this student groups have the aptitude in the programming and information system while the institute should practically increase in the database knowledge. To conclude the effect of TFIDF weighting as factor on the K-Mean clustering, can affect to get the possible student's proficiency highly.

## 5 Conclusion and Future Works

we introduce a framework of item examination. To analyze the online item, 1,000 items with 4 choices of IT questions are conducted. Based on measures of KR-20 called the reliability value and evaluated the performance of item examination three factors; difficulty index, discrimination values, and distracter efficiency. While the student's proficiency is explored by assigning IT keywords with the TF/IDF in order to clustering with K-Mean clustering algorithm. The experimental results show that the reliability is good level (0.98) and the aptitude of student is on the information technology and programming. The K-Mean clustering algorithm can help to find the possible students' proficiency with the word weighting and the distance of theirs relations. In the future works, we will use the bigger dataset and improve this item examination with adaptive-based approach for selecting the appropriated questions to each student.

### K-Mean for Finding Students' Proficiency with a Framework's Item Exam

## References

1. Beal, C. R., Cohen, P.R.:Temporal Data Mining for Educational Applications. In Ho, T.-B. and Zhou, Z.-H. (Eds.) Trends in Artificial Intelligence: 10th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2008, LNAI, pp.66–77 (2008) Springer
2. Choppin, B.: Developments in Item Banking. In the first European Contact Workshop on Monitoring National Standards of Attainment in Schools. pp.216–234 (1976)
3. Ganesh, S. H., Felciah, M. L. P., Shafreenbanu, A. K.: Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm, In Proc. of the IEEE International Conference on World Congress on Computing and Communication Technologies (WCCCT), pp.200–202 (2014)
4. Gronlund, N.E., Linn, R.L.: Measurement and evaluation in teaching (6th ed.). New York: MacMillan. (1990)
5. Hartigan, J.A., Clustering algorithms. John Wiley and Sons, Inc. (1975)
6. Kuder, Frederic G., Richardson, M. W.: The theory of the estimation of test reliability. Psychometrika. 2(3), 151–160 (1937)
7. Luhn, Hans Peter: A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of research and development (IBM). 1(4): 315 (1957)
8. Oyelade, O. J., Oladipupo, O. O., Obagbuwa, I. C.: Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance. International Journal of Computer Science and Information Security. vol.7(1), 292–295 (2010)
9. Taziz, A. Abdul, Ismail, N. Hafieza, Ahmad, Fadhilah: Mining Students' Academic Performance. Journal of Theoretical and Applied Information Technology. vol.53(3), 485–495 (2013)
10. Tongtep, N.,Theeramunkong, T.: Multi-stage automatic NE and POS annotation using pattern-based and statistical-based techniques for Thai corpus construction. IEICE Transaction on Information and Systems. vol.E96-D(10), 2245–2256 (2013)

## Acknowledgement

This work was supported by under the Research and Development Institute (RDI) of Rajamangala University of Technology Lanna, Thailand under Project of Hands-on Research and Development Number 58UR-07. We would like to thank all lecturers and students at Science Department, RMUTL Nan for all kind helps.

## Building a Semantic Ontology for Virtual Peers in Narrative-Based Environments

Ethel Chua Joy Ong<sup>1</sup>, Danielle Grace Consignado,  
Sabrina Jane Ong and Zhayne Chong Soriano

Center for Language Technologies, De La Salle University, Manila, Philippines  
[ethel.ong@de lasalle.ph](mailto:ethel.ong@de lasalle.ph)<sup>1</sup>

**Abstract.** Narrative-based environments utilize various forms of knowledge to provide an interactive space for the learner and the virtual agent to collaborate in accomplishing the learning goals. In this paper, we present the design of a semantic ontology that provides the necessary domain-based conceptual knowledge to allow a virtual peer to engage in storytelling as a form of exchange with the learner. We then show how the ontology was utilized to support the virtual peer in performing its tasks, which include generating interactive stories that teach about appropriate social behavior, and engaging in a text-based dialogue with the learner.

**Keywords:** Semantic Ontology, Virtual Peer, Story Generation, Dialogue Generation, Commonsense Knowledge

### 1 Introduction

Narrative-based environments have utilized virtual agents to serve various roles, such as playmates or learning companion [1, 2], teachable peer [3], facilitator, and tutor [4]. A virtual playmate has a similar developmental age and understands the world in similar ways as the child-user [5]. Its goal is to stimulate the child's learning by collaborating on shared tasks and even competing for producing quality work output. A teachable peer, on the other hand, reverses the concept of peer learning by having the child-user take on the tutor role, thus advocating learning by teaching. A virtual facilitator shows the student around the virtual learning environment, alerts him/her to what is new or relevant, and assists in navigating through the lesson [6]. As a tutor, the virtual agent monitors the performance and gives appropriate feedback during all pedagogical stages (acquisition, application and assessment).

For virtual agents to be effective in carrying out their roles, they must be given domain knowledge about the lessons to be covered, pedagogical knowledge about teaching strategies and remediation, linguistics knowledge to provide appropriate feedback and responses, skills for intervening at the proper time during a learning session, and even personality to motivate the student to begin, continue and complete the required learning activities. For narrative-based environments, the agents must also be given conceptual knowledge about concepts on everyday things that the learn-

ers are familiar with, and which they can use when engaging learners in storytelling as a form of pedagogical strategy or intervention.

In this paper, we present the design of a semantic ontology that provides the necessary domain-based conceptual knowledge to allow a virtual peer to engage in storytelling as a form of exchange with the learner. Interaction with virtual peers can be more natural and engaging if the peer is made to exhibit human-like behavior. In our study, we advocate storytelling as the means to achieve this natural interaction. People naturally engage in storytelling as a form of communication to narrate about daily life experiences, share beliefs and exchange information. Furthermore, storytelling remains extensively used in modern classrooms to enhance the learning experience of children [7].

The use of semantic ontology to provide the necessary domain-based conceptual knowledge to a story generation system has been explored in [8, 9, 10, 11]. The semantic ontology contains commonsense knowledge about the world that enable people to understand each other. Giving a similar body of knowledge to software agents will enable computer systems that can understand and generate text in natural language to be developed.

The paper also details how the semantic ontology was utilized to support two virtual peers in performing their tasks. Ellie, presented in Section 2, is a virtual peer that teaches appropriate social behavior to children with autism. Carla, presented in Section 3, is a conversational peer who engages the learner in a text-based dialogue. The paper ends with a discussion of the lessons learned from our study.

## 2 The Semantic Ontology of Ellie, the Virtual Social Peer

Ellie is a virtual peer designed for an interactive storytelling environment that teaches children with autism about proper social behavior. The semantic ontology built for Ellie contains assertions in the form of binary relations about concepts, events and their relationships that are relevant to the themes of the storytelling system. Currently, these themes revolve around teaching social values on *waiting for turn, greetings, tidying up, and sharing*.

A primary source of commonsense knowledge for the semantic ontology of Ellie is ConceptNet [12]. It is a publicly available, large semantic graph containing nodes of concepts that represent words or phrases in natural language, and edges that represent the relationships between two concepts. Two reasons, however, necessitate the need to build a separate ontology for Ellie.

First, ConceptNet has been populated by crowdsourcing data through the Open Mind Common Sense project [13]. As such, assertions that are not appropriate for the target audience, specifically children diagnosed with mild autism, abound. Only concepts relevant to the story themes were extracted for use by Ellie.

Second, the interactive stories narrated by Ellie adopted the Social Story structure developed by Carol Gray [14]. In a Social Story™, a situation is described in terms of relevant social cues, perspectives and common responses in order to share accurate social information in a patient and reassuring manner. The rationale behind every

desirable action, social norms and reaction to common situations is also continuously explained clearly and explicitly as facts to prepare the learners for social interactions.

## 2.1 Conceptual Relations

Given the previous requirements, the conceptual relations in the semantic ontology of Ellie were categorized into two – *classic* relations and *social* relations. *Classic* relations are those used to describe common concepts and events. They include the following relations adapted from ConceptNet: *usedFor*, *locatedAt*, *isFor*, and *can*. Table 1 provides a brief description and example assertions for these.

**Table 1.** Classic semantic relations to describe concepts and events.

Relations	Description	Example Assertions
usedFor	Indicates the purpose of an object	usedFor(ball, catch) usedFor(swings, play) usedFor(book, read)
locatedAt	Specifies where an item may be situated in the virtual world	locatedAt(swings, playground) locatedAt(teachers, school) locatedAt(book, school)
isFor	Describes an activity that may take place in the specified location	isFor(playground, everyone to share and have fun)
can	States what the specified story character can do	can(children, say "Hello!") can(teachers, smile politely)

*Classic* relations enable Ellie to generate story text that contain descriptions of objects as well as to provide possible expectations in a new situation. Consider the sample assertions in Table 1. Ellie can use the *usedFor* relations to describe the purpose of an object in the virtual story world. For example, from the *usedFor(book, read)* assertion, Ellie can generate "*You can read a book.*". The *locatedAt* relation can be used to prepare the child for what he/she may find in a given virtual world location, e.g., "*Sometimes you meet new people in places like the school. You might encounter teachers or other children.*"

*Social* relations, on the other hand, contain positive assertions that embody a “social sense” to describe commonsense knowledge relevant to social interactions [9]. Table 2 provides a brief description and example assertions with “social sense”.

*Social* relations are used by Ellie as a means of preparing the learner to engage in social interactions. Consider the sample assertions in Table 2. Assuming the setting of the story is still the school, Ellie can use the *mayFeel* and *mayThink* relations to let the learner be aware of his/her possible feelings and thoughts, and to assure that these are acceptable, e.g., "*Sometimes you may feel shy and think about leaving. That's okay.*"

The combined used of *classic* and *social* relations allow Ellie to also generate story text that teaches about social behavior. For example, the *can* relations in Table 1 can

lead to the generation of a story text that informs the learner on how to respond in the given situation, e.g., "You can say 'Hello!'"

**Table 2.** Semantic relations with "social sense" to describe social concepts.

Relations	Description	Example Assertions
mayFeel	Relates what the user might feel at a certain event	mayFeel(school, shy) mayFeel(school, nervous) mayFeel(playground, irritated)
mayThink	Suggests what the user may think at a certain event	mayThink(school, about leaving) mayThink(school, would rather play)
sharedBy	Defines who or what can share an object	sharedBy(ball, children) sharedBy(book, classmates)
expectedTo	Conveys what the user may expect at a given location	expectedTo(playground, see other people) expectedTo(playground, see other children run around)
whenUntidy	Describes the state of an object when it is untidy	whenUntidy(book, lying on the floor) whenUntidy(toys, scattered on the ground)

## 2.2 Generating Story Text

Children with autism often have difficulty generalizing what they have learned in one social setting to another [15]. The availability of variances in the semantic ontology, combined with templates, allows for the generation of stories with the same theme but situated in different settings, to help prepare learners for many possible situations. For example, the presence of two *expectedTo* assertions can lead to the generation of story text that may prepare the learner to *see other people in the playground*, or to *see other children run around in the playground*. This is similar to the notion presented by the Social Story Idea Toolkit [9], which also utilizes ConceptNet [12] as a resource to aid writers of Social Stories™.

An example template that is used to generate the body of the story is shown in Listing 1. The body of a Social Story™ contains sentences that add further description to the topic sentence that was stated in the story's introduction. The use of template-based text generation technique is necessary to ensure that the resulting sentences are appropriate for the target learners, as validated by special education teachers.

A template contains tags used to define the category of data that can be used as values for the tags. Tags in angle brackets ( $\langle \rangle$ ) are used to retrieve story world elements, such as the current location and the name of the non-playing character that the learner, who serves a "player" role, is interacting with. Tags in square brackets ( $[]$ ) are used to query the ontology.

**Listing 1.** A template to generate the story body.

---

Sometimes you meet new people in places like the <location>. You might encounter [locatedAt] or [locatedAt]. They [can] or [can]. Sometimes you might feel [mayFeel] and think [mayThink]. That's okay.

---

For example, if the <location> tag returns “school”, then the [locatedAt] tag is used to query the ontology with “locatedAt(?, school)”. Using the sample assertions in Tables 1 and 2, the resulting concept is “teachers”.

Depending on the available assertions, a set of candidate concepts may be returned by the ontology. For example, the [mayFeel] tag, which queries the ontology with “mayFeel(school, ?)”, receives the candidate concepts “shy” and “nervous”. In such a situation, the story generator randomly selects from the candidate concepts to instantiate a given template.

The template in Listing 1 may lead to the generation of any one of the two sample story text shown in Listing 2, with tags replaced by underlined words and phrases.

**Listing 2.** Sample story text generated from the template given in Listing 1.

---

Sometimes you meet new people in places like the school. You might encounter fellow classmates or teachers. They say hello or smile politely. Sometimes you might feel shy and think about leaving. That's okay.

---

Sometimes you meet new people in places like the store. You might encounter shoppers or staff members. They can say hello or smile politely. Sometimes you might feel shy and think that you should ignore them. That's okay.

---

Ellie also uses predefined dialogue templates to explain the rationale behind a good decision that the learner has made in the interactive storytelling environment, a sample of which is shown in Listing 3 for the *Greetings* theme.

**Listing 3.** A dialogue template that is used when a good decision has been made.

---

Other people can be the first one to greet you. You can initiate greetings too. People appreciate it when you greet them. Let's practice greetings! You can [can]. You can also [can] or [can].

---

All the discussions thus far portray Ellie as having the role of a teacher who adopts a narrative format to describe social situations and teach about proper behavior. But placed in an interactive storytelling environment, Ellie occasionally switches her role to a facilitator to engage the learner in a decision-making activity. Specifically, Ellie presents options on how the learner would want the story to proceed, as shown in Listing 4. The underlined phrases in the second type of decision-making (presentation of the story problem) are concepts derived from the ontology.

**Listing 4.** Decision-making points in the interactive story world.

---

Presentation of Tasks	What do you want to do now? Option 1: Look around Option 2: Play with the slide
Presentation of the Story Problem	When you want to play with the slide, others might want to play with the slide too. What should you do? Option 1: Push others away Option 2: Wait in line

---

### 3 The Semantic Ontology of Carla, the Conversational Peer

Carla is a virtual peer that is integrated to a learning environment for reading short stories and answering reading comprehension exercises. Carla is designed to engage the learner (children who are 8-10 years old) in a text-based dialogue in an attempt to shift the learner's negative affect to one that is positive. Intelligent Tutoring Systems such as AutoTutor [4] have been enhanced to take into consideration the learner's affect state, which has an effect on his/her learning performance. Carla's design posits that a positive affect can motivate the completion of the required learning activity, in this case, that of answering reading comprehension exercises.

#### 3.1 Populating the Semantic Ontology

Carla uses a semantic ontology to provide the possible topics of discourse that it can use during its conversation with the learner. The commonsense knowledge comprising this ontology has been directly sourced from ConceptNet 5 [16]. Of the 24 relations in ConceptNet, only six are presently being used. These are enumerated in Table 3 with the corresponding question that can be answered by the assertions of the given relation.

Using the questions enumerated in Table 3 as the basis, some assertions extracted from ConceptNet were replaced with more suitable relations, such as replacing *hasProperty(red, one of primary color)* with *definedAs(red, primary color)*, and *hasProperty(ice cream, made of fruit)* with *madeOf(ice cream, fruit)*. The availability of the most suitable relation for a pair of given concepts is important in order to use the correct sentence template during dialogue generation.

The current iteration of the semantic ontology of Carla consists of over 600 commonsense assertions that are age-appropriate to the target audience and that are in the domains of everyday objects and activities at home and at school, food and sports. These domains were selected to complement the reading materials.

**Table 3.** Semantic relations in the ontology of Carla. ( $C_n$  refers to Concept  $n$ )

Relation	Answerable Question	Sentence Template
isA	What kind of a thing is it?	<C1> is a kind of <C2>
definedAs	How is it defined?	<C1> is [a/an/the] <C2>
hasProperty	What property/ies does it possess?	<C1> is <C2>
hasA	What feature/s does it have?	<C1> has <C2>
madeOf	What is it made of?	<C1> is made of <C2>
locatedAt	Where can you find it?	You can find <C1> at <C2>

### 3.2 Generating Dialogue Turns

Carla uses the assertions in the semantic ontology to form the text that comprises its dialogue turn. Specifically, the conversation begins with a statement that expresses a commonsense thought, e.g., “*A fruit is good for you.*” In this instance, the topic used to start the discourse, “*fruit*”, came from the reading material which is about a young boy whose painting contains images of fruits.

An open-ended question that is related to the first statement follows, e.g., “*What else is good for you?*”, as well as a list of candidate responses. The latter is derived by querying the ontology for assertions that use the same relation as the first statement, to retrieve concepts that are semantically related from the ontology. A detailed example is shown in Table 4.

**Table 4.** Queries to the ontology to derive the contents of Carla’s dialogue.

Text	Query	Candidate Assertions
Statement	?(fruit, ?)	hasProperty(fruit, good for you) → if selected by planner locatedAt(fruit, grocery store)
Options	hasProperty(?, good for you)	hasProperty(exercise, good for you) hasProperty(sleep, good for you) hasProperty(warm bath, good for you)

Continuing from the given example, the user is presented with the following options as his/her response to the question “*What else is good for you?*”.

“*Exercise is good for you.*”

“*Sleep is good for you.*”

“*Warm bath is good for you.*”

In case more than three candidate assertions were retrieved from the ontology, three will be randomly chosen.

If the user selects “*Exercise is good for you.*”, the ontology is again queried to find candidate assertions using “*exercise*” as the seed concept. This cycle continues until

the virtual agent has determined that the learner's affect has already shifted to one that is more positive, or if the agent runs out of things to say.

Notice that in the given example, the options made by the user dictate the topic of discourse. Furthermore, the options do not necessarily relate directly to the reading material. This is intentional and is part of the storytelling process of providing opportunities for children to transfer the language they learn from stories to other personalized contexts. It is also a strategy to disrupt the user's negative affect to one that is more positive by presenting materials that a particular user may find interesting and appealing.

While Carla tries to appeal to topics that the user may find interesting, she must also ensure that the dialogue do not stray too far from the initial topic of discourse. Thus, the dialogue planner utilized by Carla performs recursive searches for semantically related concepts up to a level of three.

Another factor that Carla takes into consideration when choosing a topic of discourse is the concept's affective information. This is sourced from the polarity values stored in SenticNet [17]. Each commonsense concept in Carla's ontology is associated with a polarity value that describes how positive or negative is the affect being expressed by the concept. Carla should try to choose concepts with positive polarity values as a means of possibly shifting the learner's affect from negative to a more positive one.

Because detecting the learner's affect is outside the scope of the current research, the user has to explicitly inform Carla regarding his/her affect state anytime during the reading comprehension exercise or dialogue exchange by selecting either the "happy" or the "sad" icons. The dialogue then ends with an encouraging note to the learner to resume the learning activity, e.g., "*You seem to be feeling a lot better now. Maybe you can try working on the activity again? I'm sure you can do it now!*".

#### 4 Test Results

Table 5 shows the evaluation results from the preliminary testing of Ellie that was conducted among five (5) children diagnosed with mild autism. Children with mild autism are most capable of interacting with computer systems without requiring much assistance. They are able to live independently and possess a normal intelligence level, though they still struggle with decision-making tasks, common autism behavioral concerns such as overwhelming passion and interest towards a single topic, and social interaction and decorum.

Each participant was asked to go through the same story template five times to validate if he/she exhibited any learnings through the change in his/her choices. During this reading, the child should also interact with at least two objects or non-playing characters. Every option that has been selected is recorded. A shadow teacher is present to interpret the facial expressions or reactions of the child while using the system.

From the results in Table 5, the children seemed to have a preference for the teacher role of Ellie when they all gave a positive response in item #2. The facilitator role, on the other hand, only received affirmation from four of the children because

the fifth child had difficulty understanding the instructions given by the peer, as seen in item #3. All children liked the stories that they read (item #4), although two of the participants struggled with comprehension issues. Specifically, the language used in the stories, English, is not the primary language of the children who participated in the test. These children had to be given an ample amount of time to read and be told the stories before they were able to grasp the underlying meaning.

**Table 5.** Evaluation results of Ellie.

Questions	Yes	Somewhat	No
1. Did you understand what Ellie said?	3	2	-
2. Was Ellie able to tell you what is right from wrong?	5	-	-
3. Did you understand the instructions?	4	-	1
4. Did you like the stories?	5	-	-
5. Were you able to understand the stories?	3	2	-
6. Did you learn something new from the stories?	5	-	-

The results, however, do not provide any validation as to the effect of the Ellie's roles to student learning. Furthermore, two of the participants had to seek assistance from the shadow teacher in order to understand their dialogue with Ellie. This necessitates the need for the system to be used under the supervision of a human teacher or guardian.

Carla, on the other hand, was evaluated by 13 students. Table 6 shows the results when Carla was evaluated based on its dialogue content only.

**Table 6.** Evaluation results for Carla's dialogue content.

Questions	Yes	Maybe	No
1. Did you understand the tutor?	9 (69.2%)	4 (30.8%)	-
2. Did the tutor provide choices that make sense?	6 (46.2%)	7 (53.8%)	-

From the results, 69.2% of the learners had no trouble understanding the words used by Carla, while the remaining 30.8% partially understood the tutor. This can be attributed to the manual process of populating the semantic ontology by extracting relevant assertions from ConceptNet, and the use of predefined templates to generate the dialogue content.

For the appropriateness of the list of candidate responses that Carla provides the user, 53.8% of the participants reported concerns that range from amusing choices to choices that do not make sense. Repetitive choices are also present, such as “*flower in the park*” and “*flower at a park*”. Carla also sometimes produce redundant statements, such as “*you need an ice cream so you can eat an ice cream*”.

Separate system testing showed that, given the current size of the knowledge base, Carla can engage the learner in a 20-turn dialogue (20 dialogue questions), though a number of the questions were already repeating.

## 5 Conclusion and Further Work

We presented the design of the semantic ontology of two virtual agents – Ellie who uses interactive storytelling to teach learners about social interactions; and Carla who engages its learners in a text-based dialogue during a learning activity. Both the ontologies adapted the semantic relations from ConceptNet while supplementing these with additional knowledge to ensure that the generated text contains concepts relevant to the story themes.

The combined use of commonsense ontology and template-based story generation provided a platform for Ellie to produce a variant of stories that adhere to the narrative structure of a Social Story™ while conforming to the language needs of the target audience. In the case of Carla, this approach allowed for the generation of lengthy dialogue exchange between the agent and the learner.

To address redundancy issues in Carla, further work should consider adding a concept attribute that will be used to determine the next assertion to be expressed in the dialogue. The attribute must be dynamic, such that its value changes as the dialogue progresses. This is to prevent the dialogue planner from selecting the same set of assertions for a given set of conditions. A possible attribute is the frequency in which an assertion is expressed relative to the number of dialogue turns that have already been exchanged between Carla and the learner.

Furthermore, Carla engages the learner in only one type of dialogue, specifically information-oriented dialogue turns that ask the learner about his/her knowledge or preferences for a given topic. Generating other types of dialogue, such as persuasion or negotiation to encourage the learner to go back to his/her learning activity, has not been considered in the current implementation. Further research on motivational factors in learning and how these can be used by the dialogue planner to generate other types of dialogue will be explored in the future.

While resources such as ConceptNet are readily available from the Web, the specific requirements of the learning environments presented in this paper necessitated the need to build separate semantic ontologies. This presents a problem since the task of populating the ontologies is mostly manual and time-consuming. Future works should explore related studies in the automatic or semi-automatic population of knowledge from corpus.

Ellie was validated among a very limited population of children with mild autism. Though the results showed the potential of Ellie as a virtual companion to help chil-

dren comprehend the events in their daily lives and to prepare for social interactions, the tests did not provide conclusive evidence on the long-term effectiveness of the system as a learning environment.

Carla was validated among a slightly higher number of children compared to Ellie. Still, the test results did not provide any evidence relating to specific learning tasks and behavior, such as how the learner's bias towards the reading material may be affecting his/her attitude on the tutor's attempt at intervention, and the rationale for a conversational agent if the available reading materials are familiar to the target audience (and thus, they may have had positive affect throughout the reading activity).

Currently, both Ellie and Carla are visually portrayed as 2D graphical faces with fixed positive and negative facial expressions that are displayed at appropriate points based on the story plot. Future work can explore the use of animated virtual peers that can vary their expressions and possibly show an increasing capacity for affect that is visually authentic and appropriate to the ongoing discussions and interactions with the human user.

## References

1. Cassell, J., Tartaro, A., Rankin Y., Oza, V., Tse, C.: Virtual Peers for Literacy Learning. *Educational Technology, Special Issue on Pedagogical Agents* 47, pp. 39-43 (2005)
2. Dautenhahn K., Davis, M., Ho, W.: Supporting Narrative Understanding of Children with Autism: A Story Interface with Autonomous Autobiographic Agents. In: *Proceedings of the IEEE International Conference on Rehabilitation Robotics*, pp. 905-911, Kyoto International Conference Center, Japan. IEEE (2009)
3. Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., Schwartz, D.: Teachable Agents: Combining Insights from Learning Theory and Computer Science. In S.P. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education*, pp. 21-28. Amsterdam: IOS Press (1999)
4. D'Mello, S.K., Lehman, B., Graesser, A.C.: A Motivationally Supportive Affect-Sensitive AutoTutor. In R. Calvo and S. K. D'Mello (Eds.), *New Perspectives on Affect and Learning Technologies*. vol. 3, pp.113-126. New York: Springer (2011)
5. Ryokai, K., Vaucelle, C., Cassell, J.: Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of Computer Assisted Learning*, vol. 19 no. 2, pp. 195-208. Wiley Online Library (2003)
6. Baker, T.: Collaborative Learning with Affective Artificial Study Companions in Virtual Learning Environment. PhD Dissertation, The University of Leeds (2003)
7. Xu, Y., Park, H., Baek, Y.: A New Approach Toward Digital Storytelling: An Activity Focused on Writing Self-efficacy in a Virtual Learning Environment. *Educational Technology & Society*, vol. 14 no. 4, pp. 181-191. (2011)
8. Liu, H., Singh, P.: MakeBelieve: Using Commonsense Knowledge to Generate Stories. In: *Proceedings of the 18<sup>th</sup> National Conference on AI*, pp. 957-958, Edmonton, Alberta: National Conference on Artificial Intelligence (2002)
9. Kim, K., Picard, R., Lieberman, H.: Common Sense Assistant for Writing Stories that Teach Social Skills. In: *Proceedings of CHI EA '08 Extended Abstracts on Human Factors in Computing Systems*, pp. 2805-2810. New York: ACM (2008)
10. Cua, J., Ong, E., Pease, A.: Using SUMO to Represent Storytelling Knowledge. *Philippine Computing Journal*, vol. 5 no. 2, pp. 37-43. Computing Society of the Philippines (2010)

11. Ong, E.: A Commonsense Knowledge Base for Generating Children's Stories. In: Proceedings of the 2010 AAAI Fall Symposium Series on Common Sense Knowledge, pp. 82-87, Virginia, USA. AAAI (2010)
12. Liu, H., Singh, P.: Commonsense Reasoning in and over Natural Language. In: 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Berlin: Springer-Verlag (2004)
13. Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., Zhu, W.L.: Open Mind Common Sense: Knowledge Acquisition from the General Public. In: Proceeding of the 2002 Confederated International Conferences DOA, CoopIS and ODBASE, pp. 1223-1237 (2002)
14. Gray, C.: The New Social Story Book: 10th Anniversary Edition. Arlington, Texas: Future Horizons (2010)
15. Groden, J., LeVasseur, P., Diller, A.: Picture This!, Autism Spectrum Quarterly, pp. 18-21 (2007)
16. Speer, R., Havasi, C.: Representing General Relational Knowledge in ConceptNet 5. In: Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 3679-3686. European Language Resource Association (2012)
17. Cambria, E., Speer, R., Havasi, C., Hussain, A.: SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In: Proceedings of the 2010 AAAI Fall Symposium Series on Common Sense Knowledge, pp. 14-18. Association for the Advancement of Artificial Intelligence (2010)

## Automatic Question Generation on SQL Language Using Template-Based Method

Jittima Janphat<sup>1</sup> and Orawan Chaowalit<sup>2</sup>

<sup>1</sup> Department of Computing, Faculty of Science,  
Silpakorn University, Nakhon Pathom, Thailand  
[jittima.mink@gmail.com](mailto:jittima.mink@gmail.com)

<sup>2</sup> Department of Computing, Faculty of Science,  
Silpakorn University, Nakhon Pathom, Thailand  
[chaowalit\\_o@su.ac.th](mailto:chaowalit_o@su.ac.th)

**Abstract.** The objective of this research is to generate database questions automatically. Creating questions for learners to write SQL statement either to practice or to assess their knowledge level is time consuming for instructors. The system can reduce this workload by automatically generates questions for instructors. Template formats were created from sample exercises from database textbooks, and then the templates were filled with data from metadata about databases and data in database, which is to be put in the database management system by the instructor. After that, the system generates SQL question corresponding to the database to facilitate the instructor. Four experts from Silpakorn university who are familiar with database subject evaluated the reliability and the level of learning of questions that were generated by the system.

**Keywords:** Question generation · Metadata · Question template · SQL

### 1 Introduction

This paper presents the study of an automatic database question generation system. Most educational institutions usually teach variety of subjects in their faculty. Each subject oftentimes contains a lot of content; as a result, creating examination questions to evaluate students' level of knowledge is a lot of work and very time consuming for the instructor [1]. In addition, doing more practice questions can increase students' problem-solving skill as well as improving their study performance. This research highlights the more efficient method of teaching and learning that can benefit both learner and instructor [2]. Humans are usually not very skilled in generating questions especially if it requires more in-depth knowledge; therefore, an automatic question generation system becomes essential to generate good questions. An existing question generation system in [3] generates questions based on extracting key phrases and there are many techniques of creating automated questions. However, for a lot of questions and to cover all of the content in the subject, it may take a long time to create.

Nowadays, computers are widely used in various applications, whether it is the task of calculation, data storage, natural language processing, machine translation, information retrieval, or text summarization and so on. Consequently, we use computer system to generate questions from the subject of ‘Database System’ to evaluate students’ SQL statement knowledge. Some previous works proposed an approach to automatically generate SQL queries. SQL was specified by the instructor’s order. The user first selects a database against which to generate queries, specifies the number of related columns, column names and table name, specifies desired concept query, then the system automatically generates SQL queries used to verify the accuracy of the SQL statement; however, there were no questions to keep learners challenged and it cannot be used to test their problem solving skill [4]. The approach did not focus on Natural Language questions and it did not ask students to generate SQL statement. In this paper, we created a system that generates questions using Natural Language that ask learners to write SQL queries statement. This system is a useful tool to evaluate students’ performance and it can also be used to practice writing SQL statement. Our system also supports online courses that randomly generates SQL question for practicing.

This research paper is organized as follows; in section 2, we will introduce current research on automatic question generation, section 3, the proposed method will be presented, section 4 will be about methods of evaluation and finally, conclusion and future work will be discussed in section 5.

## 2 Theory and Related Work

### 2.1 Related Work

In automatic question generation context, questions were usually generated into different types; such as multiple choice, factoid, etc. From the best of our knowledge, these question generation methodologies can be classified into three approaches: template-based, syntax-based and semantic-based models. Information used to generate questions usually requires different methods and different data. Maha Al-Yahya [5] presented a method for multiple choice question generation using ontology. This paper queries data from two ontology domains for the evaluation: the HistOnto and SemQ. Another Research used template methodology [6] with SPARQL language to retrieve word from HisOnto and SemQ to fill in the question template. The questions were evaluated by comparing them with pre-defined rules in order to assess learning outcome according to Bloom’s Taxonomy learning theory. Nguyen-Thinh Le and Niels PINKWART [7, 8] presented a method for question generation using a relationship in WordNet. WordNet was used to create questions from a scientific text. They used the relationship of words and sentences from WordNet to fill in the question templates. Performance evaluation method was to predict whether the questions were generated by the system or by the specialist. In (Husam Ali Yllias et al.), they generated question generation from sentences [9]. Output questions were compared with the set rules. If the question does not match the rules, they updated the question and compare again. Another evaluation was using dataset: questions and answers from

TREC-2007 were used to evaluate the system. The evaluation measurements were: Precision and Recall. In (Hafedh Hussein et al.), this research generated questions of the English language content using template-based [10]. This method extracted sentences from English content by using sentence structure rules to create a question. Training systems process was to use the sentence structure form of English content and compared them with the rules in the template. If the content and template matched, this content would become the question, but if they did not resemble the existing rules, the system would create a new rule. After being trained, they have new rules of question syntax. The system started the test, loaded document data and then matched the document with the rules. After that, if the document is comparable with the rules, the system will generate the question. The criteria used to evaluate is the number of correct questions generated. Husam N. Yasin presented a method for automatic diagrammatic multiple choice question generation (DMCQ) from the knowledge [11] based on the Islamic finance contract model (IFKB), a term used to fill in the template is all about the goals. The goal is the concept of IFKB section. The system was evaluated by determining the difficulty of the questions, then built distractors and compared them with the rules. (D.S. Wang et al.) presented a method for domain-specific question answering system based on ontology and question templates [12]. Data were used to generate questions about the telephone counseling service. These questions were matched with a predefined template and showed a consistent response to the template. The template described the definition of BNF performance evaluation methods to ensure the effectiveness of the system in response to questions automatically. (Quan Do, et al.) presented a method for automatic generation of SQL queries [4]. They proposed an approach to automatically generate SQL queries using ontology from database metadata. This paper proposed question generation approach based on a template. Query metadata information was stored in a database schema comprises information; such as names of databases, names of tables in the database, attributes in a given table, database constraints, and data that the students wanted to learn to manipulate, by using SQL to query the data to fill in the questions template, the answer to the questions were in the form of SQL language for database systems.

## 2.2 SQL Language

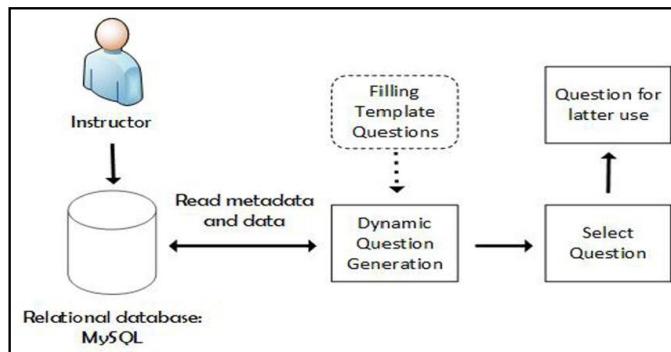
SQL stands for ‘Structured Query Language’, which is a computer language for storing, manipulating and retrieving data stored in relational database [4]. SQL is a standard language that can be used with any database management system. Additionally, SQL is a high level programming language. It is a structured program of language, which is easy to understand, uncomplicated, and has an efficient performance. SQL can work with complex queries using only a few commands. SQL statements can be separated into five different categories: Data Definition Language (DDL), Data Manipulation Language (DML), Data Control Language (DCL), Transaction Control Statement (TCS), and Session Control Statements (SCS). This research focuses on DML statement. Table 1. describes DML, which is a statement permitting users to manipulate data in the database.

**Table 1.** Data Manipulation Language

SQL statement	Meaning
SELECT	Extracts data from a database. Syntax: SELECT <column_name> FROM <table_name> WHERE <some_column = some_value>
INSERT	Inserts new data into a database. Syntax: INSERT INTO <table_name> VALUES (<value1, value2, value3,...>)
UPDATE	Updates data in a database. Syntax: UPDATE <table_name> SET <column_name> = <value> WHERE <some_column = some_value >
DELETE	Deletes data from a database. Syntax: DELETE <table_name> WHERE <some_column = some_value>

### 3 Methodology

In this section, we describe how automatic system generates questions to assess students' knowledge and to practice students' skill. The researcher gathers questions from exercises in the database fundamental textbooks to create the template formats. Then, uses data from metadata about databases to fill in the blanks in the template. Operation details are divided into two sections as shown in Fig. 1.

**Fig. 1.** System Architecture

#### 3.1 Data in Research

This research focuses on automatic question generation in the fundamental database. The questions emphasize on how to use or apply SQL language to manipulate data in

the database. We collected questions from C.J. Date's database textbook [13, 14, 15, 16]. The answers to the questions are in the form of SQL language for database systems. This research focuses on Data Manipulation Language (DML) [4] to command data management, for example INSERT, UPDATE, DELETE, SELECT, WHERE, LIKE and join the 2 tables that are related in the form of Primary key and Foreign key.

### 3.2 Create Templates

In this part, we extracted the knowledge from section 3.1 to create the templates. The templates use nature of the BNF (Backus-Naur Form) [12], [17] which is a standard model to describe the rules of the language to be used in determining a sentence that provides more convenient operation. BNF syntax can be written in the form of a short, compact form using Extended BNF. The braces symbol { } represents zero or more repetitions, brackets [ ] represent an optional construct, vertical bar | represents choice, parentheses ( ) are used for grouping, and angle-brackets < > represent that one has the exact information. Thus, there are four templates for the questions: SELECT, UPDATE, INSERT, and DELETE.

#### Template for 'Select' Question Generation

##### S1: Example question [15]

- Write a select statement to show all student grades from the Student Details table for students who are registered in Course id = "UV 10."
- Write a select statement to show all course names in uppercase, from the Course table.

The S1 template is defined as follows:

```
'Write a select statement to show all' <feild_name | description> ['in <uppercase | lowercase>'] 'from the' <table_name> 'table' ['for' <feild_name | description> 'in' <feild_name | description> '<=' | '>' | '<' ><data>]' ?'
```

##### S2: Example question [13]

- Get full details of all suppliers.
- Get full details of all suppliers in London.
- Get full details for parts supplied by a supplier in London.

The S2 template is defined as follows:

```
<command> 'full details' <conjunction> <table_name> | ['in' <data>] | ['by a' <feild_name | description> 'in' <data>]
```

##### S3: Example question [13]

- Get all shipments where the quantity is in the range 300 to 750 inclusive.

The S3 template is defined as follows:

```
<command> 'all' <table_name> 'where the' <feild_name | description> 'is in the
range' <random_number> 'to' <random_number2>
```

**S4:** Example question [14]

- List the name town and home of all students.
  - List the names of all course department and the number of credits for each.
- The S4 template is defined as follows:

```
<command> 'the' <table_name | (feild_name | description) {1 - n-1}> 'of all'
<table_name (feild_name | description){1 - n-1}> ['for each']
```

**S5:** Example question [15]

- Write a select statement to print the names of all the instructors, from the Instructor table, starts with “T.”
  - Write a select statement to print the last name of all the students, from the Student table, its name start with the letter “R” through “Z.”
- The S5 template is defined as follows:

```
'Write a select statement to print the' <feild_name | description> 'from the'
<table_name> 'table' '<starts with' <start_data> | 'its name starts with the letter'
<start_data> 'through' <end_data>>.'
```

**S6:** Example question [16]

- List all teacher names along with all the values for teacher number and course number contained in the Section table.
  - List girls and boys in the same city.
- The S6 template is defined as follows:

```
<command> <table_name1> 'and' <table_name2> 'in the same' <description_same> | <description> 'along with all the values for' <feild_name1> 'and'
<feild_name2> 'contained in the' <table_name2> 'table.'
```

**Template for ‘Update’ Question Generation**

**U1:** Example question [13]

- Change the color of all read parts to orange in the order table.
  - Change the name of number 654 to Kenneth in the Supplier table.
- The U1 template is defined as follows:

```
<command> 'the' <description> <conjunction> <<feild_name> 'is' <data> 'to'
<change_data> | 'read parts to' <change_data>> 'in the' <table_name> 'table.'
```

**Template for ‘Insert’ Question Generation**

**I1:** Example question [13, 14]

- Insert a new row into the PARC table. (1, 93, creek, king, y, 120)

- Insert a new part (name, city, 40, 179) into table P.
- Add a new record for (patti, 335, PA) to the suppliers table.

The I1 template is defined as follows:

```
<command> 'a new' <conjunction> '<into the' <table_name> 'table.'
<insert_data> | <insert_data> 'into table' <table_name> | 'for' <insert_data> 'to
the' <table_name>> 'table.'
```

#### **Template for 'Delete' Question Generation**

**D1:** Example question [13, 14]

- Remove all records from the enrolls table.
- Delete all records in the enrolls table which contain 3 grades.

The D1 template is defined as follows:

```
<command> 'all records' <conjunction> 'the' <table_name> 'table.' | ['which
contain' (<data> <feild_name> | <description>) {1 - n-1} ]
```

**D2:** Example question [13]

- Delete John from the owner table.

The D2 template is defined as follows:

```
<command> <data> 'from the' <table_name> 'table.'
```

**D3:** Example question [13]

- Delete all blue part.

The D3 template is defined as follows:

```
<command> 'all' <table_name> 'part.'
```

From the aforementioned templates, <command> represents the order of Select questions (e.g. Get, List, Select), <command> of Delete questions represents Delete or Remove, <conjunction> represents conjunction of the sentence (e.g. of all, for, record, from), <table\_name> represents the table name from the database, <feild\_name> represents attribute from the database, <description> represents description of each attribute in the database, <change\_data> represents replacement of the data in the existing database, <insert\_data> represents insertion of new data into the database, and <data> represents data in the database.

### **3.3 Generate Questions**

The system retrieves data from the database system in the form of metadata and data, which consist of three linked simple tables. The data includes table name, attribute, description, and two constraints: Primary key and Foreign key constraint. The

algorithm starts with retrieving metadata to replace in the template, which is shown in Fig. 2. The replacement of the <table\_name> in S2 template is shown in Fig. 3., and the question generated by the system is shown in Fig. 4.

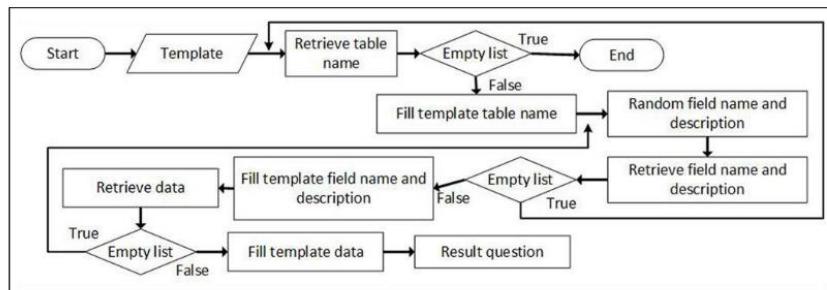


Fig. 2. Retrieve Metadata to Replace in Template

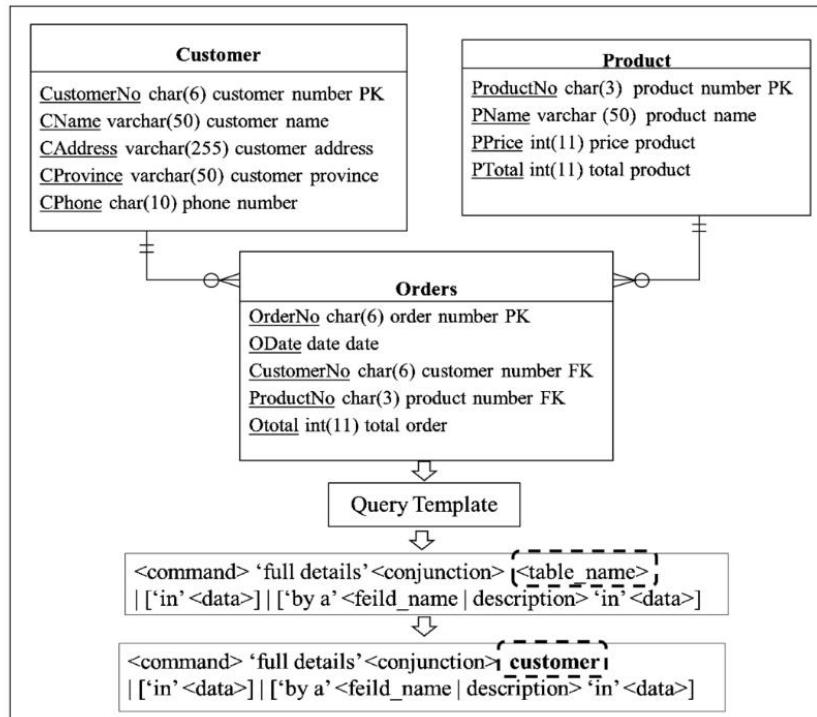


Fig. 3. Preview Template to Replace in the <table\_name>

- Write a select statement to print the customer name from the customer table starts with E.  
 - List full details of all customer by a CProvince in Chanthaburi.  
 - List all phone numbers along with all the values for PName, PTotals contained in the product table.  
 - Delete Chanthaburi from the customer table  
 - Update the customer name of CName, Jutarmas Yodrat to IQ2CFRGNK in the customer table.  
 Insert a new record for (JYRQTB, AOYZRSXFLP, GYRDUQWMBN, IRADUXEOKB, 04503187962) to the customer table.

**Fig. 4.** Question Generated by the System

## 4 Evaluation

We used two methods to evaluate the performance of the automatic database question generation that asks learners to write SQL statement. First, the level of learning of questions, and second, the efficiency and suitability of questions based on 5 evaluation criteria, Correct, Relevance, Correctness, Ambiguity and Variety.[18] Then, the evaluator evaluates the level of learning of the questions. The evaluators were 4 experts from Silpakorn University who are familiar with database subject. We mixed four question types of DML SQL statement together. There were INSERT, UPDATE, DELETE, SELECT and questions with specific conditions such as where a Primary key attributes or where the LIKE operators or mathematic operators. There are conditions and un-conditions questions. We selected 246 questions from the system and grouped them into three groups. There were 12 questions in each group. The goal is for the system to be able to create questions that are correct and suitable for all learners. The evaluator then assessed the overall performance of the system accordingly.

### 4.1 Evaluate Level of Learning of Questions

The two level of learning of questions, which were ‘understanding’ and ‘applying skill’ will be discussed in more detail below.

#### Understanding

- The questions of SELECT, UPDATE, DELETE the specified conditions, such as the database name, table names, attributes, and other details in the database, which are clearly identified by name in the database

- INSERT type questions with a single data input and a single table

#### Applying Skill

- The question of SELECT, UPDATE, DELETE without specifying the conditions, such as table names, attributes, and details in the database

- The need to analyze information, find a range of questions, find the beginning point, such as LIKE operator and mathematic operator

- Questions interrelated from two tables in the form of Primary key and Foreign key

For the level of learning of questions, the evaluator ranked the level of learning manually. The manually ranked scores are then compared with the results generated from the system. If the level matched, it is counted as 1 point, if not, it will be counted as 0 point. The total points were then divided by 12 (the number of questions for each set). The result from both set 2 and 3 questions in Table 2. comes to 83.33%. In summary, the system was able to distinguish between questions that are in the understanding category and questions in the applying skill category. However, the level of understanding and the applying skill will depend on the individual learner.

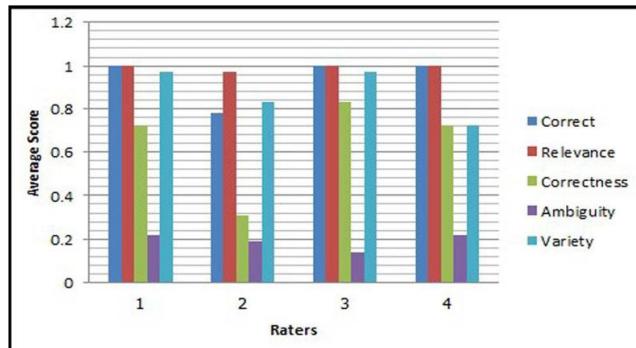
**Table 2.** Domains of Learning.

Set of questions	Domains of Learning.
Set 1	58.33 %
Set 2	83.33 %
Set 3	83.33 %

#### 4.2 Measuring Efficiency and Suitability of Questions

Professionals evaluated the efficiency and suitability of system-generated questions using five criteria: Correct, if the question meets its true objective, 1 point will be given, if not, 0 point will be given. Relevance, if the question relates to the lesson in the database, 1 point will be given, if not, 0 point will be given, Correctness, if the question is grammatically correct, 1 point will be given, if not, 0 point will be given, Ambiguity, if the question is ambiguous, 1 point will be given, if not, 0 point will be given and, Variety, if each set of questions (12 questions in each set) has variety, 1 point will be given, if not, 0 point will be given [19, 20].

The total points from the first four criteria, Correct, Relevance, Correctness, and Ambiguity were then divided by 12 (the number of questions in each set). Then combined the result with the points from the Variety criteria and divided by 3 (the number of sets of questions). Fig. 5. shows all of the five levels of learning score at between 0.31 and 1, which is satisfactory, while only ambiguity score is at between 0.14 and 0.22 the which shows ,tendency of slightly ambiguous questions [18].



**Fig. 5.** Five Levels of Learning

## 5 Conclusions and Future Work

This study describes an automatic question generation using templates. The goal of this work is to create an automatic generation of SQL question system. Instructors create any database, tables, and constraints and then insert the data. This system reads the data and generates questions for diverse learners. It is a useful tool for knowledge evaluation. The main advantage of the system is that the instructor can change the database at any time and saves time to create new questions. The limitation of the system is that it can only generate simple questions about SQL query, such as INSERT, UPDATE, DELETE, and SELECT and only 2 tables can be join. Some data fields generated may contain no meaning such as in Fig. 4. However, this problem can be ignored because the goal of question generation system is to evaluate the SQL queries writing knowledge. For future works, the system should be able to create some parts of the answer and increase the assessment criteria for different levels of students using Bloom's Taxonomy [15].

## References

1. Sarah Gibson, Lizzie Oliver, Mary Dennison: Workload Challenge Analysis of teacher consultation responses: Sixth form colleges. In: Department for Education (2015)
2. John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham: Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. In: Psychological Science in the Public Interest January 2013 14: 4-58 (2013)
3. Ming Liu, Rafael A. Calvo, Anindito Aditomo, Luiz Augusto Pizzato: Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support. In: IEEE Transactions on Learning Technologies, Vol. 5, NO. 3, pp.251-263 (2012)

4. Quan Do, Rajeev Agrawal, Luiz Dhana Rao: Automatic Generation of SQL Queries. In: American Society for Engineering Education (2014)
5. Maha Al-Yahya: Ontology-Based Multiple Choice Question Generation. Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 274949, 9 pages (2014)
6. Shiyan Ou, Constantin Orasan, Dalila Mekhaldi, Laura Hasler: Automatic Question Pattern Generation for Ontology-based Question Answering. In: Proceedings of the Twenty-First International FLAIRS Conference (2008)
7. Nguyen-Thin Le, Niels Pink want: Question Generation Using WordNet. In: International Conference on Computers in Education.Japan: Asia-Pacific Society for Computers in Education (2016)
8. r.mitkov, l.a.ha, n.karamanis: A computer-aided environment for generating multiple-choice test items. In: Printed in the United Kingdom Cambridge University Press (2005)
9. Husam Ali Yllias, Chali Sadid, A. Hasan: Automatic Question Generation from Sentences. In: TALN 2010, Montréal, 19–23 juillet (2010)
10. Hafedh Hussein, Mohammed Elmogy, Shawkat Guirguis: Automatic English Question Generation System Based on Template Driven Scheme. IJCSI International Journal of Computer Science Issues, 1694-0784 (2014)
11. Husam N. Yasin: Automatic Diagrammatic Multiple Choice Question Generation from Knowledge Bases. In: The Seventh International Conference on Information, Process, and Knowledge Management (2015)
12. D.S. Wang: A Domain-Specific Question Answering System Based on Ontology and Question Templates. In: International Conference on Software Engineering (2010)
13. C.J. Date: An Introduction to Database Systems. Pearson Education, Inc. (2004)
14. C.J. Date: Database A Primer. Addison-Wesley Publishing Company, Inc. (1983)
15. Susbi Sharma: A TUTORIAL APPROACH FOR TEACHING DATABASE CONCEPTS. Partial Fulfillment of the Requirements for the Degree of MASTER OF SCIENCE (2012)
16. Kifer, Bernstein, Lewis: Database Systems. Pearson International Education, Inc. (2006)
17. ROBERT W. SEBESTA: CONCEPTS OF PROGRAMMING LANGUAGES. Pearson Education, Inc. (2012)
18. Nguyen-Thinh Le, Tomoko Kojiri, Niels Pinkwart: Automatic Question Generation for Educational Applications – The State of Art. In: Springer-Verlag Berlin Heidelberg (2011)
19. Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, Cristian Moldovan: Question Generation Shared Task and Evaluation Challenge – Status Report. In: Proceedings of the 13th European Workshop on Natural Language Generation (ENLG) (2011)

20. Andrew M. Olney, Arthur C. Graesser, Natalie K. Person: Question Generation from Concept Maps. In: *Dialogue and Discourse* (2012)

## TSCS Monitor: Generation of Time Series Cross Section Tables from Moodle Logs for Tracking In-Class Page Views Using Excel Macros

Konomu DOBASHI†

† Faculty of Modern Chinese Studies, Aichi University  
4-60-6 Hiraike-cho Nakamura-ku Nagoya-shi Aichi-ken  
453-8777 Japan  
dobashi@vega.aichi-u.ac.jp

**Abstract.** This paper presents a method developed for viewing student access of course materials during class time and visually capturing student engagement with the materials. It focuses particularly on the development of TSCS Monitor, a set of Excel macros that automatically generates Time Series Cross Section (TSCS) tables from Moodle logs in order to monitor in-class student access of course materials. The numerical data provided by the tables can be used to identify learners who access the materials without properly following instructions or those who delay accessing the materials. It also provides data and suggestions that can be used as reference for reinforcing classroom instruction and keeping track of student engagement.

**Keywords.** time series, cross section, page views, student engagement, educational data mining

### 1 Introduction

Recently, the Course Management Systems (CMS) or Learning Management Systems (LMS) such as Moodle, currently being used in many universities are web-based and can be easily accessed from a variety of locations using a variety of devices. With such systems, the learning histories of learners are available on a large scale, offering a source of useful information with the potential to improve the quality of teaching and learning. In the education field, the logs of learning histories are large and diverse, and many kinds of data mining have been proposed [1].

A teacher is responsible for ensuring proper delivery of lessons in the classroom while simultaneously understanding the individual reactions and progress of students. Effectively satisfying these dual roles is essential to improving the quality of education. The problem is that, in a class composed of dozens of students, accurately

measuring individual reactions and progress is difficult even for experienced teachers. Data mining for the next generation of CMS should have the capacity to analyze learning histories in real time and report the status of the learner in a timely fashion.

This paper presents a method developed for viewing student access of course materials during class and visually representing student engagement with the materials. It describes the development of an Excel macro, which we call TSCS Monitor, that automatically generates Time Series Cross Section (TSCS) tables from the Moodle page view logs. The system monitors and records in-class page views and the time of viewing for all students attending the class. It also provides data and prompts suggestions that can be used to reinforce classroom instruction.

## 2 Related Research

TSCS Monitor, the approach developed in this study, is grounded in Excel pivot table functions. This makes it easy for teachers to obtain a summary of the frequency of student views of course materials. Dierenfeld and Merceron [2] and Dobashi [3] show various kinds of learning analytics methods using Excel pivot tables. Konstantinidis [4] has also developed Excel macros to process the Moodle logs in order to analyze page views and overall usage.

Moodog [5] reports that Zhang and Almeroth have developed an approach that incorporates an analysis function for logs in Moodle. Their system is able to analyze the course materials browsing rate, page views and viewing time of students. The analytical results are displayed on Moodle screens showing the interaction of the students and Moodle using graphs and the tables.

Mazza and Dimitrova have developed a system called CourseVis [6] that tracks student behavior in an online class. Such behavior can be visualized graphically, along with the status of student access to content pages following the course schedule. GISMO [7, 8] uses Moodle access histories to produce a graph of student access of course and teaching materials in order to better understand student behavior. It has been integrated into the Moodle open source learning environment and is currently being evaluated in several studies involving actual many users. Google Analytics provides a website analysis service that enables data analyses using various perspectives [9]. Whereas Google Analytics can be used only by the Moodle administrator, the method proposed in this paper can be employed by any Moodle user. TSCS Monitor is equally accessible to any Moodle course administrator.

## 3 Tracking and Processing Method

This paper targets an actual lesson for which the Moodle learning management system was used. Course materials for the lesson were uploaded on Moodle, and students could browse in-class or at home. Moodle was used to accumulate learning histories, including elements such as page views of course materials and quiz results. We developed the Excel macros used to generate the TSCS analysis featured in this

paper. The system provides TSCS analysis of student page views at regular intervals throughout the lesson from multiple perspectives (Fig. 1).

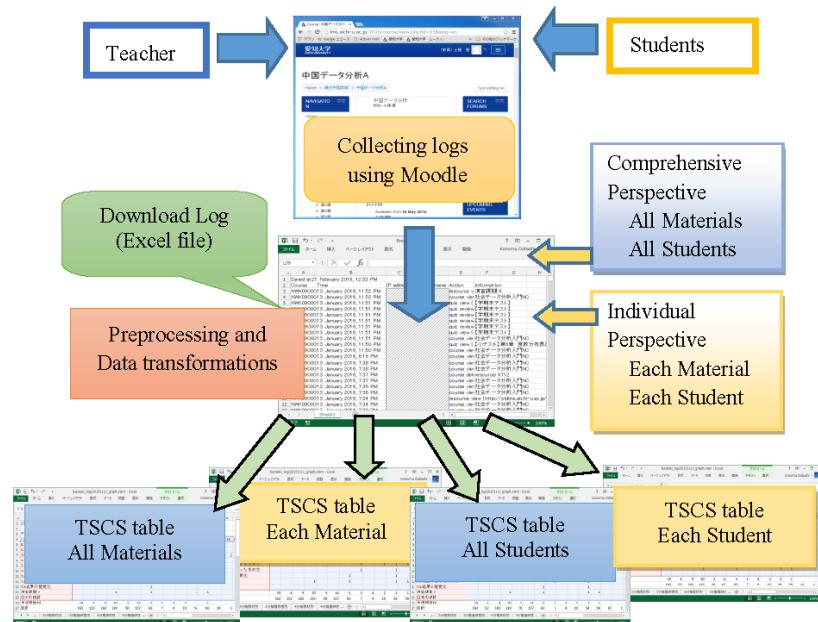


Fig. 1. Overview of the system configuration and multiple perspectives

A lesson is delivered by the teacher (who instructs students to view the prepared course items) and the teacher logs in to Moodle from his/her desk. Students log in at the beginning of the class. This method enables simultaneous recording of the teacher's page views. Because the course materials comprise 15 lessons to be delivered in half a year, sections that are not used during a given class day are also set as be able to browse items. The "used" course items are those accessed by the teacher in the day's lesson. The "unused" course items are those accessed by students but not used by the teacher in the day's lesson.

The TSCS tables are automatically generated. Excel has several features designed to process qualitative data. Especially the pivot table feature enables users to count qualitative data such as strings, create a cross section table, and quantify input data. These functions have been applied in this study. The tabulations generated here are presented in what we call a "Time Series Cross Section Table" in which an aggregated pivot table was created to incorporate time series data into the analysis.

At the beginning of the lesson, the teacher use own laptop can access Moodle via the course administrator and open a log page. Logs for download can be prepared for downloading, and the Excel format can be specified as the desired form at any time with one click. Clicking TSCS Monitor enables it to run immediately after the teacher downloads a Moodle log.

While delivering a lesson, the teacher can view student status, download Moodle logs to a specified folder, and run TSCS Monitor. The downloading of logs and macro processing take only tens of seconds. These features guarantee that sufficient time and focus is devoted to the lesson. After the lesson is completed, the teacher can run TSCS Monitor (if necessary) without having to worry about processing time during class.

In this paper, TSCS Monitor was applied in the China Data Analysis class offered at the case university to demonstrate how a TSCS table is generated. The contents of the course and the corresponding course materials contain a commentary on introductory statistics created with Excel and related exercises. On December 8, 2015, the teacher discussed the lesson on IF function and Conditional judgment for 90 minutes. The lesson was initiated at 13:00 and ended at 14:30.

#### 4 Comprehensive Perspectives

From Table 1, one can see the changes in total page views over time for each of the course items used in-class. From such a table, it is possible to identify the more heavily engaged portions and the less heavily engaged portions of the lesson. Table 1 contains TSCS data generated at 1-minute intervals. It was downloaded at 13:21 from Moodle logs and aggregated.

**Table 1.** Example of a TSCS table for used course items by minute of class time (China Data Analysis, Saved 8 December 2015, at 1:21 PM)

12/8/2015	China data analysis											
	Column label											
Rowlabel	13:00	13:01	13:02	13:03	13:04	13:05	13:06	13:07	13:08	13:09	13:10	
0.1 China data analysis	13	7	1	8	4	9	11	9	2	5	44	
0.2 Quiz Chapter 9	106	86	66	89	62	80	46	26	12	14	118	
10.0 IF function											3	
10.1 Conditional judgment											1	
Unused material	12	8	3	2	14	10	14	6		1	7	
Total Page Views	131	101	70	99	80	99	71	45	15	20	169	
												Page Views
	13:11	13:12	13:13	13:14	13:15	13:16	13:17	13:18	13:19	13:20		
	31	7	2	1			1		1		156	
	39										1	745
	4	13	25	6	2						2	56
	1	3	3	5	23	7	1		8	15	7	68
	5	7	5		4	2	1	1	2	7		111
	80	30	35	12	29	9	3	1	14	23		1136

Table 2 is a TSCS table of minute-to-minute page views for the entire list of students who attended the lesson. In Table 2, we can see the student viewing times and page views. Used when conducting a lesson, such a table can identify students whose views of a course items are delay or indicate whether a particular student is viewing course items according to the instructions of the teacher. It is also possible to use this type of table to identify students who are not engaged in the lesson. For example, one

can identify students who viewed the same course item multiple times in the same minute.

**Table 2.** Example TSCS table for the entire list of students in the class (China Data Analysis, Saved 8 December 2015, at 1:21 PM)

Row label	Column label															Page	
	13:00	13:01	13:02	13:03	13:04	13:05	13:06	13:07	13:08	13:09	13:10	13:11	13:12	13:13	13:14	13:15	Views
Student01	5	1	2	3	3						1	2					22
Student02	4	2	1	1	1	1	4					1			1	M	20
Student03	5	1	1		2	3					1	2	1	1	1	o	17
Student04	3			1	1	1	1	1	2		1	1	1	1	1	r	13
Student05				5	3	3	7	2		1	8	2			1	e	34
Student06	4		2	5			1				3				1		17
Student07	5	2	1	1	1	3					1	1	2			d	17
Student08	5	1	1	1	4	4					1	1				a	23
Student09	3	2	1	7	1					2	7	2	1	1	t	27	
Student10	5	1	5								3	3		1	1	a	21
Student11	3	1	1	1		4	6	3			3			1		c	24
Student12	1	3		4	7						3	4	2		1	u	25
Student13				4	2	4	4				1	2				t	18
Student14				4	2	2	3				3						16
Student15										14			1	1			20
Student16		4	1		2	5										h	16
	... More students data cut here...															e	16
Student51	4	2	5					2		1	1					r	7
Teacher	1	3									1	1				e	1136
Page views	131	101	70	99	80	99	71	45	15	20	169	80	30	35	12	29	51
Students	38	41	40	38	35	31	21	16	6	5	47	31	17	22	10	22	

## 5 Individual Perspectives

Table 3 is a TSCS table of page views for a particular course item (here, “10.1 Conditional judgment” as shown in the seventh row in Table 1) in the lesson. It displays page views of the targeted item by individual students during each minute of class

**Table 3.** Example of an individual course item TSCS table (10.1 Conditional Judgment)

Row label	13:07	13:08	13:09	13:10	13:11	13:12	13:13	13:14	13:15	13:16	13:17	13:18	13:19	13:20	Page
															Views
Student01															1
Student02															1
Student03															1
Student04															1
Student05															2
Student06															1
Student07															1
Student08															2
Student09															2
Student10															3
	... Part of the data cut here ...														
Student46										1					1
Student47															1
Student48										1					1
Student49										1	1				2
Student50															2
Student51															1
Teacher										1					1
Page Views	1	1			1	3	3	5	23	7	1		8	15	68
Students	1	1	0	0	1	2	3	5	21	7	1	0	8	8	49

time. As shown on the right side of the table, there is no value (in effect, a value of 0) for total page views for two of the students in the class (the student in the first row and the student in the fifth row from the bottom of the student list). However, the table does not track cases in which a student may have viewed or downloaded the item previously.

## 6 Discussion and Conclusion

Manual methods to generate the kinds of tables described here would take tens of minutes to produce. By contrast, using TSCS Monitor requires only several seconds. Beyond enabling teachers to understand the changes that underlie student page views, TSCS tables can provide data on variations in student levels of concentration. The tables have undeniable possibility of looking at the downloaded materials. Furthermore, after the teacher provides directions for opening a course material, students spend about 1 to 2 minutes accessing the resource. The aforementioned issues should be considered before the teacher advance proceeds to the next lesson.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K00498.

## References

1. Romero, C. Ventura, S.: Data mining in education. *WIREs Data Mining Knowl Discov.* vol. 3, pp. 12–27. doi: 10.1002/widm.1075 (2013)
2. Dierenfeld, H. and Merceran, A.: Learning Analytics with Excel Pivot Tables. In Proceedings of the 1st Moodle Research Conference (MRC2012). Retalis, S. Dougiamas, M. Eds. pp. 115–121 (2012)
3. Dobashi, K.: Time series analysis of the in class page view history of digital teaching materials using a cross table. In *Procedia Computer Science*, vol. 60, pp. 1032–1040 (2015)
4. Konstantinidis, A. Grafton, C.: Using Excel Macros to Analyses Moodle Logs. In: 2nd Moodle Research Conference (MRC2013), pp. 33–39. 4th and 5th October, Sousse, Tunisia (2013)
5. Zhang, H., Almeroth, K.: Moodog: Tracking Student Activity in Online Course Management Systems. *Journal of Interactive Learning Research.* vol. 21, no. 3, pp. 407–429. Jul. (2010)
6. Mazza, R., Dimitrova, V.: CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human-Computer Studies*, vol. 65, no.2, pp. 125–139. Feb. (2007)
7. Mazza, R., Milani, C.: Gismo: a graphical interactive student monitoring tool for course management systems. *International Conference on Technology Enhanced Learning*. Milan, pp. 1–8. Jan. (2004)
8. Mazza, R., Botturi, L.: Monitoring an online course with the GISMO tool: A case study. *Journal of Interactive Learning Research*, vol. 18, no. 2, pp. 251–265 (2007)
9. GOOGLE ANALYTICS, 2016. <http://www.google.com/analytics/>

## Development of Salary Prediction System to Improve Student Motivation using Data Mining Technique

Pornthep Khongchai<sup>1</sup>, Pokpong Songmuang<sup>2</sup>

<sup>1,2</sup>Department of Computer Science, Faculty of Science and Technology,  
Thammasat University, Thailand  
pornthepl.khon@thammasat.net, pokpong@cs.tu.ac.th

**Abstract.** This paper aimed to determine an efficient data mining technique for salary prediction to motivate the eagerness to study. Five data mining techniques were compared: Decision trees, Naïve Bayes, K-Nearest neighbor, Support vector machines, and Neural networks. To evaluate the relative efficiencies of the techniques, 13,541 records of graduated student data were used in 10-fold cross validation. Results showed that K-Nearest neighbor provided the best efficiency. K-Nearest neighbor was also applied as a model for salary prediction. A questionnaire survey was used to evaluate the effectiveness of the system with 50 student samples. Results indicated that the system was effective in boosting students' motivation for studying and also gave them a positive future viewpoint. The student sample registered positive satisfaction in using the system, since it was easy to use and the predictive results were simple and comprehensible.

**Keywords:** Educational Data Mining, Classification technique, Decision trees, Naïve Bayes, K-Nearest neighbor, Support Vector Machines, Neural Networks

### 1 Introduction

Nowadays, most university students enter into academic life without proper goals so they lack motivation to study in a class. Therefore, most students get bored of studying, and this causes failure in understanding lesson and examination [1]. This problem makes students to miss a required competency and knowledge which causes them retired from university. Apparently, students start their university life because they expect to have a good career with high payment after their graduation. Since comfortable life with sufficient income is one of most common human dream, the idea of this research is to use a predicted salary from graduated student history to motivate the eagerness to study and work toward their future plan.

Many studies propose salaries prediction models based on statistical models [2][3][4]. Although the proposed models perform well on a prediction of salaries task, but the problems are 1) the prediction model predicts salaries for a group, not for individual student and 2) the results from the prediction models require a person who has background in statistic to comprehend.

Therefore, in this paper, we propose a salary prediction system to predict students' future salaries based on graduated student history. We compare data mining techniques and select the best technique to conduct the salary prediction model.

The rest of this paper is organized as follows. Literature Reviews are described in Section 2. The proposed system is explained in details in Section 3. Section 4 shows experiment results including accuracy result and effect result. Finally, we conclude the paper in Section 5.

## 2 Literature Review

The goal of this paper is to find efficient data mining techniques [5] for predicting salary in order to motivate student improving their study. Therefore, at first, we review several data mining techniques in education area.

Young-joo Lee et al. [2], John jerrim [3], Karla et al. [4]. They applied regression method to predict not only salary but also satisfaction in jobs based on graduated student's history data, but the problems are 1) the prediction model predicts salaries for a group, not for individual student and 2) the results from the prediction models require a person who has background in statistic to comprehend.

Maomi Ueno [6] , Farhana Sarker et al. [7] proposes decision tree model constructed from the learning history data in the database to predict a student's future final status and predict students' academic performance.

The previous researches indicate that data mining techniques are popular to be applied as prediction models in educational research area. Moreover, the prediction models using several models are easy to understand.

However, data mining techniques have not been applied for student salary prediction. Therefore, next we compare the performance of data mining techniques for salary prediction and we apply the best one for creating a salary prediction system to motivate the eagerness to study.

## 3 Comparison of Data Mining Techniques for Salary Prediction

### 3.1 Data Mining Techniques

There are many data mining techniques that can be applied on salary prediction. Here, we select five models as follows, K-nearest neighbors (K-NN), Naive Bayes (NB), Decision tree (J48), Multilayer Perceptron (MLP) and Support Vector Machines (SVM). We compare the efficiencies of the above data mining techniques for predicting salary and install the best one in the salary prediction system described in the next section.

### 3.2 Salary Prediction System

The goal of this system is motivation student. We apply the best data mining techniques for creating a salary prediction system to motivate the eagerness to study.

Salary Prediction System

Gender: Female

Faculty: Engineering

Program: Engineering Civil

Type of Work: Employ Company

Job Training: cooperative

Certificate: Pass Practice

GPA: < 2.8

Submit Cancel

Figure 1: Interface of Salary Prediction System

Result Salary is : >18,000

Show case the best of three:

Sex	Faculty	Program	Type of Works	Job Training	Certificate	GPA	Skill	salary
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.70	Skill of Computer	25,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.64	Skill of Language	23,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Not pass the practice test	2.55	Not a Skill	20,000

Back

Figure 2: Salary Prediction Result

The system is motivation student using 1) the salary prediction from profile of students, 2) the examples of three graduated students who have similar profile and the top-ranked salaries. Accordingly, users should understand the predicted results without statistical background.

Fig. 1 illustrates the interface of Salary Prediction System. The interface contains seven attributes mentioned before. These inputs are considered as the attributes that require users to fill in the salary prediction system. After the users input the attributes and click on submit button on bottom of the interface, the system compare the inputs of attributes with the rules and show predicted salary on the top of result interface as shown in Figure 2. For example in Fig. 2, the predicted salary is *more than 18,000 baht*. Moreover, the salary prediction model compares the inputs of attributes with the profile in the database and selects the three graduated students with similar attributes with the highest salaries.

Therefore, we can conclude that the proposed salary prediction system has the potential to solve the problems of the previous systems.

## 4 Experiment and Result

In this paper, we compare data mining techniques for salary prediction using graduated student history data collected for 10 years since 2006-2015 from Rajamangala University of Technology Thanyaburi, Thailand.

### 4.1 Data Preparation

We apply linear equating techniques to adjust the previous salaries based on current base salary where our hypothesis is that the distributions of salaries are similar for every year. The salary prediction model will be created and used in the salary prediction system.

Therefore, we prepare the data as three following steps:

1.) Data Selection: We use forward selection [8] analysis to select 7 attributes from 108 attributes. The remaining 7 attributes are as follows: gender, faculty, program, type of works, job training, certificate, and GPA.

2.) Data Cleaning: Outlier data and missing values are manually removed. After data cleaning, the remaining history data are 13,541 rows.

3.) Data Transformation: We apply user specific discretization [9] of the salaries in which the salaries are divided to four levels for classes. There are four levels of salary, *less than 13,500, 13,501 - 15,300, 15,301 - 18,000 and More than 18,000*.

### 4.2 Comparison of Data Mining Techniques

We compare data mining techniques for predicting salary using data mining tool called WEKA. We use 10-fold cross-validation technique to evaluate the efficiency of the salary prediction model created using data described in previous section. Table 1. indicates the evaluation results of the salary prediction model in terms of Recall, Precision, F-measure and Overall Accuracy [10][11].

TABLE 1. SUMMARY OF SALARY PREDICTION MODEL

Class	Recall (%)				
	K-NN	NB	J48	MLP	SVM
Less than 13,500	<b>84.90</b>	50.90	75.80	45.30	37.00
13,501 - 15,300	<b>83.40</b>	49.70	71.80	41.00	58.00
15,301 - 18,000	<b>83.70</b>	21.70	71.80	0.50	23.60
More than 18,000	<b>87.10</b>	53.40	76.90	68.70	56.70
Class	Precision (%)				
	K-NN	NB	J48	MLP	SVM
Less than 13,500	<b>84.90</b>	43.90	74.60	33.30	48.70
13,501 - 15,300	<b>84.10</b>	38.70	73.40	33.10	37.10
15,301 - 18,000	<b>83.60</b>	45.10	71.50	38.30	43.00
More than 18,000	<b>86.30</b>	49.30	76.50	47.70	51.00
Class	F-measure (%)				
	K-NN	NB	J48	MLP	SVM
Less than 13,500	<b>84.90</b>	47.10	75.20	38.40	42.10
13,501 - 15,300	<b>83.70</b>	43.50	72.60	36.70	45.30
15,301 - 18,000	<b>83.60</b>	29.30	71.70	1.00	30.50
More than 18,000	<b>86.70</b>	51.30	76.70	56.30	53.70
Overall Accuracy (%)	84.69	43.63	73.96	38.08	43.71

We considered predicting each data mining techniques results from, Table 1. Summary of salary prediction model compares models shows Recall, Precision, F-measure and the average overall accuracy prediction results obtained by we can clearly see that the highest accuracy is 84.69% and the lowest is 38.08%. The highest accuracy belongs to the K-nearest neighbors (K-NN) followed by Decision tree (J48), Support Vector Machines (SVM), Naïve Bayes (NB) and Multilayer Perceptron. K-NN has the highest prediction results after run model, Recall has the best results yielded from the system are on a class of *salary More than 18,000* scored 87.10%. Then, K-NN has the highest prediction results after run model, Precision has the system are on a class of *salary More than 18,000* scored 86.30% and K-NN has the highest prediction results after run model, F-measure has the system are on a class of *salary More than 18,000* scored 86.70%.

According to the result, we apply K-NN the best overall accuracy model for create salary prediction system. The salary prediction system is application online that can salary prediction from profile of students.

#### 4.3 Evaluation of student motivation

In this part, questionnaire was selected to evaluate student motivation. It is divided into three parts. The popular samplings are 50 students from Rajamangala University of Technology Thanyaburi, Thailand. Which, it is the same source of those in use in salary prediction system.

**Part 1:** general student information

This part gathers basic student information such as gender, age, faculty and program. The obtained data are summarized as follows. For gender, we got 26 male and 24 female samplings. A range of age is 19 to 22 year-olds. From a faculty perspective, students are from 7 different faculties. For academic year, we got 7, 15, 5, and 23 student samplings from year 1 to 4 respectively.

**Part 2:** question to examine student motivation before using this part contains questions to examine student motivation.

**Part 3:** question to examine student motivation after using this part contains questions to examine student motivation.

From the Q1's: "Which level of satisfaction do you have for your faculty/ program?" mean score results show that students are satisfy in high degree for their current faculty and program. Moreover, in comparison Q1 with Q2: "How much satisfaction do you have with your current faculty/program?", we found that the predicted salary can boost up students' satisfaction for their current faculty and program. From Q3's results: "After knowing the predicted salary, how much satisfaction do you have with your enrolled faculty/program?", the predicted salary by the system makes students to realize the implicit benefit of being in the current faculty and program and an expedition to accomplish their own future plan. The Q4's results : "How much does the system motivate you to study for your expected salary?" can be inferred that with the predicted salary, students have the much clearer goal and will likely be more motivated to study to achieve the expected salary.

We evaluate detail as shown in after using the salary prediction system. In this part, Q5: "Is the system easy to use?" and Q6: "Is the system's result reliable in your opinion?", we mainly asked for an evaluation of satisfaction in using the system. By the given score, we found that the users are much satisfactory with the designed user interface as it is easy to use. This result is understandable since the prediction is for further ahead of their current state and very hard to be proved.

## 5 Summary

In this paper, we compare data mining techniques for salary prediction using graduated student history. We set up an experiment by using 13,541 records of actual graduated student data in 10-fold cross validation. The total result shows K-NN performs the best accuracy 84.69%.

After that, we apply K-NN for creating a salary prediction system to motivate the eagerness to study. The result of the system is not only a predicted salary, but also the 3-highest salary of the graduated students which share common attributes to the users. The result shows that this system can boost students' motivation in studying and also show them a positive viewpoint of their future.

## 6 Acknowledgment

The authors express their thanks to Office of Academic Resource and Information Technology, Rajamangala University of Technology Thanyaburi, Thailand.

## 7 References

- [1] Lumsden, Linda S., "Student Motivation To Learn. ERIC Digest", Number 92., 1994.
- [2] Young-joo Lee & Meghna Sabharwal , " Education-Job Match, Salary, and Job Satisfaction Across the Public, Non-Profit, and For-Profit Sectors: Survey of recent college graduates", Public Management Review, 18:1, 40-64., 2014.
- [3] John Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", Education Economics, 23:2, 162-179. 2013.
- [4] K. R. Hamlen,W. A. Hamlen,"Faculty salary as predictor of student outgoing salaries from MBA programs", Journal of Education for Business, 91:1, 38-44. 2015
- [5] ROMERO C. AND VENTURA S. 2010. Educational Data mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics. 40(6), 601-618.
- [6] M. Ueno, "Animated Pedagogical Agent based on Decision Tree for e-Learning", Proc.IEEE conference (Computer Science), ICALT, 2005.
- [7] F. Sarker, T. Tiropanis, H. C. Davis, "Students' Performance Prediction by Using Institutional Internal and External Open Data Sources", CSEDU, 2013, page 639-646. SciTePress.
- [8] M. A. Hall, G. Holmes."Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE Transactions on Knowledge and Data Engineering, Vol.15, NO.3, 2003.
- [9] P. Berka, I. Bruha,"Discretization and grouping: preprocessing steps for Data Mining", Second European Symposium, PKDD '98, Nantes, France, 1998.
- [10] O. Villacampa,"Feature Selection and Classification Methods for Decision Making: A Comparative Analysis", College of Engineering and ComputingNova Southeastern University , 2015.
- [11] G. Forman,"An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research 3, 1289-1305, 2003.

## Contents Organization Support for Logical Presentation Flow

Tomoko KOJIRI<sup>a</sup> and Yuta WATANABE<sup>b</sup>

<sup>a</sup> Faculty of Engineering Science, Kansai University

<sup>b</sup> Graduate School of Science and Engineering, Kansai University

3-3-35, Yamate-cho, Suita, Osaka, 5648680 Japan

kojiri@kansai-u.ac.jp

**Abstract.** In appropriate presentations, each topic must be explained logically. To prepare appropriate presentation slides, all of the topics related to the research must be fully elucidated. The objective of this research is to help presenters derive enough topics that cogently explain their research theme. This paper proposes the logical model for the topics in the research presentation of the computer science field. Then, content organization support system that helps presenters organize topics based on the logical model is developed. Based on the experiment, our system was effective for creating a new contents and for organizing contents.

**Keywords:** content map, causal relation, presentation support, logical presentation

### 1 Introduction

We researchers often introduce our research results as presentations. In them, we often prepare presentation slides in which discussion topics are sequentially arranged. In such presentations, the order of topics is critical so that audiences correctly understand what the presenter is arguing. If the relations among topics are difficult to grasp, audiences might get confused. In this research, we define an appropriate presentation as one in which its topics are easy to understand, and an inappropriate presentation is difficult to understand.

Of course, such presentation tools as PowerPoint, Keynote, or Prezi provide functions that support the creation of presentation slides. However, as Kohlhase argued, these functions minimize the effort and time to create beautiful and stylish slides [1] and provide layouts or animation functions to create visually understandable slides. However, they do not support the preparation of contents for appropriate presentations. Several researches have encouraged presenters to improve their slide creation skills. One approach provides an environment in which comments about a presentation's good and bad aspects are simply gathered and exchanged during rehearsals [2, 3]. In this approach, whether a presenter can improve her presentation skills depends on who provides feedback, so presenters are not always able to improve their skills.

In appropriate presentations, each topic must be explained logically. That is, all topics must be connected by logical relations that are clearly represented in the slides. To prepare appropriate presentation slides, all of the topics related to the research must be fully elucidated. Their relations with other topics must be considered; logically-connected topics should be selected that adequately explain the presentation theme in the given presentation conditions.

This research focuses on two steps: deriving topics and considering relations between topics. The objective of this research is to help presenters derive enough topics that cogently explain their research theme. In research presentations, fundamental topics should be included with which the research is composed, such as goals, background, or methods. In addition, other topics are needed that explain the validity and reliability of each topic. Such additional topics may have relationships with fundamental topics. Our research develops a system that encourages presenters to derive both fundamental and additional topics and arrange them based on logical relations. For the purpose of encouraging users to induce new ideas, several idea processing systems provide related information with already derived ideas [4, 5]. Such systems do not consider necessary types of topics. Another approach helps presenters learn presentation slide composition based on the slides of experienced presenters [6, 7]. Types of topics and their relations may be different according to the research themes. Thus, this approach is not appropriate if the presentation themes are not identical to the experienced presenters' ones.

Our research does not directly support presentations. Instead, it encourages presenters to reflect on their own research and organize the topics that explain it. If topics are organized logically, creating appropriate presentations can be simplified by selecting required both fundamental and additional topics that have logical relations to the fundamental topics.

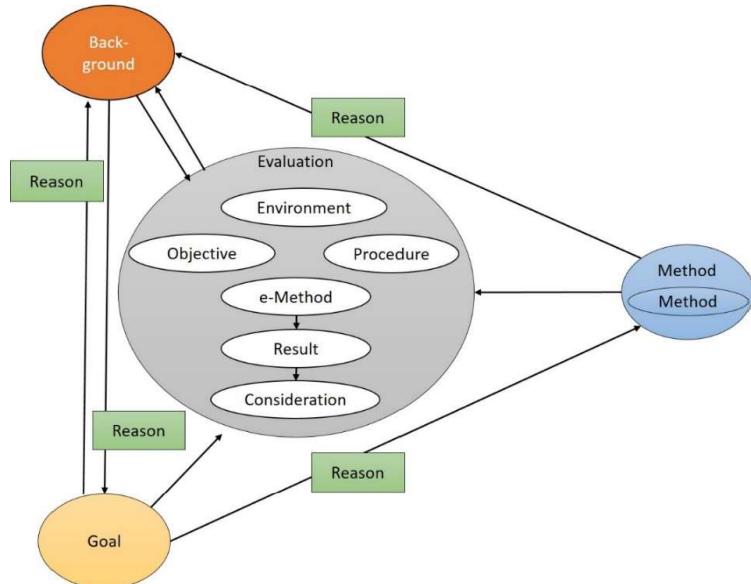
## 2 Logical Model for Research Presentations in Computer Science

A logical presentation explains its necessary topics with reasons. In research presentations, several explanations must be made, including the importance of the research's goal based on existing facts or opinions, the necessity of the proposed method by explaining how it achieves the goal, or its effectiveness by showing its evaluation results. To explain a presentation's topics, each one must be connected with reasons.

To support logically organized presentations for computer science researches, both fundamental and additional topics must be determined as well as their relations. Although Tanida et al. classified topics in the research presentations of computer science and proposed presentation semantics [4], their classification is too fine. Some of their classifications, such as logical approach and technological approach, are not necessary to create logical presentations. We eliminated some of the types and defined 11 types, which contain both fundamental and additional topics that can logically explain the fundamental topics.

First, we proposed a logical model of presentation contents in computer science (Fig. 1). In this model, links represent causal relations and nodes inside a node

indicate inclusive relations. Research presentations propose methods to accomplish goals. Therefore, method and goal are necessary components. Since method is derived to accomplish a goal, a causal relation is attached from goal to method. Since methods sometimes consist of several sub-methods, methods have inclusive relations. To describe a goal's validity, its background is explained that elucidates the target problem that must be solved or the requirements that must be satisfied in the research. A goal sometimes becomes the background of another goal, so they have causal relations in both directions. The validity of goal and background or the effectiveness of the proposed method is evaluated by an experiment called an evaluation. An evaluation consists of several factors: what to evaluate (objective), how to evaluate (e-Method), by whom and where the experiment was conducted (environment), and so on. Based on the experimental results, sometimes new problems are discovered, which become the background of other researches. In addition to these components, we add reason components for causal relation links. A reason is a special component that explains why the presenter believes that two nodes share a causal relation. This may be an important component since it reflects a presenter's belief.



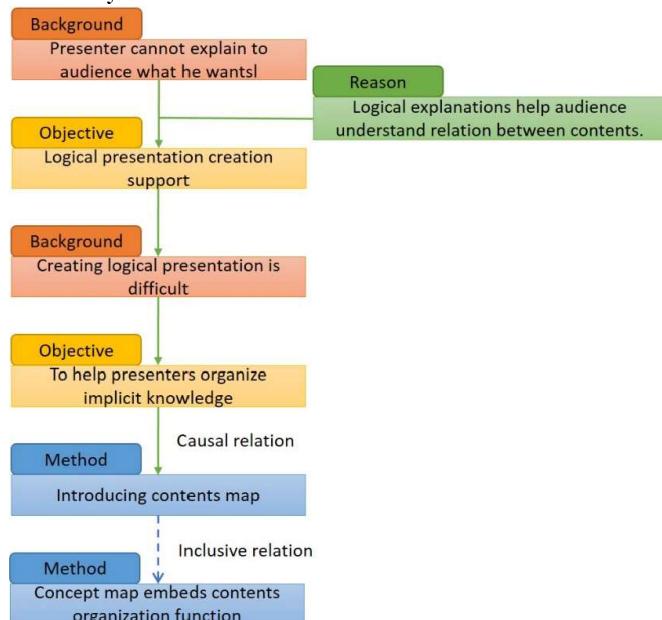
**Fig. 1 Logical model for research presentations for computer science field**

In our model, causal relations are not attached between topics. Background does not become a reason for a method, because no purpose was mentioned for deriving it. In addition, method is not a reason for establishing a goal, since a goal should solve the problems that are described in the background.

### 3 Content Map and Contents Organization Support Function

Since the topics that a presenter wants to discuss are implicit, the current topic structure, which a presenter has in his mind, is clarified and suggestions can be given for logical presentations if he externalizes his topics and their relations. By externalizing implicit knowledge, the presenter himself will notice the inappropriateness or incorrectness of his understanding of the topics.

In this research, we developed a system that 1) externalizes enough topics and 2) organizes them based on our logical model. In our system, we introduce a concept map [8], which is an externalization tool called a content map. Content maps consist of nodes and links. Nodes represent topics and links show the relations among topics. Based on a logical model, causal and inclusive relations are prepared as links. We set 11 types of nodes based on a logical model: background, goal, method, evaluation, environment, objective, procedure, e-Method, result, consideration, and reason. Fig. 2 is an example of a content map. Each topic is labeled by its types, and related topics can be connected by either causal or inclusive relation links.



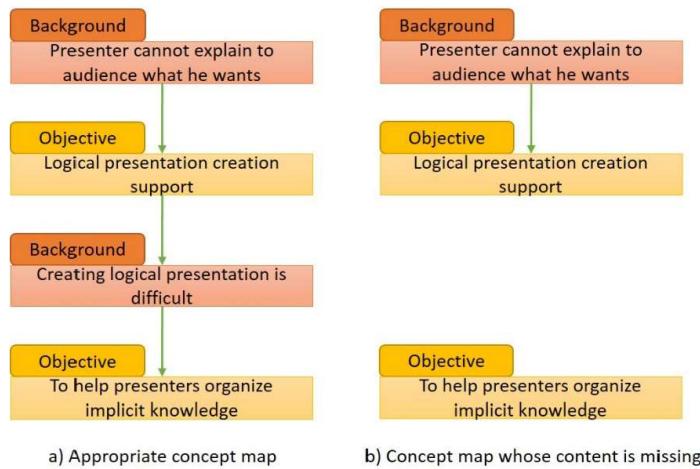
**Fig. 2 Content map**

A presenter sometimes is unable to create an appropriate contents map due to a 1) lack of topics and 2) incorrectly understanding relations. For such situations, a support function that organizes content map is useful. Several researches, which provide feedback about created concept maps [9], focus on learning activities to create correct content maps. For example, a teacher's concept map is prepared as a goal map and the differences between it and a student's concept map is given to students so they

can clarify their understanding. In our research, goal maps cannot be prepared, since only the presenter knows what he wants to represent in his content map. Its validity can only be evaluated by the presenter himself.

To increase the awareness of presenters of inappropriate relations and a lack of topics, our content map embeds a contents organization function that highlights the available contents that can be connected by the causal/inclusive relations with selected nodes. If no nodes are highlighted, no link is generated. Our content map allows links between pairs of node types to which relations are defined in the logical model. Based on this contents organization support function, a presenter can understand the necessary relations between topics in a logical presentation by observing the highlighted node types. A presenter can also notice a lack of topics if he realizes that the expected topics are not highlighted.

Here, we explain the contents organization function in Fig. 3. Assume that Fig. 3(a) suggests an appropriate content map for a presenter to create. Fig. 3(b) shows the current concept map that he created. When he wants to connect two goal nodes, our content map does not allow links between them since no links can be attached between goals in the logical mode. Based on this function, he might notice that he should create another node, such as background, to connect two goals.



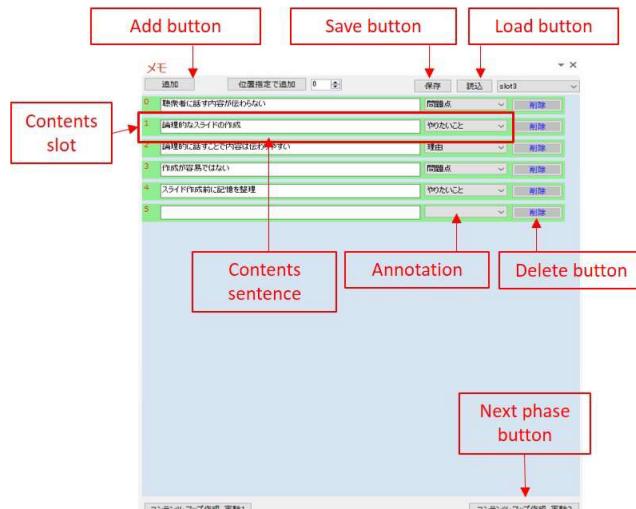
**Fig. 3 Situation contents organization support function**

#### 4 Content Map Organization Support System

We have developed a system that embeds content maps and a contents organization support function. This system, which is implemented in C# programming language, consists of two phases to create content map: contents description and contents organization. In the contents description phase, a user writes topics and annotates them. In

the contents organization phase, a user adds relations to the created contents and arranges all of the contents as a map. A user can rearrange these phases until all the topics are organized as a map.

Figure 4 shows the interface that supports the contents description phase. The contents slot represents individual topics, which consist of content sentences and annotation. A contents slot is added when the add button is pushed. Contents are created by adding a contents slot, filling in the contents sentences, and adding annotations. When the delete button next to the contents slot is pushed, the contents slot is deleted. When the next phase button is used, the interface for the contents organization phase appears and users can change to the next phase.

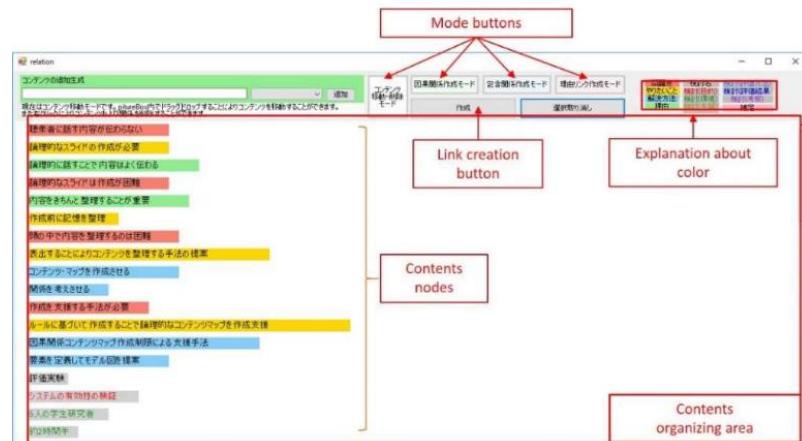


**Fig. 4 Interface for supporting contents description phase**

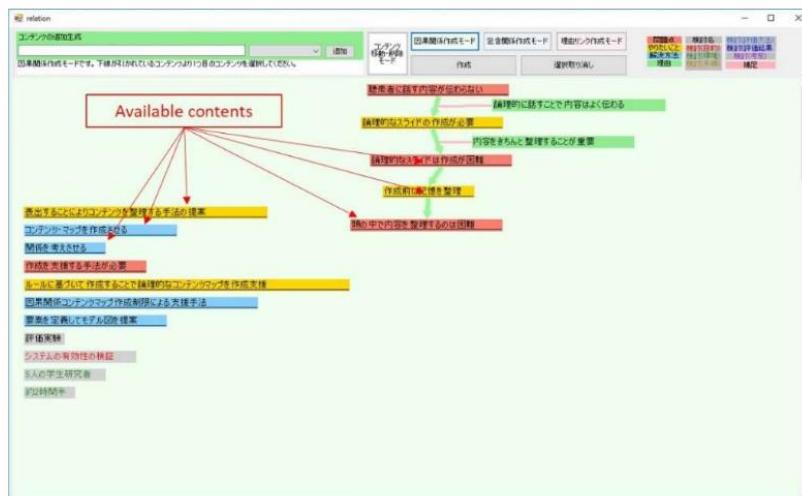
Figure 5 shows the initial state of the interface for the contents organization phase. Contents slots are transformed into contents nodes and arranged in the contents organizing area. Different colors are added to the contents nodes based on the attached annotations. For instance, backgrounds are depicted as red and methods as blue. To help a presenter recognize the meaning of these colors, explanations are shown on the top of the interface. In this interface, a presenter can drag and drop a contents node or remove a contents node by a right-button click with her computer's mouse.

A presenter changes modes to attach different relations. If a causal relation creation mode is selected, she can add causal relation links by selecting two contents nodes and pushing the link creation button. As in the contents organization support function, when selecting one contents node, contents nodes that have a causal relation with the selected one are underlined as available contents (Fig. 6). By selecting the second node from these underlined contents nodes, a link is attached that corresponds to the causal relation. However, if the contents node is selected without being underlined, no link is attached. In the same way, an inclusive relation link is attached in the

inclusive relation creation mode. Also, a reason for the link is attached by selecting types of contents and target links.



**Fig. 5 Interface for supporting contents organization phase**



**Fig. 6 Interface for indicating available contents**

## 5 Experiment

We experimentally evaluated our system's usability and the effectiveness of the contents organization support function. Four undergraduates and one graduate student (A

to *E*) created content maps of their research topics in the following three steps. All five subjects had done research in the computer science field for over a year with little experience making presentations.

**Step 1:** Create content maps using a comparison system (**CM 1**). In it, subjects can create contents nodes, but they cannot add annotations. No contents organization support function is embedded. That is, subjects can connect any two contents nodes without any restrictions.

**Step 2:** Create their content map again using our content map organization support system (**CM 2**). In this phase, they can see the content map that they created in Step 1.

**Step 3:** Answer questionnaires.

First, we discuss the effect of the contents organization support function for deriving contents nodes. Table 1 shows the number of contents nodes that were derived before and after creating each content map. Although the number of contents nodes did not increase for subjects A and B with CM 1, all of the subjects created extra contents for CM 2. This result indicates that with the contents organization support function, the subjects reconsidered their researches and found topics that are related to their topics.

**Table 1 Derived contents nodes**

<b>Subjects</b>	<b>Content map created in Step 1 (CM 1)</b>		<b>Content map created in Step 2 (CM 2)</b>	
	<b>Before</b>	<b>After</b>	<b>Before</b>	<b>After</b>
<b>A</b>	29	29	28	37
<b>B</b>	19	19	27	28
<b>C</b>	11	14	14	19
<b>D</b>	27	31	34	39
<b>E</b>	19	22	30	34

Next, we analyzed how well the contents organization support function represented appropriate relations. Even though identical contents nodes appeared in both CM 1 and CM 2, their relations are not identical. The following are the types of relation changes between the contents nodes that appeared in both CM 1 and CM 2:

**Type 1:** Relation in CM 2 was identical as in CM 1.

**Type 2:** New contents node was added in CM2 between two connected contents nodes in CM 1.

**Type 3:** Type of relations was changed.

**Type 4:** Direction of type was reversed.

**Type 5:** Relation was removed.

Table 2 shows the number of relations that correspond to these five types. For Types 2, 3, and 4, relations that were improved in CM 2 are shown in parentheses. Whether their relations improved was determined by the authors who were familiar with the research of these subjects. Pairs of contents nodes whose relations changed are Type 1, meaning that the subjects appropriately organized the contents nodes in CM 1. Only four relations of Type 2 were not improved: three relations of subject A and one relation of subject B. All other relations were appropriately changed while using the contents organization support function. Figure 7 shows an example of improved relations from CM 1 to CM 2. Since the content maps are written in Japanese, we attached English translations to each node. In CM 1, a causal relation link was attached from node **a** to node **b**. This relation was inappropriate, because both nodes were **methods**, and the **goal** that **method b** addressed was not mentioned. On the other hand, since **background d**, which was derived by **method a** and **goal c**, was added in CM 2, the reason for proposing **method b** clearly appeared. In this example, most relations were improved using contents organization function for Types 2, 3, and 4. This suggests that the contents organization support function is effective for logically arranging contents.

**Table 2 Types of relation changes: numbers in parentheses correspond to numbers of appropriately changed relations**

Subjects	Type 1	Type 2	Type 3	Type 4	Type 5
<b>A</b>	0	4 (1)	1 (1)	1 (1)	7
<b>B</b>	4	0 (0)	6 (6)	0 (0)	2
<b>C</b>	2	3 (3)	3 (3)	0 (0)	0
<b>D</b>	5	4 (3)	5 (5)	2 (2)	5
<b>E</b>	1	3 (3)	2 (2)	0 (0)	2

Figure 8 shows the part of CM 2 that was changed by Type 5. Subject D wanted to connect consideration f and goal g. In this case, consideration f contains a problem, so he regards it to be a background node. However, he failed to connect them by a causal relation link in CM 2 because they are not directly connected in the logical model. In our current logical model, problems that are derived by evaluations should be represented background, not consideration. However, in some cases, problems are sometimes contained in considerations. To cope with this drawback, two annotations

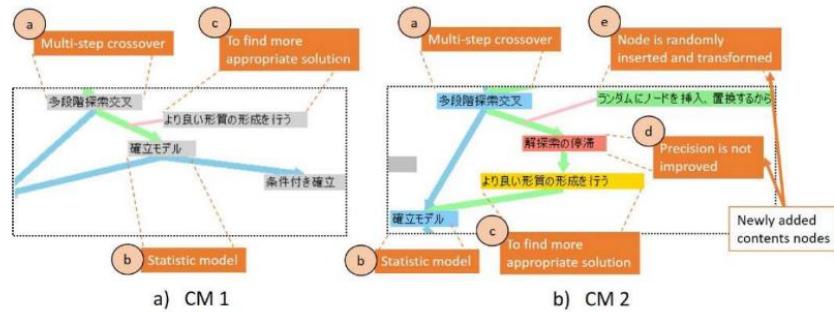


Fig. 7 Example of improved content map (Subject C)

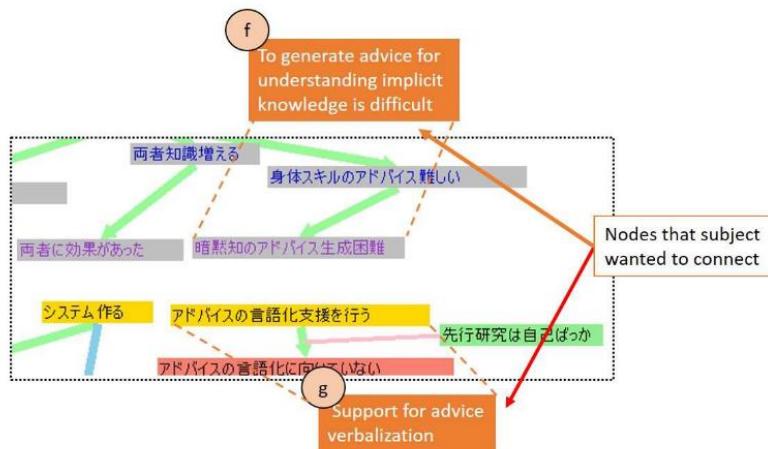


Fig. 8 Example of contents nodes for which subject failed to make a relation in CM 2 (Subject D)

can be added to one content. If the consideration contains a problem, both consideration and background should be attached to the contents node.

Table 3 indicates the questionnaire results and shows the numbers of subjects who selected individual items. Three subjects believed that they increased their understanding of their own research by creating content maps using our system and its contents organization support function. Subject E answered Yes: "By creating a content map with the system, I was encouraged to consider the relations among the existing contents and to find extra contents that were not externalized in the content map." Both of the subjects who answered No in question 1 made the following comment: "I'm not sure whether my understanding of my research deepened, but I was able to positively reflect on it." On the other hand, four subjects felt they understood their research more deeply when they added annotations. Subjects A commented: "I realized that I didn't correctly understand the role of the contents when I attached annotations to them."

Thus, both adding relations among contents and attaching annotations were effective for logically considering research contents.

**Table 3 Questionnaire results**

Questions	Yes	No
1. Did you understand your research more deeply by creating content map with our system?	3	2
2. Did you understand your research more deeply by attaching annotations?	4	1

## 6 Conclusion

This research supported presenters to create more logical presentations by organizing their topics. This paper proposed a logical model for topics in the research presentations of the computer science field and developed a concept map with a contents organization support function. Based on our experiment, our system effectively created new contents and organized them. However, we haven't yet evaluated the validity of our logical model. We must evaluate whether it appropriately represents the logical relations among topics in the computer science field.

Our system currently limits the available pairs to create links to encourage presenters to identify appropriate pairs for causal and inclusive relations. During evaluations, some presenters failed to create links because they could not derive appropriate types of nodes. To support such presenters, providing advice for generating specific types of contents might be effective. Therefore, future work will develop a function that determines the type of missing contents and encourage presenters to positively derive them.

## Acknowledgement

The work was supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) (No. 16H03089) and JSPS KAKENHI Grant-in-Aid for challenging Exploratory Research (16K12563)

## References

1. A. Kohlhase: "Semantic PowerPoint: Content and Semantic Technology for Educational Added-Value Services in MS PowerPoint," Proc. of World Conference on Educational Media and Technology, pp. 3576-3583 (2007)
2. R. Okamoto A. Tanikawa, and A. Kashihara: "Presentation Authoring Support considering Relationship between Slide Contents and Oral Expressions for Peer Review in Presentation Rehearsal," Proc. of eLearn2015, pp. 996-1001 (2015)
3. T. Kojiri, H. Nasu, K. Maeda, Y. Hayashi, and T. Watanabe: "Collaborative Learning Environment for Discussing Topic Explanation Skill Based on Presentation Slides," Proc. of 12th European Conference on e-Learning (ECEL2013), Vol. 1, pp. 199-208 (2013)
4. M. Kanakubo, and M. Hagiwara: "Creativity Support System Combining Morphological Analysis Method and Modified Input-Output Method," Joint International Conference on Soft

Computing and Intelligent Systems and 3rd International Symposium on Advanced Intelligent Systems (2002)

5. K. Nishimoto, S. Abe, T. Miyasato, and F. Kishino: "A System Supporting the Human Divergent Thinking Process by Provision of Relevant and Heterogeneous Pieces of Information Based on an Outsider Model," Proceedings of International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems", pp. 575-584 (1995)
6. A. Tanida, S. Hasegawa, and A. Kashihara: "Web 2.0 Services for Presentation Planning and Presentation Reflection," Proc. of International Conference on Computers in Education, pp. 565-572 (2008)
7. Y. Shibata, A. Kashihara, and S. Hasegawa: "Scaffolding with Schema for Creating Presentation Documents and Its Evaluation," Proc. of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, pp. 2059-2066 (2012)
8. John R. McClure: "Concept map assessment of classroom learning: Reliability, validity, and logistical practicality," Journal of research in science teaching, Vol. 36, pp. 475-492 (1999)
9. T. Hirashima, K. Yamasaki, H. Fukuda, and H. Funaoi: "Kit-Build Concept Map for Automatic Diagnosis," Artificial Intelligence in Education, Vol. 6738, pp. 466-468 (2011).

## A Framework to Generate Carrier Path Using Semantic Similarity of Competencies in Job Position

Wasan Na Chai<sup>1</sup>, Taneth Ruangraijitpakorn<sup>1,2</sup>,

Marut Buranarach<sup>1</sup>, and Thepchai Supnithi<sup>1</sup>

<sup>1</sup> Language and Semantic Technology Laboratory,

National Electronics and Computer Technology Center, Pathumthani, Thailand

{wasan.na\_chai, taneth.rua, marut.bur, thepchai.sup}@nectec.or.th

<sup>2</sup> Department of Computer Science, Faculty of Science and Technology,

Thammasat University, Pathumthani, Thailand

**Abstract.** A career path is necessary for students and workers to keep themselves in track for their career goal. However, a career path following a job standard in general is very rare. This paper presents a method to find a semantic similarity within competencies of job positions for realising a path to relate career. By a development of Thai WordNet containing terms used in competency description, a distance of classes of WordNet structure is used to determine a semantic similarity of competencies. Paths to relate job positions are assumed for the job positions sharing similar competencies, and the more they share, the more transferrable job is viable. From the usage scenario, the proposed framework proved that semantic of words is more useful than using character based similarity in competency comparison.

**Keywords:** Career Path \* Semantic Similarity \* WordNet \* Competency

### 1 Introduction

A career path is important for students and workers for planning their career towards their goal in life. In a career path, job positions are linked to other in two types, i.e. promotion and transferring job. A promotional path is a possible direction of going higher in the same job position while a transferring path is to change to another job branch. Moreover, changing a job is commonly necessary since people can be late-bloomer to find their own potential or become bored of the job. Changing to a job requiring competencies that a person already has will give him/her a good advantage in his/her new career. It is common in a company or an organization to provide a promotional path with certain criteria for its employees, but the job-transferring path is scarcely available.

There was a work mentioned on automatically generating a career path from a qualification data [1]. The work exploited certification names related to a job and their assigned competencies from Thailand Professional Qualification Institute (TPQI) [2] for linking paths to form a career path. By comparing certification names and competency names, they mainly exploited a string similarity to find commonness and assigned linking to similar concepts. They can demonstrate a career path based on a string similarity. From the result, there were some missing paths from the obtained career path since string similarity apparently overlooked the terms that are synonym or semantically related.

This paper aims to present an upgraded version of a framework to generate a career path by focusing on semantic of terms in a competency to increase coverage over the existing work. We expect to capture a semantic similarity in competencies in a different surface form of terms and create a relation between competencies. With similar competencies, job positions in a different career will be assigned with a path to inform as a transferrable job.

## 2 Background

### 2.1 Professional Qualification Standards

Professional or vocational standards or competency standards are the standards of performance individuals must achieve when carrying out functions in the workplace, together with specifications of the required knowledge and understanding [3]. They describe what an individual needs to do, know and understand in order to carry out a particular job role or function in his or her occupation. They are normally defined by a representative sample of employers and other key stakeholders and approved by the national qualification standard organization. Many countries have established occupational competence standards, to support skills development, employability and vocational education developments. This is to help enabling skills transfer and recognition, supporting skills credibility, and global economic competitiveness.

Professional qualification is a certification earned by a person to assure qualification to perform a job or task. They typically involve competence-based assessment. Competencies include all the related knowledge, skills, abilities, and attributes that form a person's job. Identifying employee competencies can contribute to improved organizational performance. Competency assessment criteria are normally based on national professional qualification standards. Qualification levels defined by the standards should be comparable between different industries. For example, a level 2 competency qualification in engineering will be a comparable achievement to a level 2 competency qualification in construction. Figure 1 shows relationship between competency, qualification of vocational competence and assessment criteria.



**Fig. 1** Relationship between competency, qualification of vocational competence and assessment criteria

## 2.2 Related Work

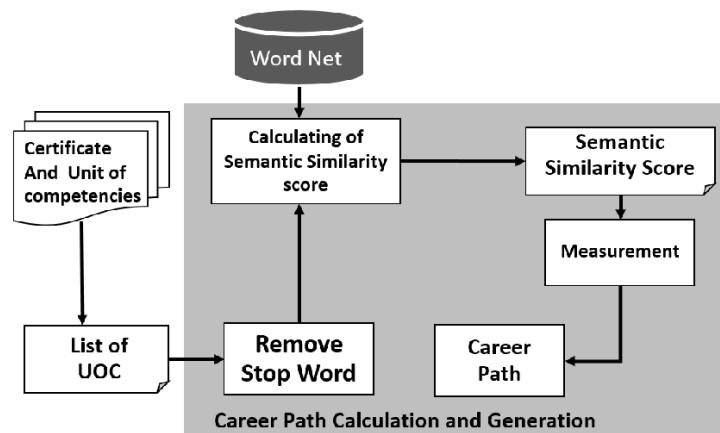
Previous work [1] presents a methodology for automatic career path generation based on certifications and their competencies provided by TPQI. The work attempted to find a path among task/role by using a similarity of characters in the certification names and competencies. The similarity score of this work is calculated by formula (1).

$$NMLCS_1(X_{i-1}, Y_{j-1}) = \frac{\text{length}(NMLCS_1(X_{i-1}, Y_{j-1}))^2}{\text{length}(X_{i-1}) \times \text{length}(Y_{j-1})} \quad (1)$$

By comparing common characters, the competencies indicated as similar were mapped together as a path to another job position. A percentage of non-union number of competencies is calculated to show a number of competences required to transfer to another career. By the experiment, this work shows that the method work fine in career path generation, but some path result are not accurate since the some paths were over-generated from long length similar characters although the competencies are not semantically related.

## 3 Methodology

This paper aims to find semantic similarity among units of competency (UoCs) for assigning a path to jobs in different career. Terms representing content (content word) are gathered and categorised into a hierarchical structure based on WordNet [4] formation as called Competency Net (CompNet). According to WordNet, this CompNet consists of a set of synonym terms called SynSet, and these SynSets are hierarchically structured in hypernym-hyponym relation. With CompNet, a semantic similarity of terms in UoC can be calculated. An overview of the framework is illustrated in Figure 2.



**Fig. 2** An overview of the proposed framework

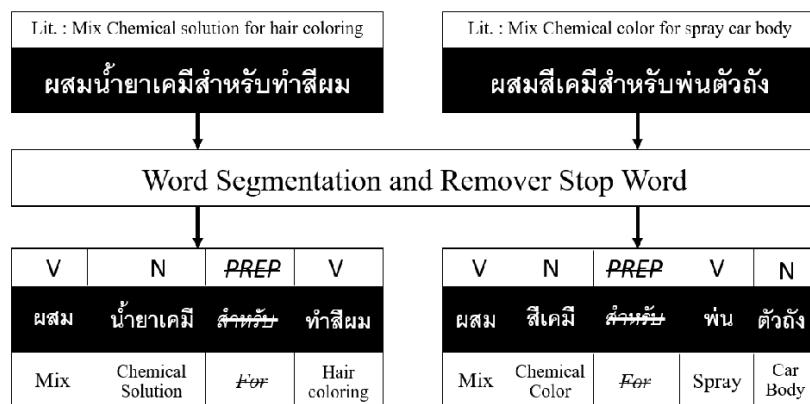
### 3.1 Pre-processing

Data from TPQI are structured as a profession for a root. For each profession, certifications are assigned to represent job position along with its level. Every certifications compose of units of competency (UoCs) which are all required to obtain a qualification for a certification. A schema of the data structure of TPQI is sketched in Figure 3.



**Fig. 3** A schema of TPQI data

An input of the framework is a list of UoCs belonging to a certain certification. UoCs are given as a Thai text describing skill and knowledge required for a job position. Since a key to inform technical skill and knowledge relies on with noun and verb, they are extracted by exploiting a word segmentation and POS tagger tool [5][6]. Hence, only nouns and verbs in UoC will later be processed in the further steps. An example of UoCs and its output from the preprocess is shown in Figure 4.



**Fig. 4** An example of the preprocess

### 3.2 CompNet Development

CompNet is a lexical resource developed according to a design of Thai WordNet [7][8]. The CompNet is designed as a representation of semantic existing specifically in UoCs.

However, some feathers provided in WordNet are not used in this work; hence, they are ignored in CompNet development such as definition and form derivation. CompNet consists of Thai terms existing in UoCs, and the terms are categorised into a synonym set (SynSet) related to one another in terms of hypernym-hyponym relation. It was carefully crafted following a hierarchy of SynSets in WordNet and extended with the specific terms mentioned in UoCs. Some lexical entries of CompNet (with their literal translation for more understanding) are illustrated in Figure 5.

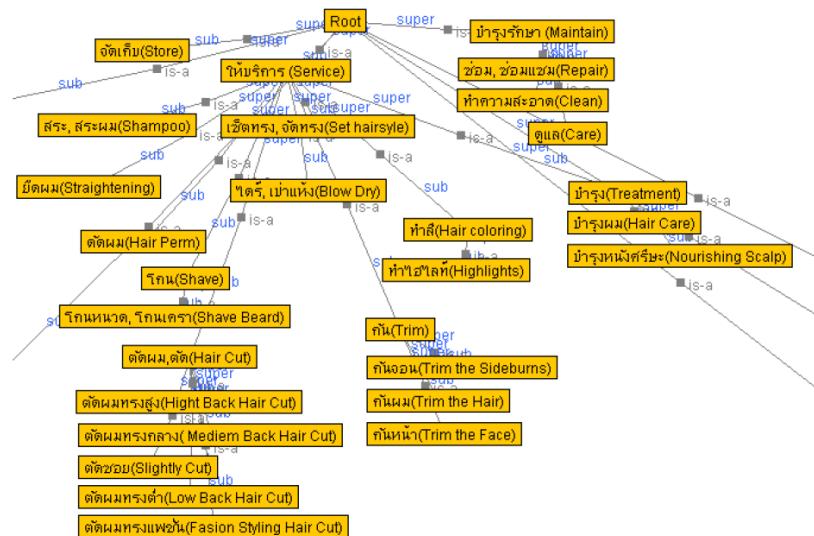


Fig. 5 Some parts of the defined CompNet

### 3.3 Semantic Similarity Calculation

From preprocess, nouns and verbs in UoCs are left over for calculating to semantic similarity score. In this work, we aim to compare an UoC with other UoCs of a certification of another profession to signify commonness in skill and knowledge to hint a transferrable path. Therefore, each word in UoC in a set will be compared to all other words of another UoCs from another certification. We exploit formula (2) to calculate a similarity score of a pair of UoCs.

$$Sim(UoC_1, UoC_2) = \frac{\sum_{i=1}^n Semantic\_Sim(\max(Wi_{UoC_1}, Wi_{UoC_2}))}{count(W_{total})} \quad (2)$$

Where :

UoC is a competency to be compared for similarity

W is a word in a competency

n is a number of word and  $W_{total}$  is the total amount

The CompNet is used to measure a distance between terms. We apply a semantic similarity measure introduced by Wu Z. and Palmer M. [9][10]. This semantic similarity measure focuses on the position of concepts in the taxonomy relatively to the position of the most specific common concept (their shared mother node). With the idea, the similarity between two concepts relies on the function of length and depth of the path.

Before further explanation, let's define notations as follows:

- 1)  $\text{len}(c_1, c_2)$ : the length of the shortest path from SynSet  $c_1$  to SynSet  $c_2$  in CompNet. In the case of the concepts of the same SynSet,  $\text{len}(c_1, c_2)=0$ .
- 2)  $\text{lso}(c_1, c_2)$ : the lowest common subsumer of  $c_1$  and  $c_2$
- 3)  $\text{depth}(c_1)$ : the length of the path to SynSet  $c_1$  from the global root concept while the depth at root is 1 and count onwards.
- 4)  $\text{deep\_max}$ : the max depth ( $c_i$ ) of the taxonomy

By using words in an UoC as a concept, the structure of CompNet can be used to measure semantic similarity. To calculate a semantic similarity using CompNet, we apply (3).

$$\text{Semantic\_Sim}(c_1, c_2) = \frac{2 * \text{depth}(\text{lso}(c_1, c_2))}{\text{len}(c_1, c_2) + 2 * \text{depth}(\text{lso}(c_1, c_2))} \quad (3)$$

From formula (3) it is noted that,

- (1) The similarity between two concepts ( $c_1, c_2$ ) is the function of their distance and the lowest common subsumer( $\text{lso}(c_1, c_2)$ ).
- (2) If the  $\text{lso}(c_1, c_2)$  is root,  $\text{depth}(\text{lso}(c_1, c_2))=1$ ,  $\text{Semantic\_Sim}(c_1, c_2) > 0$ ; if the two concepts have the same sense, the concept  $c_1$ , concept  $c_2$  and  $\text{lso}(c_1, c_2)$  are the same node.  $\text{len}(c_1, c_2)=0$ .  $\text{Semantic\_Sim}(c_1, c_2) = 1$ ; otherwise  $0 < \text{depth}(\text{lso}(c_1, c_2)) < \text{deep\_max}$ ,  $0 < \text{len}(c_1, c_2) < 2 * \text{deep\_max}$ ,  $0 < \text{Semantic\_Sim}(c_1, c_2) < 1$ .

In the case that a word amount from the comparing UoCs is different, Semantic\_Sim of  $c_1$  to  $c_2$  and  $c_2$  to  $c_1$  will be combined and divided by 2 for normalising the score. A word pair with the highest Semantic\_Sim score will be selected as aligned words from UoC<sub>1</sub> and UoC<sub>2</sub>.

#### 4 Usage Scenarios

This section mentions how the proposed method works against the string similarity score from the previous work [1]. The string similarity in comparison applies a calculation result from formula (1). There are two case scenarios.

The first scenario is the comparison of two UoCs shown in 2\*3 matrix in Figure 6. For traceability of an explanation, a related part of CompNet is also provided. This scenario shows that these UoCs gain a semantic similarity score for 0.88 while they obtain a string similarity score for 0.45. This case shows that these two UoCs share

several words with a same semantic. As for a calculation of the proposed method, we first calculated Semantic\_Sim score of each word from UoC<sub>1</sub> and UoC<sub>2</sub>. For example, the first word of UoC<sub>1</sub> and the first word of UoC<sub>2</sub> gain Semantic\_Sim as 1.0 since it found that they are the words in the same SynSet, and they are selected since they gain the highest score among all pairs. Another example is the third word of UoC<sub>1</sub> and words from UoC<sub>2</sub>. The word is not in a SynSet with other words so it is calculated using (3). Since a length of the word to the comparing words is 5 (counting from nodes), and  $\text{depth}(lso(c_1, c_2))$  is 1 since the lowest common subsumer of them is the root, the Semantic\_Sim score is calculated as 0.29. By applying (3), the result of score is 0.76 (from  $(1.0+1.0+0.29)/3$ ) for UoC<sub>1</sub> to UoC<sub>2</sub>. However, the number of words in UoC<sub>1</sub> and UoC<sub>2</sub> does not equal; hence, a similarity score of UoC<sub>2</sub> to UoC<sub>1</sub> are required to be calculated for bidirectional normalisation, and we obtain 1.0 as the score. Hence, the similarity score of UoC<sub>1</sub> and UoC<sub>2</sub> is 0.88 from normalising of 0.76 and 1.0. The scenario shows that this method can relate a semantic of words in different UoCs effectively while the string similarity calculation from the previous work ignores this pair of UoCs.

The semantic network diagram shows a hierarchical structure with the Root node at the top. Below it are several nodes connected by 'super' and 'sub' relationships. Some nodes are highlighted in yellow, including 'Service', 'Beautify', 'Blow Dry', 'Set Hair Style', and 'Fasion Hair Style'. The word matrix table compares three words from UoC<sub>1</sub> against three words from UoC<sub>2</sub>.

UoC 2		Word1	Word2
UoC 1		‘บีดร’ (Bowl Dry)	‘จัดทรง’ (Set Hair Style)
Word	Word		
word 1	เป่าแห้ง (Bowl Dry)	1	0.67
word 2	จัดแต่ง (Set Hair Style)	0.67	1
word 3	ห้องแฟชั่น (Fasion Hair Style)	0.29	0.29

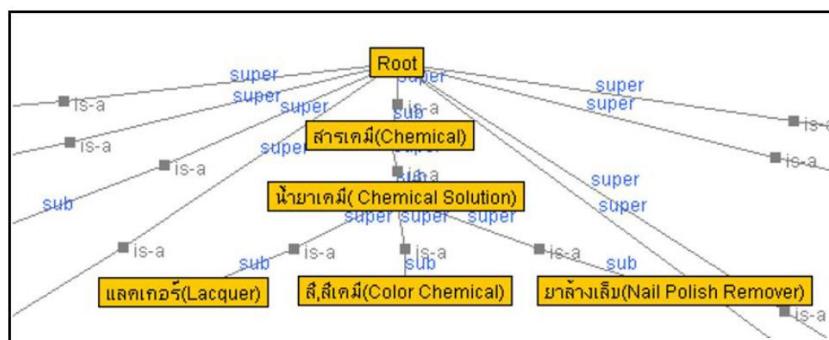
Below the table, calculations are shown:

- $\text{Sim } (\text{UoC1}, \text{UoC2}) \Rightarrow (1+1+0.29)/3 = 0.76$
- $\text{Sim } (\text{UoC2}, \text{UoC1}) \Rightarrow (1+1)/2 = 1$
- $\text{Semantic\_Sim Score} \Rightarrow (1+0.76)/2 = 0.88$

Fig. 6 Word matrix of first scenario and their details

The second scenario is the comparison of long UoCs. In this case, the words, which do not belong to POS noun and verb, are removed into 3\*4 word matrix as shown in Figure 7 and Figure 8. The interesting words in this case are ‘สีเคมี’ (chemical colour) and ‘น้ำยาเคมี’ (chemical solution). These terms are related in CompNet in the same tree while *chemical solution* is a superclass of *chemical colour*. In this case,  $\text{len}(c_1, c_2)$  of these words is 1, and  $\text{depth}(lso(c_1, c_2))$  is 3 since the lowest common subsumer is the class *chemical solution* itself at 3 levels from root. By the combination of the first two words are very high, but the difference of the rest words reduce the total score to 0.66. By the meaning, these two competencies do not supposed to be apparently related since the UoC<sub>1</sub> is about hair dying, and UoC<sub>2</sub> mentions about car painting. With the help of the rest words, these two UoCs are not recognised as related, but it could be better if we can recognise the difference in this specific technical terms automatically.

		UoC 2	Word1	Word2	Word3	Word4
			ผสุก (Mix)	สีเคมี (Chemical Color)	พ่น (Spray)	ตัวถัง (Body)
UoC 1	Word1	ผสุก (Mix)	1	0.25	0.29	0.29
Word1	Word2	น้ำยาเคมี (Chemical Solution)	0.29	0.85	0.25	0.25
Word2	Word3	ทำสีผม (Hair Coloring)	0.25	0.25	0.29	0.29
$\text{Sim } (\text{UoC1, UoC2}) \Rightarrow (1+0.85+0.29)/3 = 0.76$ $\text{Sim } (\text{UoC2, UoC1}) \Rightarrow (1+0.85+0.29+0.29)/4 = 0.61$ $\text{Semantic\_Sim Score} \Rightarrow (0.76+0.61)/2 = 0.67$						

**Fig. 7** Word matrix of second scenario**Fig. 8** Details of words and their CompNet for the second scenario

## 5 Conclusion and Future Work

In this paper, we propose a method to generate a career path using a semantic similarity of Thai words described in a competency for job position. To improve a capability of the previous work, semantic of the words is focused as the core for creating a path between job positions. CompNet (a hierarchy of words in competency following the concept of WordNet) was developed as a reference for semantic distance of the words. By calculating a score of words from different competencies, a total score of semantic similarity between UoCs is obtained. Paths to relate job positions are assumed for the job positions sharing a semantically similar competency. From the usage scenario, we found that the proposed method greatly improves the ability to generate a career path by exploiting semantic against the previous method using only string

similarity. More paths that were never recognised by the prior method are formed by the semantic.

In the future, we plan to focus on finding the terms mentioning specifically technical skill, knowledge and attitude in a competency since these terms uniquely represent specialty in their relative field and should not be related to other competencies. Moreover, we will use the result of this method for visualisation to support Thai students and workers in achieving career goal.

### Reference

1. Ruangrajitpakorn T., Na Chai W., Buranarach M., Supnithi T., Kongkachandra R.: An automatic Thai career path generation using similarity of roles and their competencies. In: 2015 International Symposium on Multimedia and Communication Technology. Phranakhon Si Ayutthaya, Thailand (2015)
2. Thailand Professional Qualification Institute Homepage, <http://www.tpqi.go.th/en/>
3. National Occupational Standards (NOS) Database, <http://nos.ukces.org.uk/>.
4. Miller G. A., Beckwith R., Fellbaum C. D., Gross D., Miller K.: WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244. (1990)
5. SegIt : Thai Word Segmentation Tool, <http://thaimt.org/lstnlp/wordseg.php>.
6. PosIt : Online Thai POS Tagger, <http://thaimt.org/lstnlp/pos.php>.
7. Thoongsup S., Robkop K., Mokarat C., Sinthurahat T., Charoenporn T., Sornlertlamvanich V., Isahara H.: Thai WordNet Construction. In: the 7th Workshop on Asian Language Resources, ACL-IJCNLP (pp. 139–144). Singapore (2009)
8. Leenoï D., Supnithi T., Aroonmanakun W.: Building a Gold Standard for Thai WordNet. In: International Conference on Asian Language Processing 2008, Chiang Mai, Thailand (2008)
9. Wu Z., Palmer M.: Verb semantics and lexical selection. In: Proceedings of 32nd annual Meeting of the Association for Computational Linguistics. June 27-30, Las Cruces, New Mexico (1994)
10. Meng L., Huang R., Gu J.: A Review of Semantic Similarity Measures in WordNet. In: International Journal of Hybrid Information Technology. Jan2013, Vol. 6 Issue 1, p1 (2013)

## Two stage travel salesman model of world tourism

Surafel Luleseged Tilahun and Jean Medard T Ngnotchouye

School of Mathematics, Statistics and Computer Science,  
University of KwaZulu-Natal, 3209, Pietermaritzburg, South Africa,  
surafelaau@yahoo.com

**Abstract.** Tourism can be defined as the commercial organization and operation of holidays and visits to places of interest. The need of visiting tourism places has increased through time with the practice of tourism and globalization. According to UNWTO's long term forecast Tourism Towards 2030, international tourist arrivals worldwide are expected to increase by 3.3% a year between 2010 and 2030 to reach 1.8 billion by 2030. Tourism contributes for the development of a country due to its contribution to GDP, employment, exports and investment. From a tourists perspective, a tourist needs to visit as many tourist destinations as possible with a budget constraint. In this paper, the problem is formulated as a travel salesman problem in which a tourist want to visit each of known tourist destination once. We assume that the cost of traveling is as a function of the distance traveled. Therefore, the problem will be minimization the travel time or distance while visiting the intended destinations. In addition tourist destinations are aggregated continent wise and geographically, as called clusters in this paper. Hence, the travel salesman problem will involve another travel salesman problem for each cluster of tourism destination. Prey predator algorithm will be used to solve the proposed approach.

**Keywords:** Tourism, travel salesman problem, two stage optimization, prey predator algorithm (PPA), metaheuristic

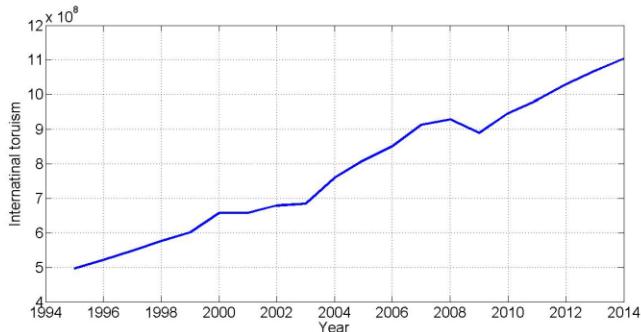
### 1 Introduction

Tourism can be recognized as long as people have traveled. Food, water, safety or acquisition of resources (trade) was the early travel motivations. But the idea of traveling for pleasure or exploration soon emerged. Travel has always been depending on technology, to provide the means or mode of travel. The earliest travelers walked or rode domesticated animals. The invention of the wheel and the sail provided new modes of transportation. Each improvement in technology increased individuals' opportunities to travel. As roads were improved and governments stabilized, interest in travel increased for education, sightseeing, and religious purposes. One of the earliest travel guides was written by Pausanias, a Greek, which was a 10 volume guide to Greece, for Roman tourists in 170 A.D [1]. However, there is no consensus concerning the definition of tourism.

The first definition of tourism was made by Guyer Feuler in 1905 [2]. He defined Tourism as “*A phenomenon unique to modern time which is dependent on the people's increasing need for a change and relaxing, the wish of recognizing the beauties of nature and art and the belief that nature gives happiness to human beings and which helps nations and communities' approaching to each other thanks to the developments in commerce and industry and the communication and transportation tools' becoming excellent.*”. In 1976, the Tourism Society of England defined Tourism as “*Tourism is the temporary, short-term movement of people to destination outside the places where they normally live and work and their activities during the stay at each destination. It includes movements for all purposes.*” [2]. Mathieson and Wall [3] created a good working definition of tourism as “*the temporary movement of people to destinations outside their normal places of work and residence, the activities undertaken during their stay in those destinations, and the facilities created to cater to their needs.*” According to [4] tourism is “*the sum of the phenomena and relationships arising from the interaction of tourists, business suppliers, host governments and host communities in the process of attracting and hosting these tourists and other visitors.*” But when it comes to explain it with the basic terms, we can sum it up as follows; Tourism is a collection of activities, services and industries which deliver a travel experience comprising transportation, accommodation, eating and drinking establishments, retail shops, entertainment businesses and other hospitality services provided for individuals or groups traveling away from home not more than one consecutive year.

The number of international tourists has increased through time. Figure 1, shows the total number of incoming international tourists in the world. United nations world tourism organization’s (UNWTO’s) long term forecast Tourism Towards 2030 forecasts that the international tourist arrivals worldwide are expected to increase by 3.3% a year between 2010 and 2030 to reach 1.8 billion by 2030 [5]. In addition, according to The World Tourism Organization (WTO) 2014 report, tourism is currently the world largest industry with annual revenues of over 1.5 trillion dollars [6].

Transportation is the critical component of tourism. The basic questions a traveler looks at are the time needed to travel to the destination and its corresponding traveling cost. Tourism developments are dependent on the ease of access and types of transportation available. Besides to this, the number of tourist has increased from time to time to visit different destination as the means of transportation improves with technology. However, most of the time tourist destination are not visited by minimizing either cost or time rather through suggestions without considering the minimum distance or cost due to lack of information about the destination and appropriate solution method to have minimum total distance or minimum total cost. For a tourist visiting a new country, visiting all major tourist destinations in the country and neighboring countries with minimum cost and minimum time is an ideal objective to achieve. Furthermore, this problem can be formulated as a travel salesman problem. Hence, in this paper a two stage travel salesman model of visiting major world tourism sites



**Fig. 1.** The number of international tourists through time based on the data from [7]

will be formulated and a metaheuristic algorithm, called prey predator algorithm [8, 9], will be used as a solution method along with enumerative approach.

The next section discusses preliminaries which includes a brief introduction to travel salesman problem and prey predator algorithm, followed by the problem formulation in section 3. Section 4 discusses simulation results and discussion followed by a conclusion in section 5.

## 2 Preliminaries

### 2.1 Travel salesman problem

The objective of Traveling Salesman Problem (TSP) is to find a tour of a given number of  $n$  cities, visiting each city exactly once and return back to the starting or initial city where the length of the tour is minimized. Currently the only known method guaranteed to optimally solve the traveling salesman problem is by enumerating each possible  $\frac{(n-1)!}{2}$  numbers of tour and search for the tour with smallest distance [10]. However, this method of computing is impossible if the number of cities,  $n$ , is large i.e. the number of tours increases factorial times. Therefore, even though TSP is an easy to understand problem, it is very difficult to solve. Researchers have proved that traveling salesman problem is Non-deterministic polynomial time complete problem (NPC) [11, 12, 13]. Many researches have been done to investigate a method to solve TSP. One of them is Brute-force method, which is based on selecting the minimum distance out of all computed tours [14, 15]. Using Brute-force method to solve TSP with a large number of cities can be frustrating and even take years or centuries to solve. In addition, it is inefficient as  $n$  gets large. Nearest Neighborhood approach is one of approximation method which is based on selecting the consequence nearest city [14, 15]. However, even though it works well and gives a good approximation solution, it does not always give the best solution as brute-force method. Greedy approach is also another type of approximation approach which is based on

choosing the  $n$  minimum distance and Hamiltonian cycle in order to find an optimal solution [16, 15]. If the tour doesn't have Hamiltonian cycle, then there is no optimal solution. This is one of the limitations of Greedy approach. Branch and bound method of solving based on dividing a problem into a number of sub problems [16, 15]. It is a system for solving a sequence of sub-problems each of which may have multiple possible solutions and where the solution chosen for one sub-problems. The algorithm depends on the efficient estimation of the lower and upper bounds of a region or branch of the search space and approaches enumeration as the size ( $n$ -dimensional volume) of the region tends zero [17].

However, all the above methods are difficult to apply when the numbers of cities increase, because it becomes impossible to find the minimum distance of every tour in polynomial time. Hence, most of recent studies use metaheuristic algorithms to find a reasonable solution within reasonable time. These algorithms have become very useful for difficult problems, especially for NP-hard optimization problems such as the traveling sales man problem.

## 2.2 Prey predator algorithm

Prey predator algorithm is swarm based metaheuristic algorithm which is inspired by the interaction between a predator and its prey [9, 8]. The algorithm has been tested in different applications and compared with other algorithm providing a promising results [9, 8, 18, 19, 20, 21, 22]. The algorithm works by categorizing the initial feasible solutions as predator and prey and updates the solutions mimicking the predator prey scenario. A solution is said to be predator if its performance is the least among current solution set, a solution is said to be best prey if its performance is better than all the current solution in the given iteration and the rest of the solutions are called ordinary prey. The predator works as an agent for exploration where it explore the solution space with bigger step length and also force the ordinary prey to explore the solution space by scaring them. The best prey, in the other hand, works as an agent for exploitation. It is considered as a prey which found a hiding place and no more affected by the predator. Hence, it totally does an exploitation of its neighborhood with smaller step length and also attracts other ordinary prey to explore its neighborhood. The movement of the ordinary prey depends on the probability of follow-up, an algorithm parameter. If a random number is less than or equal to the probability of follow-up then the prey will follow other prey with better performance compared to itself and if the random number is greater than the probability of follow-up it will run randomly away from the predator. The algorithm is originally proposed for continuous problems. Therefore, appropriate modifications need to be made to make it suitable for the travel salesman problem.

- 1. The first step is generating  $N > 2$  random feasible solutions. That will be the order of cities to be visited. Hence, it is a permutation of  $n$  natural numbers, where  $n$  is the number of cities. For instance in MATLAB, `randperm(n)` will generate random order of the  $n$  natural numbers which represent the cities.

- 2. The second step will be evaluating the performance of the randomly generated solutions based on the objective function  $f(x)$ , which can be either the sum of the cost of traveling between the consecutive cities or the sum of the distance between the consecutive cities. Based on the performance of the solutions on  $f(x)$ , the solutions will be categorized as best prey,  $x_b$ , which is the solution with better performance among all the  $N$  solutions; predator,  $x_p$ , which is the solution with the worst performance and the rest ordinary prey,  $x_i$ .
- 3. The third step will be the updating process. There will be three updating mechanism for the three classes of problems. An algorithm parameter called probability of change,  $P_c$ , will be used. For the predator, it needs to experience a good exploration behavior by changing a lot to explore different region of the solution space. Hence, bigger value of probability of change needs to be assigned and the update will be done as given in table 1.

```

for i = 1 : n
    if rand ≤ Pc
        change xp(i) randomly
        repair the solution to be feasible
    end
end

```

**Table 1.** Prey predator algorithm - updating the predator

In a similar manner, the updating of the best prey will be done with smaller probability of change,  $P_c$ .

The updating mechanism of ordinary prey,  $x_i$ , is done based on an algorithm parameter, called probability of follow-up,  $P_f$ . If a random number is less than or equal to the probability of follow-up, then it will follow better performing solutions and if it is greater than the probability of follow-up then the solution will randomly run away from the predator. A sample code is given in table 2, (suppose that the solutions are sorted in such a way that for a solution  $x_i$  all solution with higher index performs better and  $x_n$  is  $x_b$ ).

- 4. The fourth step is checking for termination criteria. If a maximum number of iteration is achieved or no improvement is recorded in consecutive iterations or a pre given tolerance target is archived then the algorithm will terminate, otherwise it will go back to step two with the updated solutions.

### 3 Two stage model of world tourism sites

A tourist would be happy to visit all the nearest tourism destination in visiting a country. The travel cost will be less compared to returning some other time to visit some of these places. For instance, for a tourist from Europe visiting Brazil,

```

if rand ≤ Pf
    for j = i + 1 : N
        for k = 1 : n
            if rand ≤ Pc
                xi(k) = xi(k) + round(rand(xj(k) - xi(k)))
                repair xi for feasibility
            end
        end
    end
else
    for k = 1 : n
        if rand ≤ Pc
            xi(k) = xi(k) + round(rand(xi(k) - xp(k)))
            repair xi for feasibility
        end
    end
end

```

**Table 2.** Prey predator algorithm - updating an ordinary prey  $x_i$ 

it would be easier to visit Chili as well compared to going back to Europe and comeback some other time to visit Chili, in terms of cost. By putting tourist destinations which are close to each other, a number of tourism clusters can be made. That will be the first stage of the model. Hence, the first stage will be optimizing clustered tourist destinations separably whereas the second stage will be finding the optimal route between these clusters. Looking at figure 2, which is the worldwide tourism arrivals in 2014, continent wise clustering is one possible way of clustering based on the geographical location of the destinations.

In order to determine the best route of these tourist destinations, the destinations are clustered according to their geographical location, specifically continent wise. The main advantage doing so is to divide the problem into sub problems and a mixed approach of heuristics and deterministic methods will be used. This intern reduce the premature convergence of the approaches to a local solution. It also help to visualize the performance of the approach on each sub problems so that additional modifications can be implemented on the components of the bigger problem, if needed.

The cluster is done based on the top tourist destinations in 2013 and 2014 continent wise as given in [23]. The distance is measure using an online tool between the capital cities (which in turn gives the distance between the airports of the cities) of the countries.

- Cluster 1: The first cluster is top ten tourist destinations in Africa. The list includes Morocco(Mo), South Africa(SA), Tunisia(Tu), Algeria(Al), Mozambique(Mz), Zimbabwe(Zi), Kenya(Ke), Uganda(Ug), Namibia(Nm) and Senegal(Sg). The distance between these countries, between the capitals is com-

## INTERNATIONAL TOURISM 2014

International tourist arrivals (ITA): 1133 million  
International tourism receipts (ITR): US\$ 1245 billion



**Fig. 2.** Tourist arrival worldwide in 2014 [5]

puted using an online tool from [24] using distance as a crow flies option. The distance is given in table 3

	Mo	SA	Tu	Al	Mz	Zi	Ke	Ug	Nm	Sg
Mo	-	4974.329	979.160	591.253	4887.456	4366.735	3736.568	3454.310	4216.508	1485.775
SA	-	-	4920.761	4989.608	1010.201	1360.506	2552.361	2540.466	789.324	4111.926
Tu	-	-	-	394.831	4580.086	4016.272	3138.937	2902.363	4131.500	2283.846
Al	-	-	-	-	4746.763	4195.347	3407.493	3150.857	4205.990	1978.865
Mz	-	-	-	-	-	571.839	1730.759	1817.826	1004.626	4387.961
Zi	-	-	-	-	-	-	1209.323	1258.824	963.944	4000.786
Ke	-	-	-	-	-	-	-	311.611	1981.547	3872.585
Ug	-	-	-	-	-	-	-	-	1896.388	3561.145
Nm	-	-	-	-	-	-	-	-	-	3482.861

**Table 3.** Distance in miles between countries for cluster 1 based on distance as a crow flies in [24]

if  $x = (x_1 x_2 \dots x_{10})$  is a candidate solution then its performance based on the fitness function of cluster 1, which is sum of distance is given by  $f_1(x)$ . The corresponding optimization problem is given in equation 1.

$$\begin{aligned}
 \min f_1(x) &= \sum_{i=1}^9 d(x(i), x(i+1)) \\
 \text{s.t. } x(i) &\in \{1, 2, \dots, 10\} \forall i \\
 x(i) &\neq x(j) \forall i, j
 \end{aligned} \tag{1}$$

where  $d(x(i), x(i+1))$  is the distance between cities  $x(i)$  and  $x(i+1)$ , and is obtained from table 3.

- Cluster 2: The second cluster is top ten tourist destinations in the middle east. The list includes Saudi Arabia (SaA), United Arab Emirates (UAE), Egypt (Eg), Iran (Ir), Jordan (Jr), Israel (Is), Qatar (Qa), Oman (Om), Lebanon (Ln) and Bahrain (Bh). The distance between these countries, between the capitals is computed using an online tool from [24] using distance as a crow flies option ad done for cluster 1. The distance is given in table 4

	SaA	UAE	Eg	Ir	Jr	Is	Qa	Om	Ln	Bh
SaA	-	502.916	1021.047	815.935	827.833	857.868	303.914	749.528	929.786	265.688
UAE	-	-	1494.349	812.696	1258.261	1295.421	211.755	247.533	1332.409	289.714
Eg	-	-	-	1234.376	307.891	264.305	1738.441	365.523	1207.833	
Ir	-	-	-	-	926.901	971.018	720.170	938.624	912.094	657.497
Jr	-	-	-	-	-	44.424	1049.894	1494.746	136.935	968.579
Is	-	-	-	-	-	-	1086.285	1533.125	146.493	1005.714
Qa	-	-	-	-	-	-	-	457.777	1128.684	86.436
Om	-	-	-	-	-	-	-	-	1561.814	530.683
Ln	-	-	-	-	-	-	-	-	-	1044.658

**Table 4.** Distance in miles between countries for cluster 2 based on distance as a crow flies in [24]

The corresponding optimization problem is given in equation 2.

$$\begin{aligned} \min f_2(x) &= \sum_{i=1}^9 d(x(i), x(i+1)) \\ \text{s.t. } x(i) &\in \{1, 2, \dots, 10\} \forall i \\ x(i) &\neq x(j) \forall i, j \end{aligned} \tag{2}$$

- Cluster 3: The third cluster is top ten tourist destinations in the Americas. The list includes United States (USA), Mexico (Mx), Canada (Ca), Brazil (Br), Argentina (Ar), Dominican Republic (DR), Colombia (Co), Chile (Cl), Puerto Rico (PR) and Peru (Pe). Similarly with what is done above the distance is given in table 5

The corresponding optimization problem is given in equation 3.

$$\begin{aligned} \min f_3(x) &= \sum_{i=1}^9 d(x(i), x(i+1)) \\ \text{s.t. } x(i) &\in \{1, 2, \dots, 10\} \forall i \\ x(i) &\neq x(j) \forall i, j \end{aligned} \tag{3}$$

	USA	Mx	Ca	Br	Ar	DR	Co	Cl	PR	Pe
USA	-	1886.052	455.915	4018.346	5223.204	1476.550	10202.967	5023.106	1558.797	3527.726
Mx	-	-	2241.731	3968.056	4597.907	1907.787	10495.036	4113.431	2158.075	2648.958
Ca	-	-	-	4386.882	5640.669	1893.677	2822.177	5466.258	1944.440	3979.356
Br	-	-	-	-	1465.447	2574.170	2338.940	1774.546	2459.885	1697.637
Ar	-	-	-	-	-	3748.972	2901.450	706.697	3705.894	1949.137
DR	-	-	-	-	-	-	1005.092	3592.727	252.154	2168.582
Co	-	-	-	-	-	-	-	2642.114	1109.175	1168.062
Cl	-	-	-	-	-	-	-	-	3604.858	1532.994
PR	-	-	-	-	-	-	-	-	-	2241.715

**Table 5.** Distance in miles between countries for cluster 3 based on distance as a crow flies in [24]

- Cluster 4: The fourth cluster is top ten tourist destinations in Asia and the Pacific. The list includes China (Chn), Hong Kong (Hk), Malaysia (My), Thailand (Th), Singapore (Sng), Macau (Mc), South Korea (SK), Japan (Jp), Taiwan (Tw) and Vietnam (Vt). Similarly with what is done above the distance is given in table 6

	Chn	Hk	My	Th	Sng	Mc	SK	Jp	Tw	Vt
Chn	-	1219.6782	2705.825	2050.737	2782.607	1237.321	595.528	1310.636	1069.403	1447.287
Hk	-	-	1567.015	1074.461	1606.573	40.925	1300.936	1798.767	506.300	542.953
My	-	-	-	738.857	192.381	1539.963	2867.826	3315.813	2013.685	1268.961
Th	-	-	-	-	887.544	1036.954	2313.353	2869.739	1578.177	614.186
Sng	-	-	-	-	-	1583.479	2903.475	3311.353	2024.287	1368.209
Mc	-	-	-	-	-	-	1328.058	1834.706	545.725	502.028
SK	-	-	-	-	-	-	-	723.918	916.990	1703.185
Jp	-	-	-	-	-	-	-	-	1309.258	2287.386
Tw	-	-	-	-	-	-	-	-	-	1038.470

**Table 6.** Distance in miles between countries for cluster 4 based on distance as a crow flies in [24]

The corresponding optimization problem is given in equation 4.

$$\begin{aligned}
 \min f_4(x) &= \sum_{i=1}^9 d(x(i), x(i+1)) \\
 \text{s.t. } x(i) &\in \{1, 2, \dots, 10\} \forall i \\
 x(i) &\neq x(j) \forall i, j
 \end{aligned} \tag{4}$$

- Cluster 5: The fifth cluster is top ten tourist destinations in Europe. The list includes France (Fr), Spain (Sp), Italy (It), Turkey (Tky), German (Gr), United Kingdom (UK), Russia (Ru), Austria (Aus), Greece (Grc) and Poland (Pld). Similarly with what is done above the distance is given in table 7

	Fr	Sp	It	Tky	Gr	UK	Ru	Aus	Grc	Pld
Fr	-	654.934	687.599	1601.585	546.237	213.228	1546.676	642.900	1303.877	850.140
Sp	-	-	848.579	1896.347	1163.086	785.458	2140.123	1125.628	1474.100	1424.422
It	-	-	-	1048.522	735.882	891.516	1477.627	475.408	653.793	818.003
Tky	-	-	-	-	1261.673	1750.901	1135.356	986.129	479.923	1020.517
Gr	-	-	-	-	-	579.820	1000.470	325.702	1121.617	321.446
UK	-	-	-	-	-	-	1555.672	768.276	1487.512	901.108
Ru	-	-	-	-	-	-	-	1038.672	1387.805	715.825
Aus	-	-	-	-	-	-	-	-	798.063	345.538
Grc	-	-	-	-	-	-	-	-	-	994.093

**Table 7.** Distance in miles between countries for cluster 5 based on distance as a crow flies in [24]

The corresponding optimization problem is given in equation 5.

$$\begin{aligned} \min f_4(x) &= \sum_{i=1}^9 d(x(i), x(i+1)) \\ \text{s.t. } x(i) &\in \{1, 2, \dots, 10\} \forall i \\ x(i) &\neq x(j) \forall i, j \end{aligned} \quad (5)$$

It should be noted that in any of the cluster the initial city doesn't affect the routes. For example if (1 3 2 7 4 10 8 9 5 6) is a solution means the best route is from Fr to It to Sp to Ru to Tky to Pld to Aus to Grc to Gr to UK and back to Fr. It is equivalent if we shift the values without changing the order, i.e. the solution is equivalent with (4 10 8 9 5 6 1 3 2 7). Therefore, it is possible to reduce the decision variable by fixing the starting city. Hence, one of our assumption is the tourist will start and finish the journey in a city. In general the first stage of the problem, for cluster  $k$ , is given in equation 6.

$$\begin{aligned} \min f_k(x^{(k)}) &= \sum_{i=1}^9 d(x^{(k)}(i), x^{(k)}(i+1)) \\ \text{s.t. } x^{(k)}(i) &\in \{1, 2, \dots, 10\} \forall i \\ x^{(k)}(1) &= 1 \\ x^{(k)}(i) &\neq x^{(k)}(j) \forall i, j \end{aligned} \quad (6)$$

Each of the five clusters will be solved separately as they are independent to each other.

The second stage of the problem is minimizing the distance to visit each clusters only once. The minimum possible distances of the clusters is given in table 8. The distance between cluster  $I$  with city indexes  $i$  and cluster  $J$  with city indexes  $j$  is computing by solving  $\min\{d(x^{(I)}(i), x^{(J)}(j))\}$ .

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	-	1297.528	3095.949	4486.205	373.141
Cluster 2	-	-	6012.621	2821.883	439.243
Cluster 3	-	-	-	6418.562	3335.106
Cluster 4	-	-	-	-	3602.180

**Table 8.** Distance in miles between the clusters based on distance as a crow flies in [24]

The nearest distance between the clusters are found to be between Tunisia & Egypt; Senegal & Brazil; Kenya & Thailand; Tunisia & Italy; Egypt & Puerto Rico; Oman & Thailand; Lebanon & Turkey; Canada & Japan; Canada & London and Russia & China. In a similar formulation done for the first stage, the mathematical formulation of the second stage which is a travel salesman model of the clusters is given in equation 7.

$$\begin{aligned}
 \min F(y^{(q)}) &= \sum_{i=1}^4 d(y^{(q)}(i), y^{(q)}(i+1)) \\
 \text{s.t. } y^{(q)}(i) &\in \{1, 2, \dots, 5\} \forall i \\
 y^{(q)}(1) &= 1 \\
 y(i) &\neq y^{(q)}(j) \forall i, j
 \end{aligned} \tag{7}$$

## 4 Simulation results

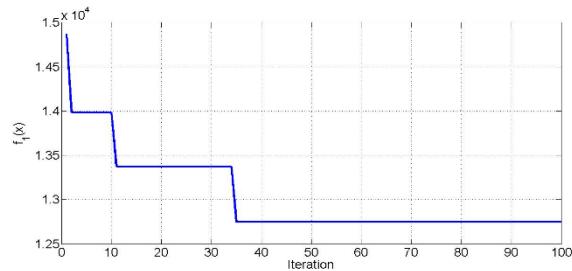
### 4.1 Simulation setup

The simulation is done using *MATLAB* 7.10.0 (*R2010a*) on *Intel(R) Core(TM) i3 – 3110M CPU 2.4 GHz, RAM 6GB and 64 bit* operating system machine. The algorithm parameters are tuned based on recommendation in the continuous problem cases from [9, 8, 18, 19, 20, 21]. The probability of follow-up and probability of change for exploration and exploitation or local search are assigned with 0.75, 0.75 and 0.25. The number of initial solutions are set to be 25 for all the simulations with a maximum number of iteration as termination criteria. The maximum number of iteration is set to be 100 for all the simulations.

### 4.2 Results and discussion

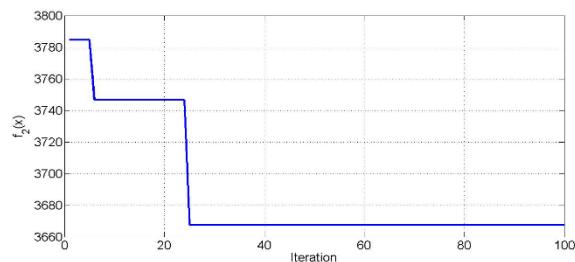
First the algorithm run to solve the five problems in the first stage. For the ten countries in each cluster there are  $(n - 1)! = 362,880$ . This is too much for checking each possibilities, as a result prey predator algorithm is implemented as discussed in section 4.1

The result found after solving cluster 1 is given by  $x^* = (1\ 4\ 3\ 8\ 7\ 6\ 5\ 2\ 9\ 10\ 1)$ . This means the order of the best tour is Morocco - Algeria - Tunisia - Uganda - Kenya - Zimbabwe - Mozambique - South Africa - Namibia - Senegal - Morocco with a total distance of about 12749 miles. The performance graph is given in figure 3



**Fig. 3.** Simulation result for cluster 1

$x^* = (1\ 2\ 8\ 4\ 9\ 5\ 6\ 3\ 7\ 10\ 1)$  is the final result after running the algorithm for cluster 2. Hence, the optimal route found is Saudi Arabia - United Arab Emirates - Oman - Iran - Lebanon - Jordan - Israel - Egypt - Qatar - Bahrain - Saudi Arabia. The final functional value is found to be 3663.3 miles. The performance graph on cluster 2 is given in figure 4.



**Fig. 4.** Simulation result for cluster 2

After running the algorithm for the third cluster the optimal solution is approximated by  $x^* = (1\ 3\ 2\ 10\ 8\ 5\ 4\ 7\ 6\ 9\ 1)$ , which gives the route United States - Canada - Mexico - Peru - Chile - Argentina - Brazil - Colombia - Dominican Republic - Puerto Rico - United States. The total distance is 14207. The performance of the algorithm as a function of the iteration is given in figure 5.

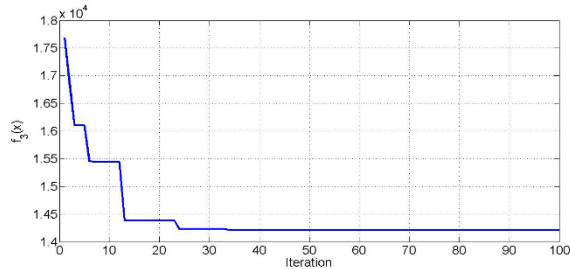


Fig. 5. Simulation result for cluster 3

The optimal solution found after running the algorithm for cluster 4 is (1 10 4 3 5 6 2 9 8 7 1) with functional value of 7752.1. Therefore, the order of the countries is given by China - Vietnam - Thailand - Malaysia - Singapore - Macau - Hong Kong - Taiwan - Japan - South Korea - China. The performance graph is given in figure 6.

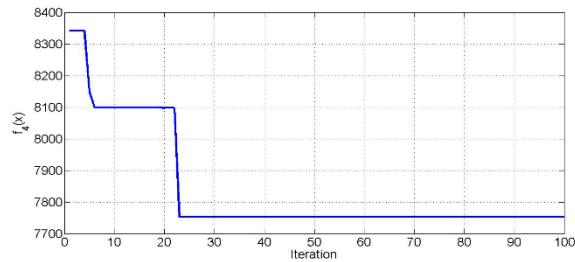


Fig. 6. Simulation result for cluster 4

The result for cluster 5 is given by  $x^* = (1 \ 6 \ 5 \ 8 \ 10 \ 7 \ 4 \ 9 \ 3 \ 2 \ 1)$ , with functional value of 5952.7. The order of the countries is given by France - United Kingdom - German - Austria - Poland - Russia - Turkey - Greece - Italy - Spain - France. The performance graph is given in figure 7.

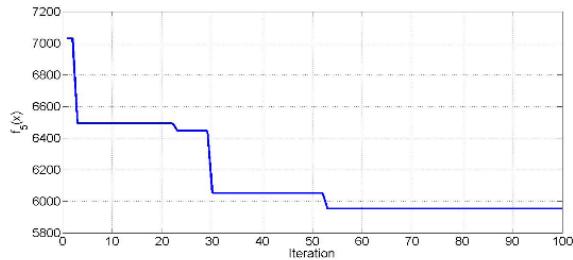


Fig. 7. Simulation result for cluster 5

The next step is solving the second stage of the problem which is solving a travel salesman model for the clusters. Since there are 5 destinations or clusters, the total number of possible routes is 24. Hence, enumerating approach in which the cost of each route can be calculated to find the best route. A MATLAB code is written to check the cost of these costs and the optimum solution is found to be (1 3 4 2 5 1), implying the best tour is cluster 1 - cluster 3 - cluster 4 - cluster 2 - cluster 5 - cluster 1, i.e. Africa - Asia and Pacific - Middle East - Europe - Africa. The optimum functional value is found to be 13148.778.

The assumption in this study is that it is possible to go from any country to another in the list. Furthermore, the distance between two countries is measured by the distance between their capital cities. Perhaps, the tourist destination may not be the capital city and the distance may change. That is for possible future work along with adding another stage in which different tourist destinations within a country is formulated in a similar way. Furthermore, actual traveling distance or time can replace the distance values for accurate and better results, as our simulation is based on a direct distance between the capitals of the countries from an online tool. It should also be noted that it is like a case study where a complex similar problem with many cities or countries can be solved with the same approach.

## 5 Conclusion

In this study the world top ten tourist destination of different regions are clustered and formulated as two staged travel salesman problem. The first stage finding the minimum sum of distance between capital cities in the same cluster. The second stage is solving a travel salesman formulation of visiting each of the clusters once with the same initial and final destination. A swarm based metaheuristic algorithm called prey predator algorithm is used. The algorithm is originally designed for continuous problems, but modified to suite the problem at hand. The simulation results shows that it is a promising approach. However, another stage of finding the best route for different tourist destinations in a city can be further studied.

## Bibliography

- [1] A. Satta. Towards sustainable tourism in the mediterranean: Situation and challenges at a regional level. In *Regional experience sharing workshop: Moving towards sustainable tourism in Medeterranean MPAs, 24 -25, Novemebr 2015, Sinis, Italy.* [Online] accessed May 2016 from [http://www.medpan.org/documents/10180/0/presentation\\_tourism\\_25nov15\\_satta/0c6a1ee4-075b-4eea-9e75-052db62e9208](http://www.medpan.org/documents/10180/0/presentation_tourism_25nov15_satta/0c6a1ee4-075b-4eea-9e75-052db62e9208), 2015.
- [2] C. Bogahavatta. Introduction to tourism and cultural resources. In [http://si.archaeology.lk/materials/Introduction-to-Tourism-and-Cultural-Resources\\_2013\\_04\\_01.pdf](http://si.archaeology.lk/materials/Introduction-to-Tourism-and-Cultural-Resources_2013_04_01.pdf). 2013.
- [3] A. Mathieson and G. Wall. *Tourism: economic, physical and social impacts*. Longman, Harlow, United Kingdom, 1982.
- [4] R. W. McIntosh and C. R. Goeldner. *Tourism: principles, practices, philosophies*. Wiley, NY, USA, 1986.
- [5] UNWTO. Unwto, world tourism highlights 2015 edition. In <http://www.e-unwto.org/doi/pdf/10.18111/9789284416899>. [Online] accessed May 2016, 2016.
- [6] UNWTO. World tourism organization (unwto) annual report 2014. In [http://cf.cdn.unwto.org/sites/all/files/pdf/unwto\\_annual\\_report\\_2014.pdf](http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_annual_report_2014.pdf). [Online] accessed May 2016, 2015.
- [7] The World Bank. International tourism, number of arrivals. In <http://data.worldbank.org/indicator/ST.INT.ARVL>. [Online] accessed May 2016, 2015.
- [8] S. L. Tilahun and H. C. Ong. Prey predator algorithm: A new metaheuristic optimization algorithm. *International Journal of Information technology & Decision Making*, 14:1331 – 1352, 2015.
- [9] S. L. Tilahun. Prey predator algorithm: A new metaheuristic optimization approach. In *PhD thesis, School of Mathematical Sciences, Universiti Sains Malaysia, Malaysia*. USM, 2013.
- [10] S. Kumbharana and G. Pandey. Solving travelling salesman problem using firefly algorithm. *International Journal for research in Science and Advanced Technologies*, 2:53 – 57, 2013.
- [11] M. Garey and J. David. *Computers and Intractability: A guide to the Theory of NP-Completeness*. Bell Laboratories Murray Hill, New Jersey, USA, 1979.
- [12] C. H. Papadimitriou. The euclidean travelling salesman problem is np-complete. *Theoretical Computer Science*, 4:237 – 244, 1977.
- [13] R. Harp. Reducibility among combinatorial problems. In *University of California at Berekly*. 1972.
- [14] W. Cook. The travel salesman problem. In <http://www.tsp.gatech.edu/index.html>. [Online] accessed May 2016, 2010.
- [15] A. Maredia. History, analysis, and implementation of traveling salesman problem (tsp) and related problems. In *Senior Project*

- in University of Houston-Downtown. [Online] accessed May 2016, <http://cms.uhd.edu/faculty/redlt/annemseniorproject.pdf>, 2010.
- [16] H. Calgor. The bruth force algorithm. In <http://www.ctl.ua.edu/math103/hamilton/hamilton.html>. [Online] accessed May 2016, 2010.
  - [17] A. Land and A. Doig. An outomotive method of solving discrete programming problems. *Econometrics*, 28:409 – 520, 1960.
  - [18] N. Hamadneh, S. L. Tilahun, S. Sathasivam, and H. C. Ong. Prey-predator algrorithm as a new optimization technique using in radia basis function neural networks. *Research Journal of Applied Sciences*, 8:383 – 387, 2013.
  - [19] S. L. Tilahun and S. Melese. Prey predator with adaptive step length. In *the 8th International Congress of Industrial and Applied Mathematics, 9 - 13 Aug. 2015, Beijing, China*. 2015.
  - [20] S. L. Tilahun. Fuzzy graph representation of bus timetabling problem and its solution method using prey-predator algorithm. In *International Workshop on Optimal Network Topologies (IWONT) 2012, 27 - 29 July 2012, Institut Teknologi Bandung, Bandung, Indonesia*. 2012.
  - [21] S. L. Tilahun and J. M. T. Ngnotchouye. Prey predator algorithm with adaptive step length. *International Journal of Bio-Inspired Computing [in press]*.
  - [22] S. L. Tilahun, H. C. Ong, and J. M. T. Ngnotchouye. Extended prey-predator algorithm with a group hunting scenario. *Advances in Operations Research*, vol. 2016:14 pages, 2016.
  - [23] Wikipedia. World tourism rankings. In [https://en.wikipedia.org/wiki/World\\_Tourism\\_rankings](https://en.wikipedia.org/wiki/World_Tourism_rankings). [Online], accessed on May 2016, 2016.
  - [24] Free Map Tools. How far is it between. In <https://www.freemaptools.com/how-far-is-it-between.htm>. [Online], accessed on May 2016, 2016.

## Extracting and Characterizing Functional Communities in Spatial Networks

Takayasu Fushimi<sup>1</sup>, Kazumi Saito<sup>2</sup>, Tetsuo Ikeda<sup>2</sup>, and Kazuhiro Kazama<sup>3</sup>

<sup>1</sup> University of Tsukuba, Japan, [takayasu.fushimi@gmail.com](mailto:takayasu.fushimi@gmail.com)

<sup>2</sup> University of Shizuoka, Japan, [{k-saito,t-ikeda}@u-shizuoka-ken.ac.jp](mailto:{k-saito,t-ikeda}@u-shizuoka-ken.ac.jp)

<sup>3</sup> Wakayama University, Japan, [kazama@ingrid.org](mailto:kazama@ingrid.org)

**Abstract.** We address the problem of extracting and characterizing functional communities consisting of functional similar regions in spatial networks such as urban streets. Such characteristics of regions will play important roles for developing and planning city promotion, travel tours and so on, as well as understanding and improving the usage of urban streets. In order to analyze such functionally similar regions, based on a previous algorithm for extracting functional communities for each network, we propose a new method consisting of a technique for simultaneously comparing these functional communities of several networks, and an effective way of visualizing these communities calculated from OpenStreetMap data, by especially focusing on a fact that the maximum degree of nodes in spatial networks is restricted to relatively small numbers. In our experiments using urban streets of six cities, we show that our method can produce a series of useful visualization results accompanied with interpretable functional communities. Moreover, we empirically confirm that our results are substantially different from those obtained by representative centrality measures.

### 1 Introduction

Studies of the structure and functions of large complex networks have attracted a great deal of attention in many different fields such as sociology, biology, physics and computer science (Newman, 2003). As a particular class of them, we focus on spatial networks embedded in the real space, like urban streets, whose nodes occupy a precise position in two or three-dimensional Euclidean space, and whose links are real physical connections (Crucitti et al., 2006). In this paper, we address the problem of extracting and characterizing functional communities consisting of functionally similar regions in spatial networks such as urban streets. Such characteristics of regions will play important roles for developing and planning city promotion, travel tours and so on, as well as understanding and improving the usages of urban streets.

In this paper, we mainly consider spatial networks constructed by mapping the ends and intersections of streets into nodes and the streets between the nodes into links. For these networks, we consider the node functionality defined through a probability vector obtained from a random walk process (Fushimi et al., 2012).

Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama

We believe that it can be a quite important research topic to extract groups of functionally similar nodes in a spatial network, which are referred to as functional communities. Examples of such functional communities might include parts of streets constructed in planned cities like lattices, those reflected by geographical restrictions like cul-de-sacs, and so on.

In order to analyze such functionally similar regions, based on an existing algorithm for extracting functional communities for a single network, we propose a new method consisting of a technique for simultaneously comparing these functional communities of several networks, and an effective way of visualizing these communities. More specifically, by using the degree and community distributions, and we first calculate each Z-score with respect to the degree and community, and then characterize each functional community in terms of these Z-scores by especially focusing on a fact that the maximum degree of nodes in spatial networks is restricted to relatively small numbers. As for our visualization way, based on the actual location of each node mapped from intersections of streets, we plot these nodes by assigning an individual color to each community and consistently using this color scheme for several networks.

In our experiments using several types of urban streets of six cities, we evaluate the extracted functional communities by our proposed method and discuss the characteristics of these results in comparison to those obtained by representative centrality measures. This paper is organized as follows: after explaining related work in Section 2, we describe a detail of our proposed method in Section 3. Then, in Section 4, we qualitatively and quantitatively evaluate the characteristics of the extracted functional communities, and describe our conclusion in Section 5.

## 2 Related Works

As mentioned earlier, the structure and functions of large spatial networks have been studied by many researchers (Burckhart and Martin, 2012, Crucitti et al., 2006, Montis et al., 2007, Opsahl et al., 2010, Park and Yilmaz, 2010, Wang et al., 2012). From structural viewpoints, centrality measures have been widely used to analyze this type of networks (Crucitti et al., 2006, Park and Yilmaz, 2010), especially by extending the conventional notions of centrality measures on simple networks into those of weighted networks (Montis et al., 2007, Opsahl et al., 2010). From functional viewpoints, traffic usage patterns in urban streets have been investigated (Burckhart and Martin, 2012, Wang et al., 2012). Unlike these previous studies, in this paper, we attempt to extract functional communities as intrinsic properties of these spatial networks. Here, also note that our study naturally combines structural and functional viewpoints in terms of functional communities.

Studies of community extraction can be another main stream of complex network analysis. As mentioned above, we employ a method of extracting functional communities (Fushimi et al., 2012). This is because representative methods for extracting communities as densely connected subnetworks, which include the

### Extracting and Characterizing Functional Communities in Spatial Networks

Newman clustering method based on a modularity measure (Newman, 2004), cannot directly deal with such functional properties. Also, conventional notions of densely connected subnetworks such as  $k$ -core (Seidman, 1983) and  $k$ -clique (Palla et al., 2005) cannot work for this purpose. Namely, it is naturally anticipated that these representative methods have an intrinsic limitation for extracting functionally similar nodes. Moreover, it might be difficult to straightforwardly apply these conventional methods to spatial networks, because the maximum degree of nodes in each network is generally restricted to a relatively small number, i.e., densely connected subnetworks are unlikely to appear in these networks.

### 3 Proposed Method

As an essential component of our proposed method, we first revisit the existing algorithm for extracting functional communities (Fushimi et al., 2012), which consists of two steps, calculation of functional vectors and clustering of these vectors. Let  $G = (V, E)$  be a network constructed from urban streets, and for each node  $u \in V$ , we denote the set of its adjacent nodes by  $\Gamma(u)$ . Let  $y_s(u)$  be a visiting probability of node  $u$  at step  $s$ , which satisfies  $y_s(u) > 0$  and  $\sum_{u \in V} y_s(u) = 1$ ; then, after initializing the  $|V|$ -dimensional probability at step  $s = 0$  to  $\mathbf{y}_0 = (1/|V|, \dots, 1/|V|)^T$ , we can consider the following random walk process, just like the PageRank algorithm (Langville and Meyer, 2004),

$$y_s(u) = \sum_{v \in \Gamma(u)} \frac{y_{s-1}(v)}{|\Gamma(v)|}, \quad (1)$$

where  $|V|$  means the number of  $V$ 's elements. Let  $S$  be the final step of the random walk process, which is set to a reasonably large number, say  $S = 1,000$  in our experiments; then, for each node  $u \in V$ , we can calculate an  $S$ -dimensional vector  $\mathbf{x}_u = (y_1(u), \dots, y_S(u))^T$ , which is referred to as the functional vector of node  $u$ , where  $\mathbf{a}^T$  means a transposed vector of  $\mathbf{a}$ .

We divide all nodes into the  $K$  groups of functional communities by employing the  $K$ -medoids algorithm (Vinod, 1969) due to its robustness. More specifically, by using a cosine similarity function between each pair of the functional vectors  $\mathbf{x}_u$  and  $\mathbf{x}_v$ , defined by  $\rho(u, v) = \mathbf{x}_u^T \mathbf{x}_v / (\|\mathbf{x}_u\| \|\mathbf{x}_v\|)$ , we maximize the following objective function with respect to  $\mathcal{R} \subset V$ :

$$f(\mathcal{R}) = \sum_{v \in V} \max_{r \in \mathcal{R}} \rho(v, r), \quad (2)$$

where  $\|\mathbf{x}_v\|$  means the standard L2 norm and  $|\mathcal{R}| = K$ . In order to maximize the objective function  $f(\mathcal{R})$ , we employ a greedy algorithm, i.e., after initializing  $k \leftarrow 1$  and  $\mathcal{R} \leftarrow \emptyset$ , we select  $r_k = \arg \max_{w \in V \setminus \mathcal{R}} \{f(\mathcal{R} \cup \{w\})\}$ , and set to  $\mathcal{R} \leftarrow \mathcal{R} \cup \{r_k\}$  and  $k \leftarrow k + 1$  during  $k \leq K$ . From the obtained  $K$  representative objects  $\mathcal{R} = \{r_1, \dots, r_K\}$ , we can calculate each functional community as  $V^{(k)} = \{v \in V; r_k = \arg \max_{r \in \mathcal{R}} \{\rho(v, r)\}\}$ . Here note that in virtue of the submodularity

Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama

of the objective function, it is guaranteed that we can obtain a unique greedy solution with reasonably high quality (Nemhauser et al., 1978).

Now, based on the above algorithms for extracting functional communities, we propose a new method which incorporates a technique for simultaneously comparing these functional communities of several networks, and an effective way of visualizing these communities. Let  $V_j = \{u \in V ; |\Gamma(u)| = j\}$  and  $V_j^{(k)} = \{u \in V^{(k)} ; |\Gamma(u)| = j\}$  be the sets of nodes with degree  $j$  and those belonging to the functional community  $k$ , respectively. By defining the degree and community distributions by  $p_j = |V_j|/|V|$  and  $p^{(k)} = |V^{(k)}|/|V|$ , respectively, we can calculate the following Z-score  $Z_j^{(k)}$  with respect to the degree  $j$  and community  $k$ :

$$Z_j^{(k)} = \frac{|V_j^{(k)}| - |V|^2 \times p^{(k)} \times p_j}{\sqrt{|V|^2 \times p^{(k)} \times p_j \times (1 - p^{(k)} \times p_j)}}. \quad (3)$$

Evidently, when  $Z_j^{(k)}$  is large (or small), we can conjecture that a significantly large (or small) number of nodes with degree  $j$  appear in the community  $k$ . In our proposed method, we use these Z-scores as a measure in order to characterize each functional community. Recall that the maximum degree of nodes in spatial networks is restricted to relatively a small number.

As for our visualization way, based on the actual location of each node mapped from intersections of streets, we plot these nodes by assigning an individual color to each community. In our experiments, we set the number of communities to  $K = 5$ , and assign each color of red, lime, blue, yellow and magenta to  $V^{(1)}, \dots, V^{(K)}$  in this order. Here, we consistently use the same color scheme for different networks in order to contrast differences of each network.

## 4 Experiments

We obtained OSM (OpenStreetMap) data of six cities, i.e., Barcelona, Brasilia, Cairo, Washington D.C., New Delhi and Richmond, which were obtained from Metro Extracts<sup>2</sup> in August, 2015. These cities were selected as a subset of those studied in (Crucitti et al., 2006), but note that in our experiments, each area of these cities is more than 100 times larger than 1-square mile area used in the previous study. Then we extracted all highways and all nodes from the OSM data of each city, and each spatial network was constructed by mapping the ends and intersections of streets into nodes and the streets between nodes into links. Here, note that in order to simplify our analyses, we deleted nodes used for representing curve-segments of highways by directly connecting both sides of deleted ones. Table 1 shows the basic statistics of the networks for six cities, where  $C$  and  $L$  denote the averages of clustering coefficients and shortest path lengths over each network, respectively. We can see that although the areas and the numbers of nodes and links,  $|V|$  and  $|E|$  are substantially different, degree

<sup>2</sup> <https://mapzen.com/data/metro-extracts>

## Extracting and Characterizing Functional Communities in Spatial Networks

**Table 1.** Basic statistics as network.

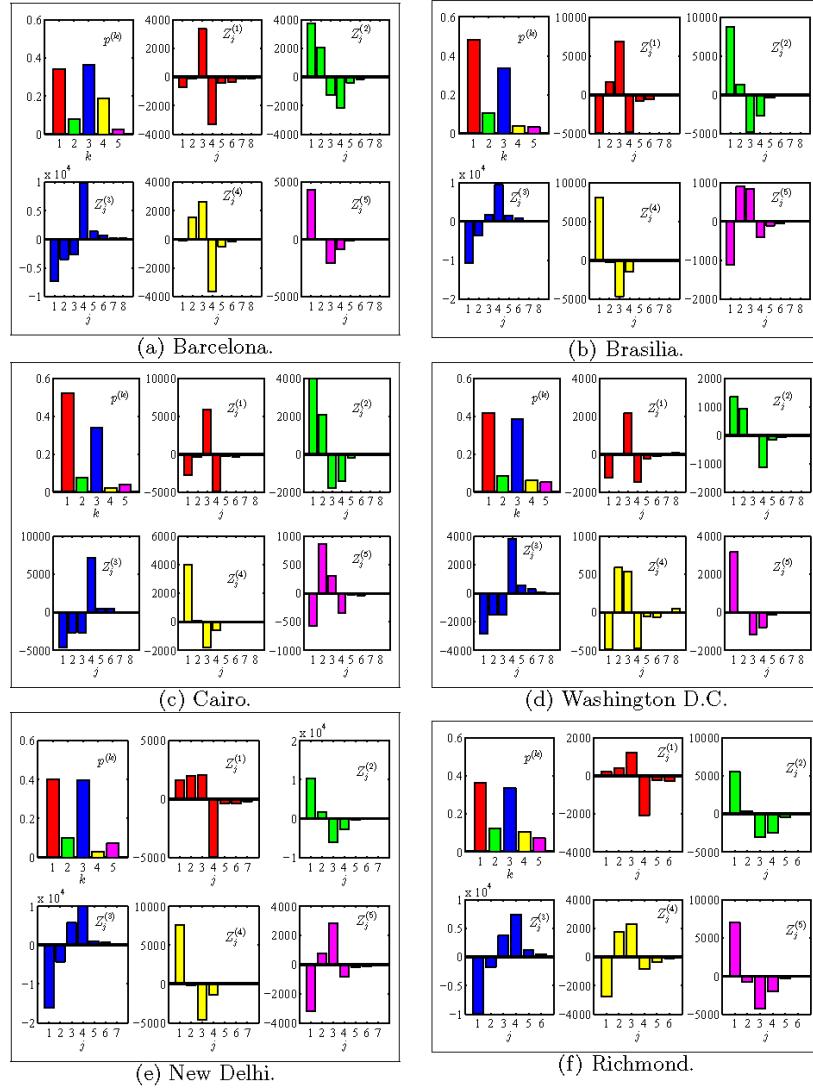
city	area	$ V $	$ E $	$p_1$	$p_2$	$p_3$	$p_4$	$p_{>4}$	$C$	$L$
Barcelona	45km × 30km	66,790	198,774	.103	.031	.659	.207	.006	0.06	53.07
Brasilia	120km × 104km	95,811	273,910	.133	.025	.694	.146	.002	0.04	92.94
Cairo	87km × 86km	56,781	171,188	.068	.025	.733	.172	.002	0.04	58.80
Wash. D.C.	23km × 18km	24,564	76,106	.096	.028	.571	.293	.012	0.05	51.89
New delhi	109km × 75km	116,905	333,486	.138	.017	.702	.142	.002	0.03	78.04
Richmond	54km × 35km	66,739	175,140	.249	.045	.543	.160	.003	0.04	65.47

distributions defined by  $p_j$ , as well as  $C$  and  $L$ , are quite similar as common characteristics of these spatial networks.

Figure 1 shows the community distribution  $p^{(k)}$  together with the obtained Z-score distribution  $Z_j^{(k)}$  for each of six cities by setting  $K = 5$ , where we assigned each color of red, lime, blue, yellow and magenta to  $V^{(1)}, \dots, V^{(5)}$  in this order, just like shown later in our visualization results. From the community distribution results, we can see that the union of the 1st and 3rd functional communities occupied more than around 80% of total nodes. From the Z-score distribution results, we can observe that for all of the six cities, the 1st, 2nd and 3rd functional communities ( $V^{(1)}$ ,  $V^{(2)}$  and  $V^{(3)}$ ) have the commonly similar characteristics, i.e., significantly larger Z-scores for nodes with degree  $j = 3$ ,  $j = 1$  and  $j = 4$ , respectively. In what follows, we refer to these community regions simply as 3-intersection, cul-de-sac and lattice. On the other hand, we can observe that the 4th and 5th functional communities ( $V^{(4)}$  and  $V^{(5)}$ ) do not have the commonly similar characteristics. We consider that this differences may be caused by individual characteristics of these cities, which are reflected by geographical restrictions, and/or historical and cultural backgrounds. These characteristics of the extracted communities can also be naturally explained by the nature of the greedy algorithm employed in our proposed method. Namely, this algorithm generally selects the first community having with some average characteristics like 3-intersection regions, and then the successive communities having some salient characteristics like cul-de-sac and lattice regions. Thus, we can conjecture that the former three communities ( $V(1)$ ,  $V(2)$  and  $V(3)$ ) reflected the common characteristics for these cities, while the latter two communities ( $V(4)$  and  $V(5)$ ) reflected the individual characteristics of each city.

Figure 2 shows our visualization results for six cities, where recall that each color of red, lime, blue, yellow and magenta is consistently used in our proposed method. From these results, we can observe that all of six cities have the commonly similar characteristics, i.e., the red 3-intersection regions ( $V^{(1)}$ ), which were surrounded by lime cul-de-sac regions ( $V^{(2)}$ ), surround blue lattice regions ( $V^{(3)}$ ). This observation must be naturally interpretable from the aspects of geographical restrictions, and suggests the practical usefulness of our proposed method. Moreover, as another advantage of our visualization results, we can intuitively understand the detailed regions of each city in terms of characteristics

Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama



**Fig. 1.** Community and Z-score distributions for six cities.

### Extracting and Characterizing Functional Communities in Spatial Networks

of interpretable functional communities. In addition, through our experiments by changing the focused area size of Barcelona to smaller one ( $30\text{km} \times 15\text{km}$ ) and larger one ( $69\text{km} \times 47\text{km}$ ), we have confirmed that our method could robustly produce some results with almost consistent characteristics regardless of the area sizes, although these results are not shown in this paper due to a space limitation. Therefore, our proposed method is expected to work as a useful tool for developing and planning city promotion, travel tours and so on, as well as understanding and improving the usage of urban streets.

By focusing on representative three centrality measures, Bonacich, closeness and betweenness, we characterize our method in comparison to these centrality measures. Note that, for a given network, Bonacich centrality gives a rank to each node by the principal component of an adjacency matrix, closeness centrality by the inverse of the sum of shortest path lengths, and betweenness centrality by the passing rate over shortest paths between any pair of two nodes (Wasserman and Faust, 1994). Figures. 3, 4 and 5 show our experimental results using these centrality measures, where we plotted each node with a gradation color from red to blue according to the rank by each centrality measure. From these figures, as common characteristics for all of six cities, we can see that Bonacich centrality gave high ranks to some regions (faces) of relatively high degree nodes typically in city centers, closeness centrality to some streets (lines) of continuously adjacent nodes typically on arterial roads, and betweenness centrality to some points of isolated nodes scattered widely all over the network. Here we should note that the city center of Brasilia consists of relatively low degree nodes. In contrast, our method can more detailedly characterize these regions in terms of interpretable functional communities, where only the 3rd functional community  $V^{(3)}$  might roughly coincide with the highly ranked regions by Bonacich centrality. Namely, we could empirically confirm that our results were substantially different from those obtained by representative centrality measures.

## 5 Conclusion

We addressed the problem of extracting and characterizing functional communities consisting of functionally similar regions in spatial networks such as urban streets. To this end, we proposed a method consisting of a technique for characterizing these functional communities and an effective way of visualizing these communities, based on an existing algorithm for extracting functional communities. In our experiments using urban streets of six cities, we showed that our method could produce a series of useful visualization results accompanied with interpretable functional communities. Moreover, we empirically confirmed that our results are substantially different from those obtained by representative centrality measures. These promising results may suggest that we have taken some important steps to tackle the interpretation problem of extracted communities (or clustering results), one of fundamental problems in data mining and machine learning researches. In future, we plan to evaluate our method using various

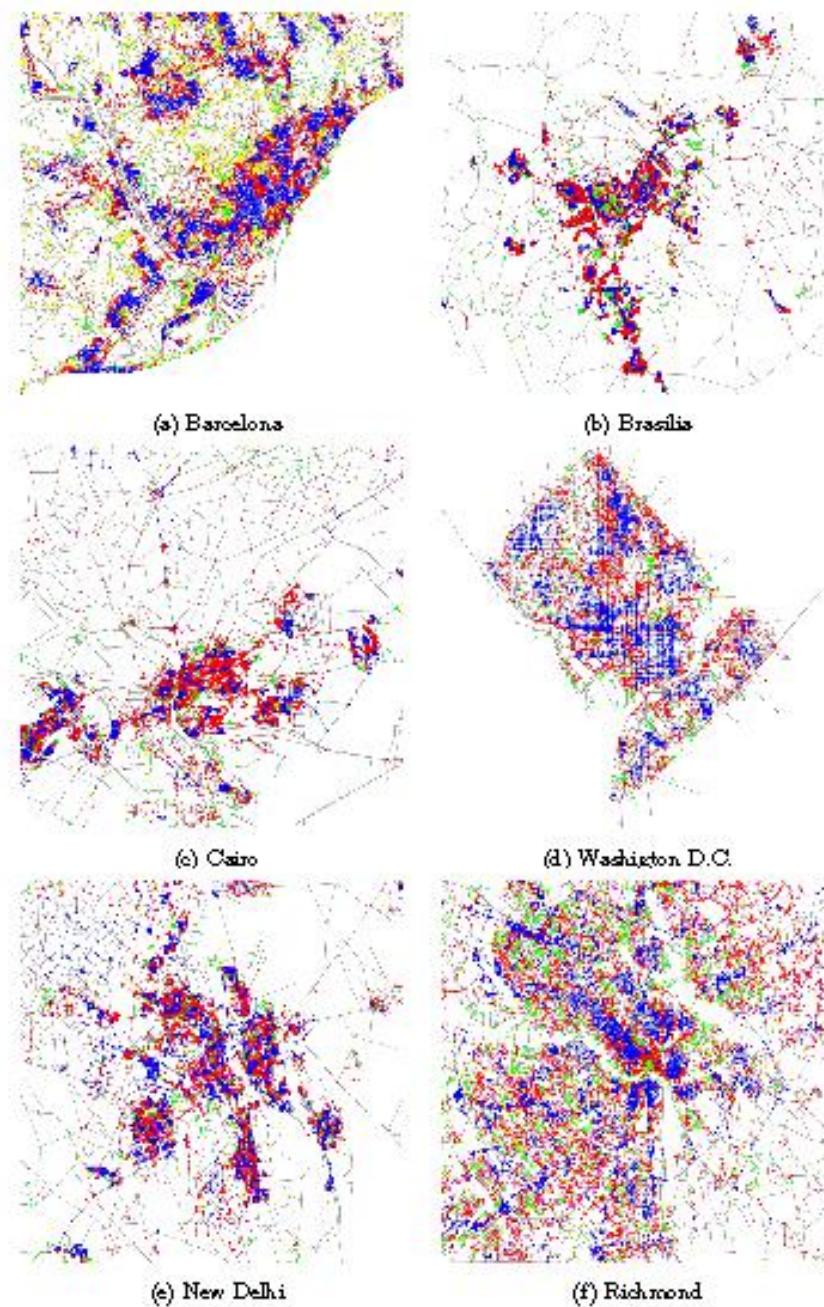
Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama

spatial networks, and attempt to establish more useful tools for developing and planning city promotion, travel tours and so on. In addition, we need to detail extracted functionally similar regions, especially ,  $V(4)$  and  $V(5)$  which reflect the individual character of each city. Furthermore, in order to deal with large scale spatial networks, we plan to develop some techniques for improving the scalability of the  $K$ -medoids algorithm used in our proposed method.

## References

- Burckhart, K. and Martin, O. J. (2012). An Interpretation of the Recent Evolution of the City of Barcelona through the Traffic Maps. *Journal of Geographic Information System*, 4(4):298–311.
- Crucitti, P., Latora, V., and Porta, S. (2006). Centrality Measures in Spatial Networks of Urban Streets. *Physical Review E*, 73(3):036125+.
- Fushimi, T., Saito, K., and Kazama, K. (2012). Extracting Communities in Networks based on Functional Properties of Nodes. In Richards, D. and Kang, B. H., editors, *Proceedings of the 12th Pacific Rim Knowledge Acquisition Workshop (PKAW2012)*, pages 328–334, Berlin, Heidelberg. Springer-Verlag.
- Langville, A. N. and Meyer, C. D. (2004). Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–380.
- Montis, D. A., Barthelemy, M., Chessa, A., and Vespignani, A. (2007). The Structure of Interurban Traffic: A Weighted Network Analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924.
- Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An Analysis of Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, 14:265–294.
- Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256.
- Newman, M. E. J. (2004). Detecting Community Structure in Networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330.
- Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks*, 32(3):245–251.
- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*, 435:814–818.
- Park, K. and Yilmaz, A. (2010). A Social Network Analysis Approach to Analyze Road Networks. In *Proceedings of the ASPRS Annual Conference 2010*.
- Seidman, S. B. (1983). Network Structure and Minimum Degree. *Social Networks*, 5(3):269 – 287.
- Vinod, H. (1969). *Integer Programming and The Theory of Grouping*, volume 64. An Official Journal of the American Statistical Association.
- Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., and Gonzalez, M. C. (2012). Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, 2(arXiv:1212.5327):1001. 47 p.
- Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

## Extracting and Characterizing Functional Communities in Spatial Networks

Fig. 2. Visualization results by proposed method ( $K = 5$ ).

Takayasu Fujimi, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama

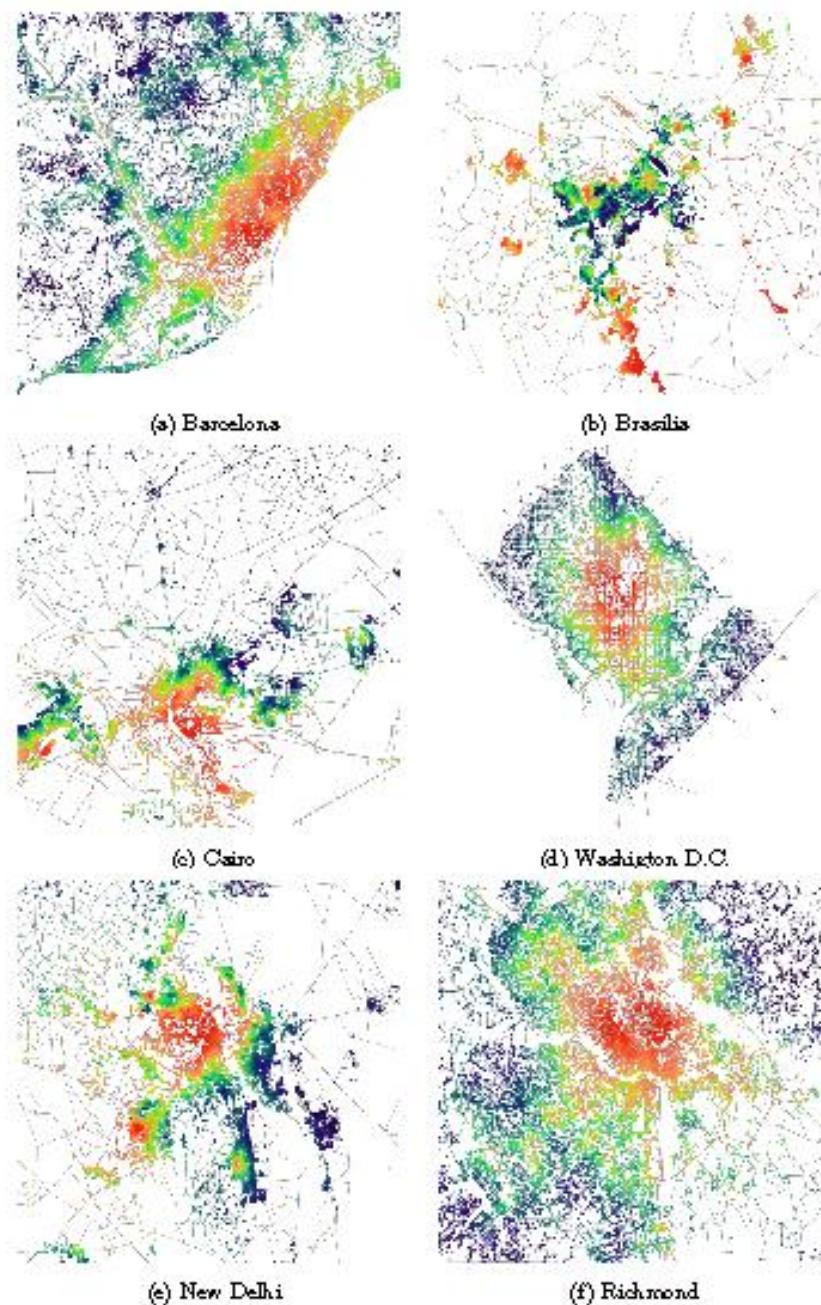


Fig. 3. Visualization results by Bonacich centrality.

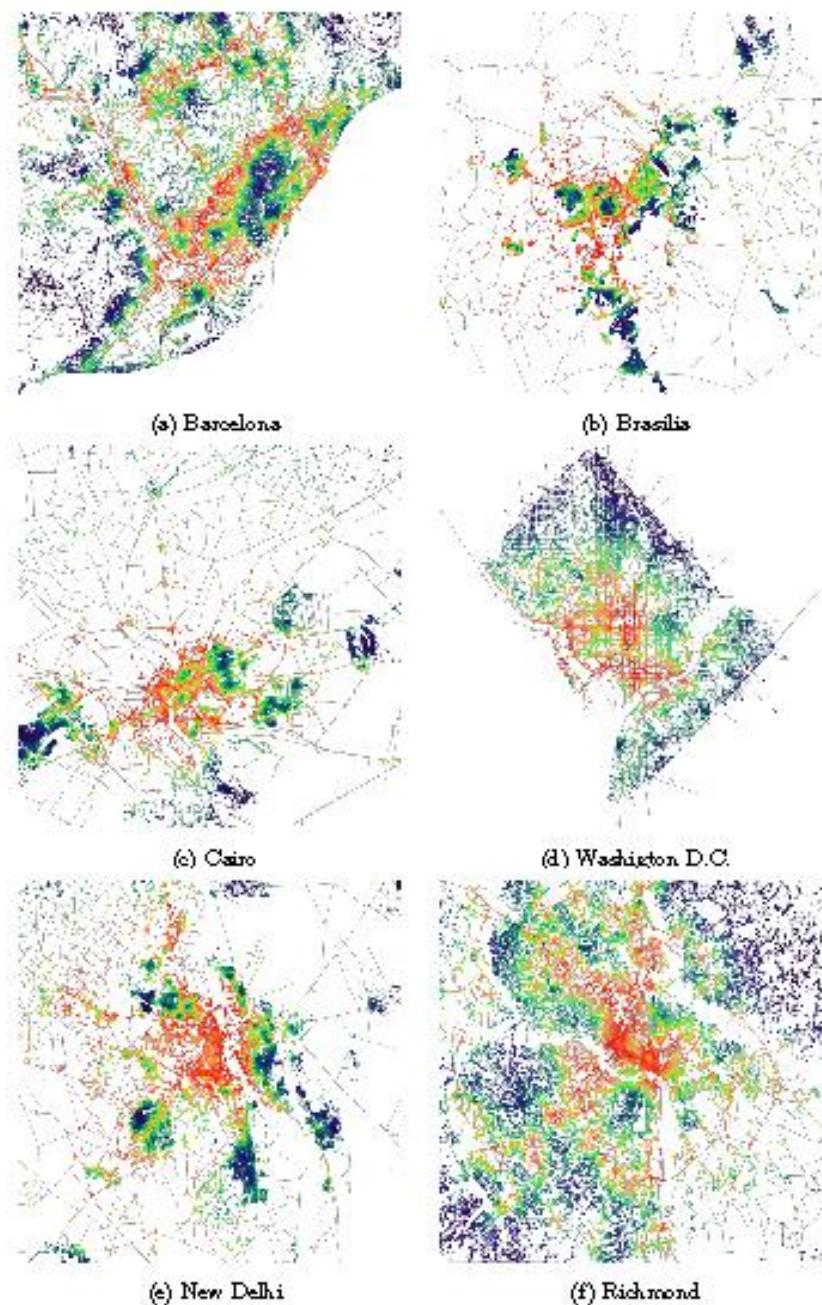
*Extracting and Characterizing Functional Communities in Spatial Networks*

Fig. 4. Visualization results by closeness centrality.

Takayasu Fujinami, Kazumi Saito, Tetsuo Ikeda, and Kazuhiro Kazama

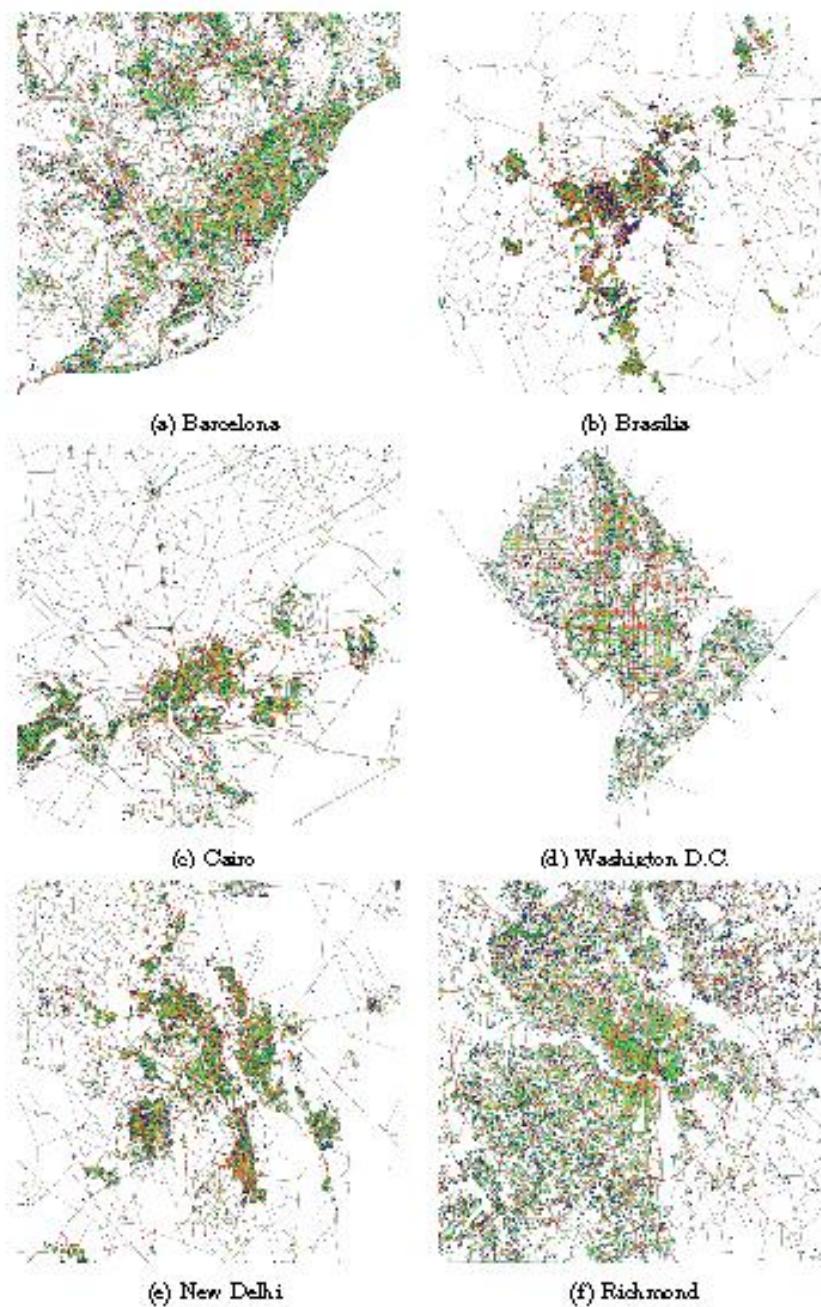


Fig. 5. Visualization results by betweenness centrality.

## Travellers' Behaviour Analysis Based on Automatically Identified Attributes from Travel Blog Entries

Kazuki Fujii<sup>1</sup>, Hidetsugu Nanba<sup>1</sup>, Toshiyuki Takezawa<sup>1</sup>, Aya Ishino<sup>2</sup>,  
Manabu Okumura<sup>3</sup> and Yohei Kurata<sup>4</sup>

<sup>1</sup> Hiroshima City University, 3-4-1 Ozuka-higashi, Asaminami-ku,  
Hiroshima, 731-3194, Japan

{fujii, nanba Travellers' Behaviour Analysis Based on Automatically Identified At-  
tributes from Travel Blog Entries, takezawa}@ls.info.hiroshima-cu.ac.jp

<sup>2</sup> Hiroshima University of Economics, 5-37-1 Gion, Asaminami-ku,  
Hiroshima 731-0192, Japan  
ay-ishino@hue.ac.jp

<sup>3</sup> Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku,  
Yokohama, 226-8503, Japan  
oku@pi.titech.ac.jp

<sup>4</sup> Tokyo Metropolitan University, 1-1 Minamiosawa, Hachioji, Tokyo, 192-0397, Japan  
ykurata@tmu.ac.jp

**Abstract.** We propose a method to analyse travellers' behaviour using automatically identified travellers' attributes, such as gender and language, from travel blog entries. We consider that travel blog entries are a useful information source for obtaining travel information, because many bloggers' describe their travel experiences in them. Several studies have analysed travellers' behaviour using travel blog entries. However, they used a small number of manually identified travellers' (bloggers') attributes. In our work, we identify travellers' attributes automatically using natural language processing techniques, and conduct a large-scale travellers' behaviour analysis.

**Keywords:** travel blog, behaviour analysis, travellers' attributes,

### 1 Introduction

Being aware of travellers' needs is crucial for tourism planning. Traditionally, such analyses were conducted using questionnaires, but these are costly and time-consuming. Recently, travel blog entries have been used instead of questionnaires. In travel blog entries, various travellers' experiences and opinions are described, and they can help identify travellers' needs. For example, Wenger [1] analysed travel blog entries written by travellers visiting Austria, and found that many females visited Austria to enjoy dining experiences. However, few blog entries were used in this analysis, because the analysis was conducted manually. In this paper, we propose a method for analysing travellers' behaviour from vast numbers of blog entries using natural language techniques. In our approach, we identify travellers' (bloggers') at-

tributes automatically using method based on machine learning. Then, we use the attributes to characterize travellers' behaviour.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our method of identifying travellers' (bloggers') attributes in travel blog entries. To investigate the effectiveness of our method, we conducted some experiments, and Section 4 reports the experimental results. Using our method, we conducted travellers' behaviour analysis, and Section 5 reports the results. We present some conclusions in Section 6.

## 2 Related Work

Questionnaires have been frequently used in marketing surveys to determine tourist policies. Xia *et al.* [2] asked 464 visitors to Phillip Island about their gender, tourist spots they planned to visit, and their residence, and conducted behaviour analysis using decision-tree methods. Similarly, Jonsson *et al.* [3] also conducted a questionnaire investigation of 163 visitors to a Caribbean island, and analysed travellers' behaviour. However, it is costly and quite time-consuming to conduct questionnaire investigations. Therefore, we aim to analyse travellers' behaviour automatically from travel blog entries, because there are many descriptions of travellers' experiences and many opinions in travel blog entries.

The number of studies using travel blog entries for travellers' behaviour analysis has been increasing. Mack *et al.* [4] and Akehurst [5] reported that travel blog entries are useful for tourism marketing, although they are not as reliable as traditional word of mouth. Li and Wang [6] analysed impressions of China from Taiwanese travellers using travel blog entries in a Chinese travel portal site. They manually classified blog entries to several categories, such as "scenery", "purchase", and "stay", and analysed entries in each category. They reported that the impressions of hot springs were good, and suggested that these hot springs should be promoted as a major tourist attraction in China. Classifying travel blog entries for travellers' behaviour analysis is a common point with our work. However, our work classifies entries automatically, which enables us to conduct a larger-scale analysis.

In related research using travel blog entries, Wenger [1] conducted travellers' behaviour analysis based on travellers' attributes, such as gender and age. They used the dataset of TravelBlog<sup>1</sup>, in which some bloggers reveal their attributes in their profiles. However, few bloggers explicitly describe these attributes. We therefore identify bloggers' attributes automatically, and use them for travellers' behaviour analysis.

Recently, several methods for identifying bloggers' attributes, such as gender, age [7, 9], and residential area [8] have been proposed. We identify travellers' gender based on Ikeda's method. With respect to travel-related documents, Saeki *et al.* [10] identified travellers' languages in each geotagged tweet using a Java language-detection<sup>2</sup> program (`langdetect`), and conducted foreign travellers' behaviour analysis

---

<sup>1</sup> <https://www.travelblog.org/>

<sup>2</sup> <https://code.google.com/p/language-detection/>

in Japan. In the same way, we also apply the program to travel blog entries, and use them for travellers' behaviour analysis.

### 3 Automatic Identification of Travellers' Attributes

We aim to analyse travellers' behaviour based on the travellers' gender, language, and behaviour using the TravelBlog dataset. Unfortunately, this information is not explicitly written in the dataset. Therefore, we identify attributes automatically using natural language processing techniques. We explain how to identify a blogger's (a traveller's) gender, language, and behaviour (content type of each blog entry) in Sections 3.1, 3.2, and 3.3, respectively.

#### 3.1 Identification of Blogger's Gender

To identify bloggers' genders, we propose three methods: (1) semi-supervised learning approach based on Ikeda's method [7] (SSL), (2) cue-word-based method (CUE), and (3) a combination of the CUE and SSL methods.

Ikeda *et al.* (2008) assumed that each blogger has a writing style based on his/her attributes, such as gender and generation. They identified these attributes from two kinds of data: (i) a small number of blog entries in which bloggers explicitly describe their attributes; and (ii) a large number of entries with no explicit attributes using a semi-supervised technique. They experimentally confirmed that their approach obtained an accuracy score of 0.890 for the identification of bloggers' genders, while a supervised approach using the data (i) obtained 0.760. Therefore, we examine the semi-supervised approach based on the Ikeda's method (SSL).

In addition to the SSL method, we also examine a cue-word-based method. This method focuses on the differences in words that male and female bloggers use. We created two cue word lists by collecting words that are frequently used in blog entries written by male or female bloggers, and used them for the identification of bloggers' genders (CUE).

As our third approach, we examine the combination of the SSL and CUE methods. The basic procedure is the same as that for the SSL method, except that the SSL+CUE method does not use all words in blog entries, only those that appear in two cue-word lists. We believed that using the cue-word lists could identify the features of blog entries more clearly, and would improve the SSL method. In the experiment in Section 4, we will report that our SSL+CUE method did actually improve the SSL method.

#### 3.2 Identification of Travellers' Languages

Many entries in the TravelBlog database were written in English, but other languages, such as French and German, are also used. We focus on the languages that bloggers used, and analyse the travellers' behaviours in each city they visited in terms of their languages. This analysis is particularly useful for tourism policy. For example, if we

find that there are many visitors who use French in a particular city, then it is effective for the city authorities to provide signs in French.

To identify which language a blogger uses in a blog entry, we use the langdetect program. This library automatically estimates probabilities of each language for a given text. We attempt to identify travellers' languages in two different ways: (1) identify the language that has the highest probability score (Top method); and (2) identify all languages having probabilities higher than a threshold value (Threshold method). This threshold value was determined in a pilot study.

### 3.3 Identification of Content Type

Even when travellers visit the same destination, the purpose of the visit is not always the same. Some travellers might visit tourist spots as their primary purpose, while others might visit to enjoy local dining. We aim to make the visitors' purposes clear by identifying the content type of each blog entry, as shown in Table 1. These types were originally proposed by Fujii *et al.* [11], and they devised a system that identifies one or more content types relevant to a given blog entry written in Japanese. In their method, they first collected cue words that are useful for identifying content types using the information gain (IG) method. Second, they applied Support Vector Machine (SVM) to identify content types using cue words as features for the SVM. In our work, we developed a system that can identify content type for a given blog entry written in English using two different methods. In the first method, we collect 100 cue words for each content type using IG and apply the SVM, using the same procedure as that used by Ishino *et al.* for Japanese blog entries. In the second method, we employ a machine translation method (MT) by combining the first method with a Japanese identifier devised by Ishino *et al.* In the MT method, we translate an English blog entry into Japanese using the Microsoft Translator API<sup>3</sup>, and then identify its content type using the Japanese identifier (MT). Then, we use the result as one of the features of SVM, and identify the content type of the given English blog entry (IG+MT).

**Table 1. Content types and their descriptions**

Content type	Criterion
Watch	Sightseeing for watching enjoyment
Experience	Experience (scuba diving, dance)
Buy	Shopping or souvenir stores
Dine	Drinking and dining
Stay	Accommodation

---

<sup>3</sup> <https://datamarket.azure.com/dataset/bing/microsofttranslator>

## 4 Experiments

To confirm the effectiveness of our method, we conducted three experiments: (1) identification of blogger's gender; (2) identification of the language used; and (3) identification of the content type of each blog entry. We describe them in Sections 4.1, 4.2, and 4.3, respectively. In the experiments, we used blog entries from the TravelBlog dataset, which we mentioned in Section 3.

### 4.1 Identification of Traveller's Gender

#### 4.1.1 Experimental Settings

##### Data

We used 228 bloggers for this experiment, 77 males and 151 females. We obtained the gender information from each blogger's profile. In this experiment, we used blog entries written in English.

##### Machine Learning and Evaluation Measure

We employed SVM with a linear kernel for machine learning, and conducted two-fold cross-validation. We used accuracy determined by the following equation as an evaluation measure.

$$\text{Accuracy} = \frac{\text{Number of travellers whose genders are correctly identified}}{\text{Number of travellers}}$$

##### Alternatives

We examined the following four methods.

- **Baseline:** Identify all bloggers as female.
- **SSL:** Semi-supervised learning based on Ikeda's method [7].
- **CUE:** Use two different cue word lists as features for machine learning. These lists were created by collecting approximately the top 100 words that appeared in blogs written by males or females.
- **SSL+CUE:** A combination of the SSL method and the CUE method. When we conduct the SSL method, we use words in the above two lists as features for machine learning.

##### Results

We show the experimental results in Table 2. Of the four methods, our method SSL+CUE obtained the best accuracy score.

We expected Ikeda's method to obtain a high accuracy score, because they also identified bloggers' genders, and obtained a score of 0.890. Unfortunately, their method obtained 0.667 with our data, which is much smaller than their experimental result. We consider that this must be because of the different characteristics of the two datasets. Ikeda's experiment used blog entries about various topics, such as sport, politics, automobiles, beauty salons, and sweets, while we only used blog entries

about travel. As a result, Ikeda's method could not capture the gender-related features of blog entries, and did not work well in our dataset. On the other hand, the two cue word lists assisted to capture the gender-related features of blog entries, and as a result, SSL+CUE outperformed SSL.

**Table 2. Evaluation results for the identification of bloggers' genders**

Method	Accuracy
Baseline	0.662 (151/228)
SSL	0.667 (152/228)
CUE	0.776 (177/228)
SSL+CUE	0.877 (195/228)

## 4.2 Identification of Traveller's Language

### 4.2.1 Experimental Settings

#### Data

We manually identified the language that each of 109 bloggers used in their blog entries, and used these for our experiment. The statistics are shown in Table 3. Note that some bloggers used more than one language.

**Table 3. The number of bloggers for each language**

Language	Bloggers	Language	Bloggers
English	83	Portuguese	1
German	10	Swedish	1
Spanish	9	Afrikaans	1
Dutch	9	Hungarian	1
French	6	Finnish	1
Danish	5	Slovene	1
Italian	2	Romanian	1
Japanese	2		

#### Evaluation Measure

We used recall and precision as evaluation measures. Of these, we consider that precision is more important, because a low precision score causes inaccurate analysis, as we will describe in Section 5. In addition, as the number of blog entries has been increasing, the low recall can be improved.

#### Alternatives

For the identification of the languages that each blogger uses, we used the langdetect program, with the following two methods.

- **Top:** languages having the highest probability among automatically identified languages by the langdetect program.
- **Threshold:** languages with probabilities higher than a threshold value, which was determined in a pilot study.

### Results

We show the experimental results in Table 4. The Top method obtained the higher precision score of the two methods. In the analysis in Section 5, we used the Top method.

**Table 4. Evaluation results for the identification of bloggers' languages**

Method	Precision	Recall
Top	<b>0.972</b>	0.797
Threshold	0.887	0.887

### 4.3 Identification of Content Types of Each Blog Entry

#### Data

We examined using 660 travel blog entries, all written in English. For these entries, we assigned content types. A summary of the data is shown in Table 5. Here, more than one content type was assigned to several entries, so the total number of entries in Table 6 is larger than 660.

**Table 5. Summary of the data using in the examination of content type identification**

Content type	Buy	Dine	Experience	Stay	Watch	Other
Entries	30	97	143	61	316	155

#### Machine Learning and Evaluation Measures.

We employed SVM with a linear kernel. We evaluated our methods and a baseline method by recall and precision.

#### Alternatives

We examined the following methods.

- IG: SVM-based approach with cue words, which were identified using information gain method.
- MT: Machine translation-based approach using Microsoft translator and Japanese content-type identifiers [11].
- IG+MT: SVM-based approach using cue words and the MT result as features.
- Baseline: SVM-based approach using all words as features.

### Results

We show the experimental results in Table 6. Our method, IG+MT, outperformed the baseline method by 0.449 points, confirming the effectiveness of our method.

**Table 6. Evaluation results for identification of blog entry type**

Method	Precision	Recall
IG	0.574	0.296
MT	0.458	0.406
IG+MT	<b>0.597</b>	0.336
Baseline	0.148	0.702

## 5 Analysis of Travel Blogs Based on Bloggers' Attributes

We analysed 7,490 blog entries focusing on bloggers' attributes and the content types of each blog entry. These blog entries were written by 1,302 bloggers who had visited Japan.

### 5.1 Basic Statistics of Visitors to Japan Based on the Automatically Identified Attributes

First, we show some basic statistics of foreign visitors to Japan based on the attributes of visitors and content types of blog entries.

#### Gender

We show the number of bloggers for each gender in Table 7.

**Table 7. The number of automatically identified bloggers for each gender**

Gender	Travellers
Male	513
Female	789

#### Language

We show the number of bloggers for each language in Table 8. We removed the English-speaking travellers from the results, because, the number of English-speaking travellers in TravelBlog is much larger than others, as we mentioned previously. From the results in Table 8, French and German are most often used after English.

**Table 8. The number of automatically identified travellers for each language**

Language	Travellers
French	22
German	13
Dutch	10
Spanish	7
Finnish	7

### Content Type

We analysed the content types of the 7,490 blog entries automatically. We show the results in Table 9. The primary purpose of visitors to Japan is watching. On the other hand, only one blog entry had the content type are “buy”.

**Table 9. Automatically identified content types of blog entries**

Content type	Entries
Buy	1
Dine	1134
Experience	315
Stay	319
Watch	3213

## 5.2 Analysis Based on Attributes of Travellers and Content Types

Using automatically identified travellers’ attributes, we conducted travellers’ behaviour analysis. In this report, we mainly focused on Japan as a test case, but it is possible to conduct the same analysis in any city in any country, because we have already identified all travellers’ attributes.

### 5.2.1 Analysis Based on Travellers’ Languages and Content Types of Travel Blog Entries

We analysed travellers’ languages in each prefecture. The top three prefectures with the largest numbers of languages used by travellers are shown in Table 10.

**Table 10. Top three prefectures that the number of languages**

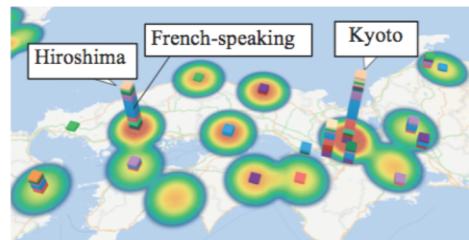
Prefecture	Visitors	Languages
Tokyo	68	18
Kyoto	31	13
Hiroshima	21	10

Among these three prefectures, in the following, we compare Kyoto and Hiroshima in terms of travellers’ languages and content types of travel blog entries. Fig. 1 shows the proportion of blog entries with content type “watch” as a heat map. Red cities indicate that they have more “watch” entries. In this figure, we also show travellers’ languages as bar charts<sup>4</sup>. From the results in Fig. 1, we found that the proportion of “watch” entries was higher in Hiroshima and Kyoto than in the other prefectures.

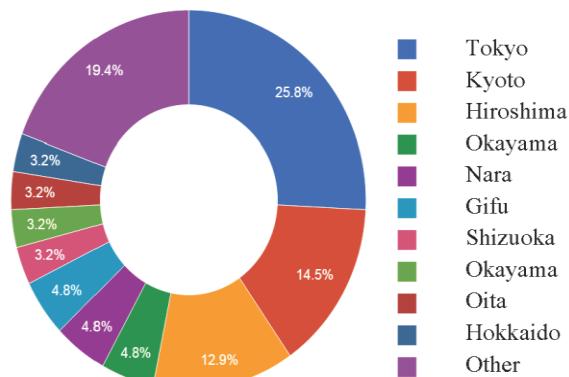
We read the “watch” entries in Kyoto and Hiroshima, and found that many French-speaking travellers visited world heritage locations in both prefectures. From the result, we further investigated the ratio of French-speaking travellers for each prefecture. We show the results in Fig. 2.

---

<sup>4</sup> When we created this chart, we eliminated English-speaking travellers, because most of the travellers (bloggers) in the TravelBlog dataset used English.



**Fig. 1.** The proportion of “watch” travel blog entries and travellers’ languages for each prefecture



**Fig. 2.** The proportion of French-speaking travellers among all

From the results in Fig. 2, we found that approximately 30% of French-speaking travellers visited either Kyoto or Hiroshima. We also found that French-speaking travellers tend to visit world heritage locations in other prefectures. From these results, we can conclude that promotion of world heritage locations and providing signs in French are more important in these areas.

### 5.2.2 Analysis Based on Travellers’ Gender and Content Types of Travel Blog Entries

To compare the different travel purposes between males and females, we calculated the proportions of each content type for each gender in Japan. We show the results in Fig. 3 (a). The results indicate that there are small differences between males and females. However, when we focus on particular areas, we found obvious differences between genders. Fig. 3 (b) shows the proportions of each content type for each gender in Ehime prefecture. The figure shows that the secondary purpose of males who visit Ehime is to dine, while that of females is to experience. Ehime prefecture is famous for its hot springs, and we found that many female visitors enjoyed them. From

these results, we can conclude that the purposes of visiting a particular area are not always the same for each gender, and different promotions are required. For example, in Ehime prefecture, providing toiletries and bath towels will please females, which seems likely to increase the number of female visitors to Ehime prefecture.

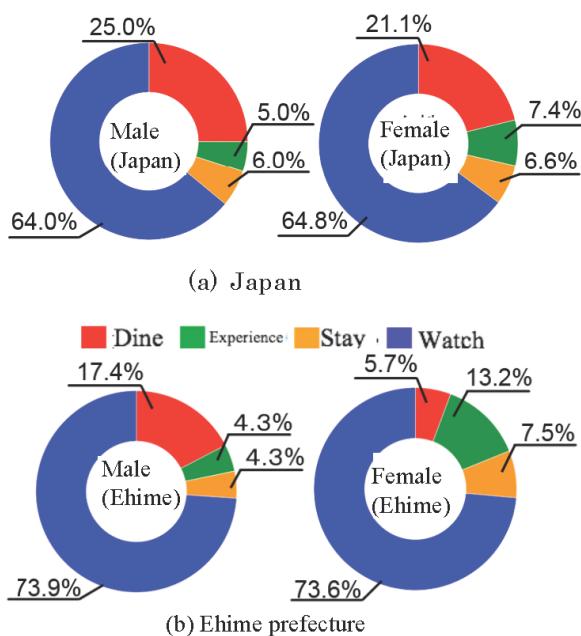


Fig. 3. Proportions of content types of travel blog entries for each gender

## 6. Conclusion

In this paper, we have analysed travellers' behaviour using travel blog entries. For this analysis, we used travellers' attributes identified automatically from travel blog entries. To identify gender, our method, SSL+CUE, obtained an accuracy score of 0.877, which outperformed the baseline method by 0.223. To identify languages, we used the langdetect program, and confirmed that it obtained precision of 0.972 and recall of 0.797. To identify the content type of each blog entry, our method, IG+MT, obtained precision of 0.597 and recall of 0.327. Using 7,490 travel blog entries with these attributes, we conducted behaviour analysis of 1,302 travellers, who visited Japan. We were able to derive useful information for tourist authorities.

## 7. Acknowledgements

This study was carried out under support of Ministry of Internal Affairs and Communications' SCOPE (Strategic Information and Communications R&D Promotion Programme).

## References

1. Wenger, A.: Analysis of Travel Bloggers' Characteristics and their Communication about Austria as a Tourism Destination, *Journal of Vacation Marketing*, 14(2), pp. 169-176 (2008)
2. Xia, J., Ciesielski, V. and Arrowsmith, C.: Data Mining of Tourists' Spatio-temporal Movement Patterns - A Case Study on Phillip Island, *Proceedings of the 8th International Conference on GeoComputation*, pp. 1-5 (2005)
3. Jonnson, C. and Devonish, D.: Does Nationality, Gender, and Age Affect Travel Motivation? A Case of Visitors to the Caribbean Island of Barbados, *Journal of Travel and Tourism Marketing*, 25(3-4), pp. 398-408 (2008)
4. Mack, R. W., Blose, J. E. and Pan, B.: Believe it or not: Credibility of Blogs in Tourism, *Journal of Vacation Marketing*, 14(2), pp. 133-144 (2008)
5. Akehurst, G.: User Generated Content: the Use of Blogs for Tourism Organizations and Tourism Consumers, *Journal of Service Business*, 3(1), pp. 51-61 (2009)
6. Li, Y.R. and Wang, Y.Y.: Exploring the Destination Image of Chinese Tourists to Taiwan by Word-of-Mouth on Web, *Proceedings of World Academy of Science Engineering and Technology*, 7, pp. 977-981 (2013)
7. Ikeda, D., Takamura, H. and Okumura, M.: Semi-Supervised Learning for Blog Classification, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1156-1161 (2008)
8. Yasuda, N., Hirao, T., Suzuki, J. and Isozaki, H.: Identifying Bloggers' Residential Areas, *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 231-236 (2006)
9. Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.: Effects of Age and Gender on Blogging, *Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199-205 (2006)
10. Saeki, K., Endo, M., Hirota, M., Kurata, Y., and Ishikawa, H.: Language-Specific Analysis of Domestic Places Visited by Foreign Tourists Using Crawled Twitter Data, *Tourism and Informatics*, 11(1), pp. 45-56 (2015) (in Japanese)
11. Fujii, K., Nanba, H., Takezawa, T., and Ishino, A. Enriching Travel Guidebooks with Travel Blog Entries and Archives of Answered Question, *Proceedings of ENTER 2016*, pp. 157-171 (2016)

## Inferring Tourist Behavior and Purposes of a Twitter User

Yuya Nozawa<sup>1,\*</sup>, Masaki Endo<sup>1,2</sup>, Yo Ehara<sup>3</sup>, Masaharu Hirota<sup>4</sup>, Syohei Yokoyama<sup>1</sup>,  
and Hiroshi Ishikawa<sup>1</sup>

<sup>1</sup>Graduate School of System Design, Tokyo Metropolitan University, Japan  
 {nozawa-yuya, endo-masaki}@ed.tmu.ac.jp,  
 ishikawa-hiroshi@tmu.ac.jp

<sup>2</sup>Division of Core Manufacturing, Polytechnic University, Japan

<sup>3</sup>National Institute of Advanced Industrial Science and Technology, Japan  
 y-ehara@aist.go.jp

<sup>4</sup>Department of Information Engineering, National Institute of Technology, Oita College, Japan  
 m-hirota@oita-ct.ac.jp

<sup>5</sup>Faculty of Informatics, Shizuoka University, Hamamatsu, Japan  
 yokoyama@inf.shizuoka.ac.jp

### Abstract

The importance of tourism information such as tourism purposes and tourist behavior continues to increase. However, obtaining precise tourist information such as the tourist destination and tourism period is difficult, as is applying that information to actual tourism marketing. We propose a method to classify Twitter user into tourist behavior and tourism purposes, extracting related information from Twitter posts. Our experiments demonstrated a 0.65 *F*-score for multi-class classification, showing accuracy for inferring tourist behavior and tourism purposes.

**Keywords:** Attribute estimation, Travel information,

### 1 Introduction

Recently, tourism occupies an important position in many countries as a key industry. Consumption activities related to tourism positively affect industries such as transportation, lodging, and manufacturing. Therefore, increasing the number of tourists is an important issue for governments and companies.

Providing tourists with sufficient information is demanded to increase tourism. For example, to obtain information related to tourist attributes and destinations, Japan's government has conducted surveys of tourism markets<sup>1</sup>. In this survey, inbound tourists state where they plan to sightsee, by questionnaire. The most important benefit of this approach is that the result has high reliability because the investigation method is face-to-face in almost all cases. Another advantage is that it is easy to ob-

---

<sup>1</sup> <http://www.mlit.go.jp/kankochosiryou/toukeisyouhityousa.html>

tain adequate information because the questioner can set the contents of questions as intended. However, this questionnaire method presents some shortcomings. First, the amount of data is limited by time and monetary costs for researchers. Second, it is difficult to set questions to obtain sufficient information in line with what researchers want to analyze. Third, even if a questioner sets complete question items to obtain the necessary information, the analyzer cannot use it flexibly after obtaining the results because of the fine granularity of the resulting information. Therefore, it is difficult to use questionnaire results obtained using the investigation method.

Therefore, some researchers specifically examine social media sites to extract tourist information instead of survey by questionnaires. Tourists have been posting reports of their impressions and opinions related to tourist attractions to social media sites. In Twitter, users post the contents on the fly and note impressions that one might experience at a tourist destination. Extracting information from Twitter have some benefits that are not provided by questionnaires administered to extract tourism information. First, to the method obviates the dispatch of research workers for the survey. Second, the analyzer can obtain huge amounts of data at low cost. Finally, no need exists to determine in advance what to analyze. One can determine issues of granularity, such as time and space of the data examined, to meet the analytical goals. However, it is difficult to obtain information such as population and distribution of sampling from Twitter. Therefore, Twitter is not suitable for research such as statistics, but suitable for research such as impression, behavior and senses.

**Table 1. Examples of Behaviors of "sightseeing" and "business"**

sightseeing	tour, tourism, spa, nature viewing, museums, concerts, movies, theater appreciation, participation events, botanical gardens, theme parks.
business	visit such as the head office, branches and factories, suppliers visit, participation in training and seminars, professional sports activities.

Therefore, we propose a method to extract information related to tourist behavior and tourism purposes from Twitter. To infer tourist behavior, we classify each tweet posted during a tourism period to classes of tourist behaviors. Then, to infer the tourism purpose, we classify a group of tweets posted during a tourism period to classes of tourism purposes. We define a user who posted tweets another area from the biosphere during a short period of time. Then, we regard that tourism purpose as representing the reason why a user does tourist activities. Tourism purpose classes are "sightseeing", "business" and "other purpose". Additionally, in this research, tourism purposes of "business" and "sightseeing" are inferred according to Table 1 created by the tourism Purpose of the Travel and Tourism Consumption Trends Survey of the Tourism Agency<sup>2</sup> as a reference. In addition, tourist behavior represents what a user was doing at the time of posting a tweet. For this research, we set five classes of tourist behavior as "sightseeing", "business", "eat", "buy" and "other behavior". In our approach, we infer the tourism purpose and tourist behavior in separate steps. For example, a user who goes on a business trip might post some tweets about liking

<sup>2</sup> <http://www.mlit.go.jp/kankochosiryou/toukei/shouhidoukou.html>

eating and sightseeing. The tourism purpose should be classified as "business". However, each activity should not be classified as "business", and those are classified as "sightseeing" and "eat". In other words, the tourism purpose and tourist behavior show a gap on tourism information that a researcher can analyze. Therefore, when obtaining detailed information of the user related to tourism, one must elucidate both classes separately.

## 2 Related research

Some researchers proposed the inference of user attributes. The main targets of estimation are gender [3], [4], age [4], [5], political-orientation [4], [6], residence [7], and occupation [8]. Classification of tourist behavior and tourism purpose to address in this research are also features of research related to the estimation of user attributes.

Research to extract tourist information from the information related to the Web has been conducted actively. Li et al. [9] proposed a method to divide travel purpose words into seven types, such as "scenic spots", "shopping", and "food". Zamal et al. [10] inferred "age", "gender", and "political preference" by combining estimation results and classifiers using the data and lexical identity of Twitter friend relationships. Ishino et al.[11] extracted traveler's transportation information automatically from travel blog entries written in Japanese using machine-learning techniques. Methods used in the estimation of user attributes are diverse. When classifying the attributes using supervised learning among them, Support Vector Machine (SVM) [12] is used particularly often. Benevenuto et al. [13] defined "spammers" who continue to send spam on Twitter. They are classified using SVM as "spammers" and "non-spammers". Pennacchiotti et al. [14] identify "political affiliation" and "particular ethnicity" and "Starbucks fans" from Twitter.

For this research, we propose a method to classify each tourist behavior as "sightseeing", "business", "eat", or "buy". The tourism purposes of "business" and "sightseeing" use SVM to target Twitter. A point of novelty of this research is that tourists of "business" purposes also do "sightseeing". To classify tourism only regarding purposes of "business" and "sightseeing", we classify the behavior what tourists are doing in tourism. In addition, there is a novelty even to the point of using multi-label SVM. We consider that the most tweets belong to more than one class.

## 3 Proposed method

In this section, we describe a method to infer Twitter user tourist behavior and tourism purposes using tweets. First, we extract the tourism period of a Twitter user from the tweets. Second, to infer tourist behavior during the tourism period, we classify each tweet to their tourism classes based on the tweet text. Finally, to infer the tourism purpose, we propose two methods: one based on text classification using tweets, and one using aggregation of the inferred results of tourist behavior.

As described in this paper, we define five classes of tourist behavior with four behavior classes "sightseeing", "business", "eat", and "buy", and "other behavior".

Additionally, we define two purpose classes as "sightseeing" and "business", and "other purpose". We define classes that represent other purposes for two reasons. First, sometimes, a user posts tweets without doing any tourism during travel period. Second, even after eliminating such noisy tweets, inferring appropriate classes is difficult because user behavior and purposes are widely varied.

### 3.1 Preprocessing

To obtain the tweets, we used the Twitter Streaming API<sup>3</sup>. At that time, we eliminated tweets that had been posted from countries other than Japan. Next, we apply preprocessing to the obtained tweets. We delete tweets including auto-generated texts from other social media sites. Additionally, we delete replies, retweets, URLs, and pictograms from the body of the acquired tweet.

### 3.2 Extraction of tourism-destination-related tweets

This section describes the procedure for extracting user tweets posted from the vicinity of a specific tourism destination. As described in this paper, we regard Tokyo as specific tourism destination. Therefore, we extract tweets of a Twitter user who stays outside Tokyo regularly, and who stays in Tokyo for a short time.

1. First, we sort the user tweets in chronological order. Additionally, we set a latitude-longitude bounding box of tourism destinations, and obtain all tweets that the user posted related to the tourism destination.
2. Second, we ascertain whether the user is a tourist or not. We assume that when the number of tweets posted at the tourism destination is sufficiently smaller than all tweets, then the tweets were posted during the travel period. Therefore, we calculate the percentage of the tweets posted at the tourism destination among all tweets. The user is defined as a tourist if the result is below the threshold. We tried each threshold as 0.1, 0.2, ..., 0.5 and confirmed that the best threshold is 0.3 by visual inspection. Here, tweets posted at the tourism destination by the tourist are defined as a "tourism tweet".

In almost all cases, the tourism tweets are continuous in all tweets. Therefore, we combined the texts of consecutive tweets posted during a tourist period to produce a single document. We designate the document as the "tourism period document". Tourism tweets are used to infer the tourist behavior. In addition, the tourism period document is used to infer the tourism purpose using SVM.

### 3.3 Vectorization

This section describes a means of vectorizing texts of the tourism tweets and tourism period documents. First, we apply morphological analysis to the text of tourism

---

<sup>3</sup> <https://dev.twitter.com/overview/documentation>

tweets. As described in this paper, we use nouns, verbs, and adjectives from extracted morphemes. For this research, we use MeCab<sup>4</sup> as a morphological analyzer, and mecab-ipadic-NEologd<sup>5</sup> to morphological analyzer. Next, we apply tf-idf to the words. Then, we apply Latent Semantic Analysis (LSA) [15], to reduce the tf-idf dimensionality. The process of vectorizing tourism period documents is the same.

### 3.4 Filtering of the "other behaviors" and "other purposes"

In this section, we describe the method used to filter "other behavior" and "other purpose". Both other classes represent the noisy class. Tourism tweets that are not classified as "sightseeing", "business", "eat", and "buy" is defined as "other behavior". In addition, tourism period documents that are not classified as "sightseeing" or "business" are defined as "other purpose". In the following steps, we classify tourism tweets and tourism period documents into four and two classes. Therefore, to improve performance of the steps, we filter those other classes preliminarily.

We use Support Vector Machine (SVM) to filter others. We classify tourism tweets and tourism period documents into "non-other behavior" and "other behavior" by SVM using a manually generated training dataset. Similarly, we filter tourism period documents. In the following steps, we apply our method to tourism tweets and tourism period documents without others classified in this step.

### 3.5 Inference of tourist behavior

In this section, we describe a method to classify a tourism tweet into the four classes of "sightseeing", "business", "eat", and "buy". We use the Multi-label SVM to classification using a feature vector created in Section 3.3. We use multi-label SVM because we infer that a tweet often includes several behaviors. For example, the tweet "I finished meeting at Tokyo, and will go to lunch" includes classes "business" and "eat". For that reason, multi-label classification is a more suitable method for inferring tourist behavior than multi-class classification. Here, multi-label SVM, which we use, is implemented in a combination of two-class SVM.

### 3.6 Inference of tourism purpose

This section presents a method to infer the tourism purpose of a user in tourism period. Classes of tourism purpose are three: "sightseeing", "business" and "other purpose". We propose approaches of two types: using results of tourist behavior calculated in Section 3.4; and using multi-class SVM based on tourism period documents. We present the following procedure for using tourist behavior results.

1. Documents are classified as "business" if at least one tourism tweet in a tourism period document was classified as "business". This is true because we assume that a

---

<sup>4</sup> <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

<sup>5</sup> <https://github.com/neologd/mecab-ipadic-neologd>

tourist entering the country for business might do something corresponding to the class of "sightseeing" after business matters are completed. However, a sightseeing tourist does not correspond to "business".

2. For a tourism period document that does not contain the class of "business" in a tourism tweet, a document that includes "sightseeing" is classified as "sightseeing".
3. The other case, a tourism period document is classified into "other purpose".

Next, we describe another method using tourism period documents and multi-class SVM. We apply multi-class SVM to a tourism period document. Then, we classify the document as "sightseeing", "business", or "other".

## 4 Evaluation

### 4.1 Experimental conditions

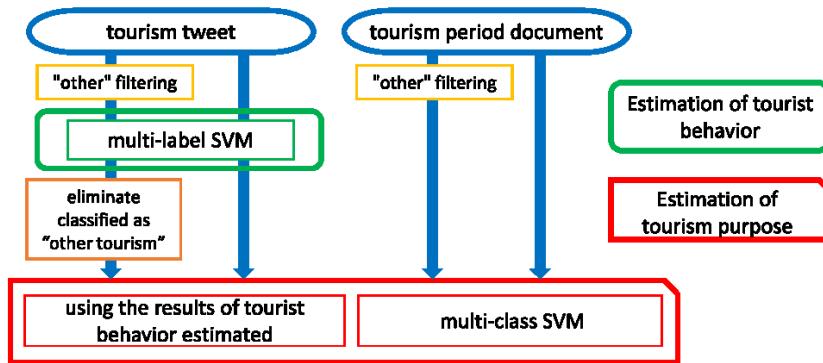
We use tweets posted in Japan posted during March 11, 2015 through October 28, 2015. By the processes described in Section 3.1 and Section 3.2, we obtained 68,588 tourists to Tokyo, 706,221 tourism-related tweets, and 218,052 tourism period documents. To prepare a ground truth of classification, we annotate tourism labels ("sightseeing", "business", "eat", "buy", and "other") to tourism tweets, and purpose labels ("sightseeing", "business", and "other"). In this process, the author annotates labels to tourism tweets and tourism period documents based on the contents to fit their natural feelings. Table 2 presents the number of each labels.

**Table 2. Number of classes and labels of tourism tweets and tourism period documents**

	sightseeing	business	eat	buy	other	total
tourism tweet	1,073	702	928	659	1,754	5,107
tourism period document	381	468			349	1,198

We vectorize the tourism period documents and tourism tweets using the process described in Section 3.3. The dimensions are increased from 100 to 500 dimensions by 100 dimension steps. Based on experiments conducted with multi-class SVM, we used 400-dimensional in following experiments, which scores the highest accuracy.

All SVMs employed for the experiments used a Gaussian kernel. Additionally, we experiment using five cross-validation, and hyperparameters  $c$  and  $\gamma$ , which have the highest accuracy in each experiment. We used Accuracy, Precision, Recall, and the  $F$ -Score as evaluation criteria to compare classification results.



**Fig. 1. Overview of the experiment.**

#### 4.2 Experimental procedures

This section presents a description of the method of the experiments performed in this paper. First, Figure 1 presents an overall view of the experimental.

For evaluation of tourist behavior inferences, we experiment with two methods. First, we apply "other" filtering to tourism tweets, and infer tourist behavior by a multi-label SVM. It is named "filtering and SVM". Second, we do not apply "other" filtering to tourism tweets, and infer tourist behavior. It is named "only SVM".

For tourism purpose inference, we experiment with four methods. First method is that we infer the tourism purpose using the inferred class of tourist behavior in "filtering and SVM". At the time, we eliminate the tweets classified as "other tourism". It is named "eliminate and using result of SVM". Second method is inference of tourism purposes using all tourism tweets (including "other tourism"). It is named "only using result of SVM". Third method is applied to "other" filtering to the tourism period documents. We infer tourism purposes using a multi-class SVM. It is named "filtering and using document". In addition, fourth method is one by which we infer tourism purpose using a multi-class SVM without a procedure to eliminate "other" filtering from tourism period documents. It is named "only using document".

#### 4.3 Experiment results and discussion

##### 4.3.1 Evaluation of "other" filtering

First, we evaluate "other" filtering to eliminate "other tourism" and "other purpose". Our method uses "other" filtering in "filtering and SVM", to tourism tweets, and in "filtering and using document", to tourism period documents. Table 3 portrays the classification results of "other" filtering for tourism tweets and tourism period documents. Table 4 shows the number of tourism tweets and tourism period documents erased by "other" filtering. Here, Accuracy, Precision, Recall, and F-Score show average values in each class. In Table 3, recall of "other" is low in both the tourism

tweets and tourism period documents. Moreover, in Table 4, the most eliminated class is "other" of tourism tweets and tourism period document about 33%. In addition, the other class accounts for a few percent. Although this indicates that eliminating "other" is insufficient, this procedure is effective for "other behavior" and "other purpose" because the recall of "non-other" is high (i.e., misclassification of those classes is less, but this procedure can eliminate "other" classes).

The reason why the recall of "other" is low is the diversity of tweets. Results show that the contents of those tweets are various. It is unrealistic to train the characteristics of those "other" tweets to eliminate "other" classes. Therefore, we infer that "other" filtering is more important to classify "other" classes mistakenly into other classes.

**Table 3. Results of "other" filtering: evaluation index**

applies	accuracy	class	precision	recall	F-score
tourism tweet	0.742	other	0.852	0.326	0.472
		non-other	0.722	0.970	0.828
		avg. /total	0.770	0.744	0.700
tourism period document	0.730	other	0.442	0.228	0.266
		non-other	0.752	0.940	0.834
		avg. /total	0.664	0.730	0.670

**Table 4. Results of "other" filtering: number of erasures**

applies	class	Before	after	erasure rate (%)
tourism tweet	business	702	665	5.270
	sightseeing	1,073	1,038	3.262
	buy	659	643	2.428
	eat	928	917	1.185
	other	1,745	1,177	32.550
tourism period document	business	468	447	4.487
	sightseeing	381	351	7.874
	other	349	269	22.922

#### 4.3.2 Evaluation of tourist behavior inference

We describe the performance of tourist behavior inference. For tourist behavior inference, two types of the experimental methods are shown in Section 4.2. The experimentally obtained results for each of the method are shown in Table 5. There, the average *F*-score of all classes in "filtering and SVM" and "only SVM" is about 0.65. The *F*-score of "sightseeing" in "filtering and SVM" and "only SVM" is low, compared to each attribute. Here, from Table 1, many types of behavior corresponding to the "sightseeing" than "business". Therefore, "sightseeing" is more diverse than the others and the vector of the tourism tweets is a variation to classify. As a result, the *F*-score of the "sightseeing" in classification is regarded as lower than other attributes.

**Table 5. Classification results of tourist behavior**

method	Accuracy	label	precision	recall	<i>F</i> -score
filtering and SVM	0.544	business	0.758	0.646	0.698
		sightseeing	0.642	0.556	0.596
		buy	0.788	0.666	0.722
		eat	0.744	0.652	0.694
		other	0.616	0.572	0.592
		avg. / total	0.698	0.612	0.648
only SVM	0.570	business	0.718	0.650	0.676
		sightseeing	0.638	0.536	0.582
		buy	0.724	0.676	0.696
		eat	0.728	0.644	0.684
		other	0.710	0.706	0.710
		avg. / total	0.704	0.646	0.672

Tables 6 and 7 present results and the labels of ground truth of "filtering and SVM" and of "only SVM". In those tables, "correct" represents the labels of ground truth, which were annotated manually. "estimate" represent the labels inferred using the proposed method. In those tables, almost all tourism tweets of "buy" were classified correctly because many words such as "buy" or "souvenirs", which are clearly related to buying behavior, appear in the text of the tourism tweet of the class "buy". In addition, some tourism tweets related to "sightseeing" were misclassified as "other behavior". We consider that vector variation of tourism tweets can lead to responses of "other behavior" and "sightseeing", as described earlier. The class of "sightseeing" is divided to sub-classes that contain more details of behavior.

Furthermore, in Tables 6 and 7, some tweets are classified into "unlabeled". The multi-label SVM is configured with multiple two-class SVMs in each attribute. The SVM classifies a tourism tweet according to its attributes or not. Therefore, the SVMs do not often classify the tweet into any class. However, the tweet is originally classified as "business". This method misclassifies the tweet as "unlabeled".

Here, we discuss that we apply "other" filtering to tourism tweets or not. In Table 5, the average of each criterion of "filtering and SVM" is less than "only SVM". In addition, the value of "other tourism" in "filtering and SVM" is less than "only SVM", although "filtering and SVM" eliminates tourism tweets "other tourism". The evaluation value of the classes other than "other tourism" in "filtering and SVM" is higher than "only SVM". This result presents the performance of "other" filtering is sufficient, but "other tourism" in "filtering and SVM" is difficult to eliminate because classifying easily the tourism tweets is already finished. In addition, the reason why the evaluation value of "other tourism" in "filtering and SVM" is less than "only SVM" is the average contains score of classifying tourism tweets into "other tourism" or not. However, because our purpose is inferring the tourist behavior, we regard the evaluation values of classes other than "other tourism" are more important. Therefore, we regard "filtering and SVM" as better than "only SVM" to infer tourist behavior.

**Table 6. Results of labeling using “filtering and SVM”**

correct\estimate	business	sightseeing	buy	eat	other	unlabeled
business	432	53	16	70	88	79
sightseeing	42	579	52	78	206	155
buy	19	42	430	49	49	90
eat	71	89	44	598	85	113
other	78	175	42	81	671	212

**Table 7. Results of labeling using “only SVM”**

correct\estimate	business	sightseeing	buy	eat	other	unlabeled
business	452	47	25	79	115	82
sightseeing	48	577	57	76	232	169
buy	23	49	444	52	68	94
eat	85	82	49	597	105	127
other	64	186	76	92	1,230	213

#### 4.3.3 Evaluation of tourism purpose inference

We describe the performance of the tourism purpose inference. In Table 8, we present results of tourism purpose inference performance. This evaluation compares 4 methods. Compared to each method, it is apparent that classification performance using the tourist behavior is better than that of multi-class SVM using the tourism period documents. In the proposed method, although we regard all tweets during the tourism period as tourism period documents, the key tweets used to ascertain the class of tourism purpose are few tweets during the tourism period document. Therefore, tweets other than the key tweets might adversely affect classification performance. As a result, “eliminate and using result of SVM” and “only using result of SVM” are better than “filtering and using document” and “only using document” to infer tourism purposes.

Next, comparison of methods reveals that the tourism purpose inference performance is improved by eliminating “other purpose” because the classification performance of the other class is improved by erasing “other purpose”. In addition, a tourism period document that does not contain “sightseeing” and “business” during the tourism period is not classified as “other purpose”. In Section 5.2, we confirmed that the classification performance of other attributes improves by eliminating the previous “other”. However, in the tourist behavior, it is regarded as the “other purpose” is a set of “other behavior”, “eat”, and “buy”. Therefore, the classification performance of “other purpose” is considered to have improved in the tourism purpose. Consequently, it is probably useful to erase “other purpose” for tourism purpose inference.

**Table 8. Classification results of tourism purposes**

method	accuracy	class	precision	recall	<i>F</i> -score
eliminate and using result of SVM	0.760	business	0.830	0.770	0.800
		sightseeing	0.740	0.720	0.730
		other	0.700	0.780	0.740
		avg. / total	0.760	0.760	0.760
only using result of SVM	0.720	business	0.790	0.740	0.760
		sightseeing	0.750	0.640	0.690
		other	0.620	0.770	0.690
		avg. / total	0.730	0.720	0.720
filtering and using document	0.654	business	0.704	0.796	0.744
		sightseeing	0.664	0.640	0.650
		other	0.554	0.452	0.482
		avg. / total	0.654	0.654	0.646
only using document	0.652	business	0.654	0.852	0.742
		sightseeing	0.682	0.546	0.602
		other	0.634	0.504	0.556
		avg. / total	0.658	0.654	0.642

## 5 Conclusion

For this research, we proposed a method to infer user tourism purposes and tourist behaviors at a tourist destination using tweet geo-tagging and text posted to Twitter as one approach to extract tourist information. Our proposed method classifies a feature vector by a multi-label SVM as tourist behavior related to "sightseeing", "business", "eat", "buy" and "other behavior". Subsequently, we classify tourism period documents into tourism purposes of "tourism", "business", and "other" using tourist behavior inference results. The evaluation experiment showed *F*-scores of tourist behavior and tourism purposes were both about 0.65.

As future applications, we expect to visualize classification results obtained using the proposed method based on time. The obtained valid information might increase tourist arrivals by enabling them to visualize the area, and to analyze tourist behavior and position at the time of that behavior according to tourism purposes.

### ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 16K00157, 16K16158, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas "Research on social big data". We are grateful for the assistance by Yoshiyuki Shoji.

### REFERENCES

1. Fang, Guanshen, Sayaka Kamei, and Satoshi Fujita. "How to extract seasonal features of sightseeing spots from Twitter and Wikipedia (Preliminary Version)".

*Bulletin of Networking, Computing, Systems, and Software*, Vol. 4, No. 1, pp. 21-26, 2015.

2. Bannur, Sushma, and Omar Alonso. "Analyzing temporal characteristics of check-in data". *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web*, pp. 827-832, 2014.
3. Burger, John D and Henderson, John and Kim, George and Zarrella, Guido. "Discriminating gender on Twitter". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301-1309, 2011.
4. Rao, Delip and Yarowsky, David and Shreevats, Abhishek and Gupta, Manaswi. "Classifying latent user attributes in Twitter". *Proceedings of the Second International Workshop on Search and Mining User-Generated Contents*, pp. 37-44, 2010.
5. Burger, John D., and John C. Henderson. "An Exploration of Observable Features Related to Blogger Age". *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 15-20, 2006.
6. Pennacchiotti, Marco, and Ana-Maria Popescu. "Democrats, republicans and starbucks aficionados: user classification in Twitter". *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 430-438, 2011.
7. Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating Twitter users". *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759-768, 2010.
8. Sloan, Luke and Morgan, Jeffrey and Burnap, Pete and Williams, Matthew. "Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data". *PloS One*, Vol.10, No. 3, 2015.
9. Li, Y. R., and Y. Y. Wang. "Exploring the Destination Image of Chinese Tourists to Taiwan by Word-of-Mouth on Web". *Proceedings of World Academy of Science, Engineering and Technology*, No. 79, pp. 977, 2013.
10. Al Zamal, Faiyaz, Wendy Liu, and Derek Ruths. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors". *International Conference on Weblogs and Social Media*, Vol. 270, 2012.
11. Ishino, Aya and Nanba, Hidetsugu and Takezawa, Toshiyuki. "Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries", *ENTER*, pp. 113-12, 2011.
12. Vapnik, Vladimir Naumovich, and Vlaminir Vapnik. "Statistical learning theory". Vol. 1, 1998.
13. Benevenuto, Fabricio and Magno, Gabriel and Rodrigues, Tiago and Almeida, Virgilio. "Detecting spammers on Twitter". *Collaboration, Electronic Messaging, Anti-Abuse AND Spam Conference*, Vol. 6, pp. 12, 2010.
14. Pennacchiotti, Marco, and Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification". *International Conference on Weblogs and Social Media*, Vol.11. No.1, 2011, pp.281-288.
15. Deerwester, Scott and Dumais, Susan T and Furnas, George W and Landauer, Thomas K and Harshman, Richard. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391, 1990.

## Application of Annotation Smoothing for Subject-independent Emotion Recognition based on Electroencephalogram

Nattapong Thammasan<sup>1</sup>, Ken-ichi Fukui<sup>2</sup>, and Masayuki Numao<sup>2</sup>

<sup>1</sup> Graduate School of Information Science and Technology, Osaka University  
Suita-shi, Osaka 565-0871, Japan

[nattapong@ai.sanken.osaka-u.ac.jp](mailto:nattapong@ai.sanken.osaka-u.ac.jp)

<sup>2</sup> Institute of Scientific and Industrial Research (ISIR), Osaka University  
Ibaraki-shi, Osaka 567-0047, Japan

[{fukui, numao}@ai.sanken.osaka-u.ac.jp](mailto:{fukui, numao}@ai.sanken.osaka-u.ac.jp)

**Abstract.** In the construction of computational models to recognize emotional state, emotion reporting continuously in time is essential based on the assumption that emotional responses of a human to certain stimuli could vary over time. However, currently existing methods to annotate emotion in temporal continuous fashion are confronting various types of challenges. Therefore, the manipulation of the annotated emotion prior to labeling training samples is necessary. In this work, we present an early attempt to manipulate the emotion annotated in arousal-valence space by applying three different signal filtering techniques to smooth annotation data; moving average filter, Savitzky-Golay filter, and median filter. We conducted experiments of emotion recognition in music listening tasks employing brainwave signals recorded from an electroencephalogram (EEG). Smoothed annotation data were used to label the features extracted from EEG signals to train emotion recognizers using classification and regression techniques. Our empirical results indicated the potential of the moving average filter that could increase the performance of emotion recognition evaluated in subject-independent fashion.

**Keywords:** Emotion recognition, Electroencephalogram, Music-emotion, Annotation, Smoothing

### 1 Introduction

Among the endeavors of building computational models capable of perceiving human emotion, recent researchers have emphasized the necessity of estimating emotional state continuously over the course of time (Gunes and Schuller, 2013). In particular, the empathic computing systems are expected to be capable of capturing emotional fluctuation of users and properly respond almost instantly. Beyond the continuity in time, psychological researchers proposed to describe emotion in multi-dimensional continuous space as a dimensional continuous model could resolve the ambiguity issues occurred in using discrete categories to describe emotion. Among these, arousal-valence model (Russell, 1980)

was one of the most commonly exploited models to describe emotional states; arousal indicates emotional intensity ranging from calm to activated emotion, whereas valence, on the other dimension, describes positivity of emotion ranging from unpleasant to pleasant. Considering the combined continuity of emotion in time and space, the efforts to track emotion variation in the continuous arousal-valence space have emerged.

However, the approach is still in its infancy and confronting with a variety of challenges. In particular, it lacks standard methodology to report emotion continuously in time and space. The existing emotion annotation approaches have certain limitation (Metallinou and Narayanan, 2013). For instance, unfamiliarity with annotation tools could lead to the inconsistency and the contamination of noise in emotion annotation data while familiarizing the user with annotation tools by increasing practicing time could be annoying and lead to the impracticality of the systems. In addition, the delay of annotation could result in a mismatch between the emotional cues and the reported emotions (Mariooryad and Busso, 2013). Therefore, completely relying on the annotation data reported by the user would degrade the performance of emotion recognition systems. By these reasons, the manipulation of emotion annotation data prior to proceeding to emotion recognition system construction is undoubtedly essential but was usually not taken into consideration in previous research.

In this work, we propose an early attempt of manipulating user's emotion annotation data continuously self-reported through time in the continuous arousal-valence space. In particular, we introduce an application of signal smoothing techniques to alleviate the adverse effect of annotation noises. While continuous emotion annotation can be performed using a variety of tools (e.g., mouse-moving (Cowie et al., 2000) and joystick (Soleymani et al., 2012)), we focus on mouse-clicking emotion annotation approach for the sake of simplicity and less cognitive processing requisition. In our experiment, electroencephalogram (EEG) was used to record brainwave signals in music listening tasks, then informative features were extracted from EEG responses and such features were labeled with the smoothed annotation data. Afterward, the continuous emotion recognition models were established by machine learning approaches using classification technique (Thammasan et al., 2016) and regression technique (Soleymani et al., 2016). The performance of emotion recognition was evaluated by using subject-independent approach aiming the border goal toward generic emotion recognition systems.

## 2 Research Methodology

### 2.1 Experimental protocol

Fifteen healthy male subjects aged 22-30 years (mean age = 25.52 years, SD = 2.14 years) participated in the experiment. All of them were students of Osaka University and had a minimal formal musical education. At the beginning, each subject was instructed to select 16 MIDI songs from a 40-song music collection (mean length = 106.3 s, SD = 16.2 s, range = 73–147 s). For further investigation,

the selected songs were controlled to be comprising of an equal number of familiar and unfamiliar songs.

Afterward, the selected songs were presented to the subject as synthesized sounds using the Java Sound APIs MIDI package<sup>3</sup>; a 16 s silent resting period was inserted at the interval between each song to reduce any effect of the previous song. During music listening, brainwave signals were recorded using Waveguard EEG<sup>4</sup> placed in accordance with the 10-20 international system, and acquired at a sampling rate of 250 Hz. Positions of the twelve selected electrodes (Fp1, Fp2, F3, F4, F7, F8, Fz, C3, C4, T3, T4, and Pz) were nearby frontal lobe, which is believed to be play an importance role in emotion processing (Koelsch, 2014), whereas the vertex electrode (Cz) served as a reference electrode. Over the entire course of EEG recording, the impedance of each electrode was maintained below 20 kΩ. EEG signals were amplified by Polymate AP1532 amplifier<sup>5</sup> and visualized on APMonitor<sup>6</sup>. Each subject was instructed to keep his eyes close and limit body movement during music listening to alleviate any effect owing to unrelated artifacts. The EEG signals were filtered using a 60-Hz notch filter to suppress the noise of the electric power line followed by a 0.5–60 Hz band-pass filter to eliminate the unrelated noises. Eye-movement artifacts were corrected by applying the independent component analysis of the EEGLAB toolbox (Delorme et al., 2011).

After finishing the listening session of all sixteen songs, each subject proceeded to the emotion annotation session without EEG recording. The annotation was conducted through our developed software. Each subject was instructed to annotate the emotions that were perceived in the previous session by continuously specifying at a corresponding point on the arousal-valence space displayed on a monitor screen (Fig. 1). At the end of the emotion annotation of each song, each subject rated the confidence level, on a scale of ranging from 1 to 3, of the correspondence between the emotions perceived during the first listening phase and the annotated emotions. Each subject was also encouraged to perform the emotion annotation of a particular song again in the case that the annotated data for that song was not satisfied yet. A brief guideline of arousal-valence emotion model was provided throughout annotation session to acquaint each subject with the model. Arousal and valence annotation data were recorded independently as numerical values ranging from -1 to 1. Eventually, the annotation data was associated with the artifact-corrected EEG signals via timestamps. Unfortunately, we discarded all data from the two subjects who reported drowsiness during EEG recording.

## 2.2 Annotation Smoothing

The assumption for the mouse-clicking emotion annotation is that annotated emotion is stable for a certain duration until the next annotation is clicked.

<sup>3</sup> <http://docs.oracle.com/javase/7/docs/technotes/guides/sound/>

<sup>4</sup> <http://www.ant-neuro.com/products/waveguard>

<sup>5</sup> <http://www.teac.co.jp/industry/me/ap1132/>

<sup>6</sup> Software developed for Polymate AP1532 by TEAC Corporation

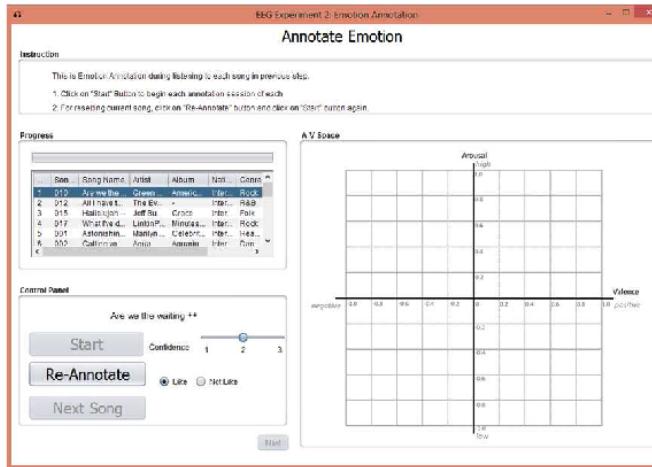


Fig. 1: A screenshot of the annotation software

However, this assumption is quite far from realistic because emotion in real life gradually changes from one to another state rather than abruptly shifts from one to another. Therefore, our mouse-clicking emotion annotation tool has limited practicability despite our encouragement for subject to annotate emotion in a high frequency. Consequently, adjusting the curve of emotion annotation that could provide a smoother curve reflecting the gradual changing of emotion might be necessary as the curve could be considered as more corresponding to the real characteristic of emotion. In addition, the noises owing to unfamiliarity to the tools of some annotators and unintentionally mouse clicking could inevitably contaminate the annotation data. Therefore, adjusting the annotation curve by using a certain approximation function could also correct these noises.

To resolve the issues, we hereby propose the techniques to manipulate the annotation data by applying time-series data smoothing techniques directly to the annotation curves. Inspired by signal processing techniques, we introduce an application of three commonly used filtering techniques implemented by MATLAB Signal Processing Toolbox<sup>7</sup> to smooth our annotation curve. As the characteristic of the smoothed curves highly relies on the size of filter frame, we also examined the influence of the size to our emotion recognition performance by varying the filter frame size from 501 to 8001 points (equivalent to 2.004 and 32.004 s respectively) at a step of 500 points.

**Moving Average Filter** Owing to its simplicity, the moving average is one of the most commonly used filter in digital signal processing to smooth out short-term fluctuations while preserving long-term trends. It is a premier filter for

<sup>7</sup> <http://www.mathworks.com/products/signal/>

time domain encoded signals. Each of the output points is obtained by taking the average of a number of points from the input signals within a sliding filter frame. However, applying the moving average filter results in a delay by the half of filter frame size. To handle the delay, we shifted the annotation curve forward to the corresponding point and replicated the end point in the annotation data of a particular song to compensate the missing data with a size of the delay owing to the shifting technique.

**Savitzky-Golay Filter** Savitzky-Golay filter is another signal smoothing technique without greatly distorting the signal (Savitzky and Golay, 1964). Different from applying the moving average filter, the high-frequency components of signals can be successfully preserved. The filter operates by fitting a low-degree polynomial to a set of data points in a sliding filter frame by the method of local least-squares polynomial approximation and then evaluating the resulting polynomial at a single point within the approximation interval. In this work, we applied Savitzky-Golay filter to smooth annotation data with the two different levels of polynomial degree — three (cubic) and four (quartic) — to examine the effect of the polynomial degree to the smoothed annotation data and the performance of the following emotion recognition.

**Median Filter** Unlike the other filters that mainly eliminate the sharp edges of input signals, a median filter is an alternative filter that has prominence in smoothing input signals while preserving the edge of signals. Within a sliding filter frame, smoothed signals are obtained by deriving the median of a number of points from the input signals.

### 2.3 Experiments of Emotion Recognition

Inspired by the successful results in previous EEG-based emotion recognition studies (Sourina et al., 2012; Thammasan et al., 2016), the fractal dimension (FD) approach was exploited to extract informative features from the EEG signals. FD value is a non-negative real value that characterizes the complexity and irregularity of a time-varying signal and it could be used to indicate brain states from EEG signals (Sourina et al., 2011). In this study, we implemented Higuchi algorithm (Higuchi, 1988) to calculate FD values. As reported as effective features to estimate emotional states (Koelstra et al., 2012; Thammasan et al., 2016; Lin et al., 2014), asymmetric features were also included in our original feature set; the feature was derived from the difference of FD value extracted from an electrode on the left hemisphere and that from the symmetric lateral electrode on the right hemisphere. As there were five symmetric electrode pairs, the total number of the features extracted from EEG signals is, therefore, 17. To capture temporal dynamics of the emotional states, we applied a non-overlapping 4 s sliding window segmentation technique; the size of 3-6 s was reported to achieve the highest performance in emotion classification in literature (Candra et al., 2015).

Afterward, the extracted features were labeled by the corresponding annotation data prior to constructing emotion recognition models.

Next, we conducted the experiments of emotion recognition by applying classification and regression techniques. For the sake of simplicity, the outputs of emotion classification are the binary classes of arousal and valence, while the outputs of regression are numerical numbers ranging from -1 to 1 of arousal and valence estimation in the two-dimensional emotion space.

**Classification** Despite the continuity of the arousal-valence space, we converted emotion recognition into the binary classification of arousal and valence independently. Arousal classification was to classify high and low arousal, while valence classification was to classify positive and negative valence, whereas the sign of the annotated numerical rating was used to define the class of arousal and valence. A majority method was adopted to determine the emotional label for a particular window containing emotional class shifting. The features were scaled for each subject between [0,1] using *min-max* strategy. As a classifier, we adopted support vector machine (SVM) based on Gaussian radial basis kernel function (RBF) using MATLAB Statistics and Machine Learning Toolbox<sup>8</sup>; the SVM was found to be popular and successful classifier in the research of EEG-based emotion recognition (Kim et al., 2013). The kernel scale of RBF kernel was set as 0.5.

The evaluation of emotion classification was performed by using subject-independent approach. This approach trains and tests emotion classification model using the aggregated data from all subjects. It can apparently reflect the degree of generalization of the emotion recognition model. In this work, we adopted the leave-one-subject-out validation method to evaluate the performance of classification. In each trial, the classifier was trained by using the combined data from twelve subjects and then the trained classifying model was tested against the data from the remaining subject. The results from each trial were averaged to derive overall performance.

**Regression** Utilizing the continuity of arousal-valence space, we also performed emotion recognition by using regression technique to recognize arousal and valence independently as the technique could provide an estimated emotion as numerical values of arousal and valence in the arousal-valence space. Compared to classification, regression could provide finer detail of the estimated emotion and the different emotions belonging to the same quadrant in the arousal-valence space can be distinguished. In this work, we applied the support vector machine regression based on Gaussian kernel implemented by using MATLAB Statistics and Machine Learning Toolbox<sup>3</sup> was used to estimate emotion; the kernel scale of Gaussian kernel was set as 0.5. To label the features extracted from a particular sliding window with an emotional tag, the averaged values of arousal and valence in that window were used.

<sup>8</sup> <http://www.mathworks.com/products/statistics/>

The performance of emotion recognition using regression was evaluated by using the Pearson correlation to reflect the similarity between the estimated curves and the annotation data, and the mean square error (MSE) to represent the disparity (loss) between the estimated emotion and the ground truth. Similar to emotion classification, the overall performance of emotion recognition was evaluated in subject-independent strategy by using leave-one-subject-out cross-validation.

### 3 Results

For clarity, the main purpose of this current work is to study the feasibility of applying the smoothing technique to the annotation data aiming to enhance the performance of emotion recognition. First, we examined the shapes of the smoothed annotation curves. Next, we assessed the performance of emotion recognition with the smoothed annotation data using subject-independent evaluation. The high averaged confidence level (2.4109, SD = 0.6676) of the annotation across subjects suggested the validity of the annotation data.

#### 3.1 Annotation Smoothing Results

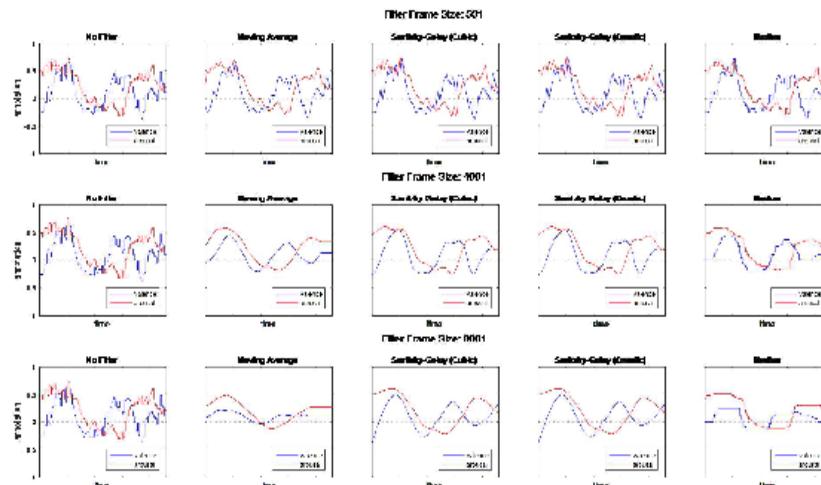


Fig. 2: An example of the resulting smoothed annotation data compared to the original annotation data of a song by Subject 12 (song length = 91.6 s)

To illustrate the results of applying the smoothing technique to the annotation data, we exemplify by using the emotion annotation data of a song by

Subject 12. We varied the size of filter frame from small (501 points) to medium (4001 points) and large (8001 points) frame. The comparison between original annotation data, which contains the variation of the arousal and valence over the course of time, and the smoothed annotation data is shown in Fig. 2. As can be seen, the smoothed annotation data was similar to the original annotation data when using a small filter frame size. Specifically, the moving average filter could noticeably remove the high-frequency fluctuation in the annotation. However, when the filter frame size was enlarged to medium level, we found that the trend of annotation curve was preserved but the high-frequency fluctuation was dramatically reduced for any filtering technique. Interestingly, flat plateaus were also found in the resulted annotation data when applying the median filter. Nevertheless, when we applied filtering technique with large filter frame size, the shape of the annotation data was highly distorted in comparison to the original annotation data, especially for the moving average filter and the median filter. Specifically, the height of the resulted curve was distinctly dissimilar to the original curve when applying either filter, while the Savitzky-Golay filter could still preserve the height of the curve regardless of the polynomial order.

### 3.2 Results using Classification

The averaged classification accuracies across subjects are illustrated in Fig. 3. As can be seen, the obtained accuracies were higher than the random classification (50% accuracies for two-class classification) but the limited performance suggested that the classification might suffer from the inter-subject variability in EEG signals and/or in emotion annotation strategies. The enhancement of the emotion classification performance as filter frame size enlarged was found, especially by using the moving average filter. The Savitzky-Golay filter also achieved the improved classification results, where the cubic filter could outperform the quartic filter. In overall, the results suggested the promise of the moving average filter to upgrade the performance of emotion recognition.

### 3.3 Results using Regression

The performance of emotion recognition using regression is shown in Fig. 4. Similar to emotion classification, the low performance of the subject-independent emotion recognition suggested the adverse effect of the inter-subject variability. Despite the limited performance, all of the approaches could enhance the performance of arousal estimation, especially when increasing the filter frame size to a certain extent. Moving average filter achieved the improved performance in majority cases of arousal and valence recognition suggesting the promise of this technique. The proper size of filter frame was approximately between 18 s (4501 points) and 24 s (6001 points). In addition, the median filter was found to be another potential approach to upgrade arousal recognition performance. Furthermore, the lower polynomial order of Savitzky-Golay filter can be preferred as the cubic filter achieved better performance in comparison to the quartic filter.

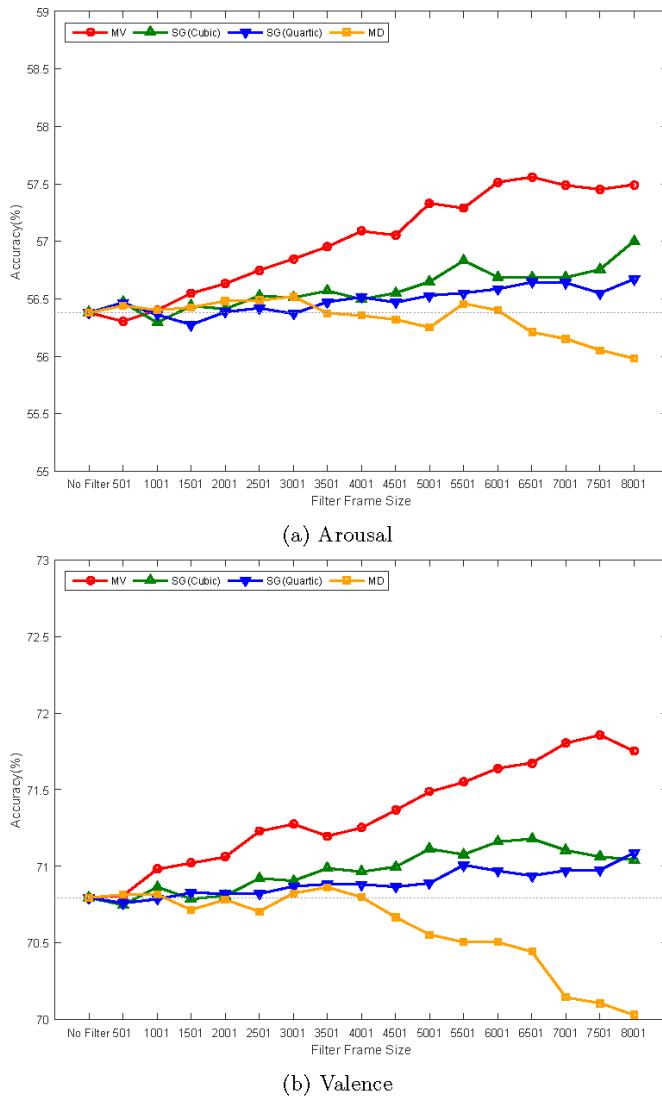


Fig. 3: Results of subject-independent emotion classification using the annotation data smoothed by the moving average (MV), Savitzky-Golay (SG), and median (MD) filters

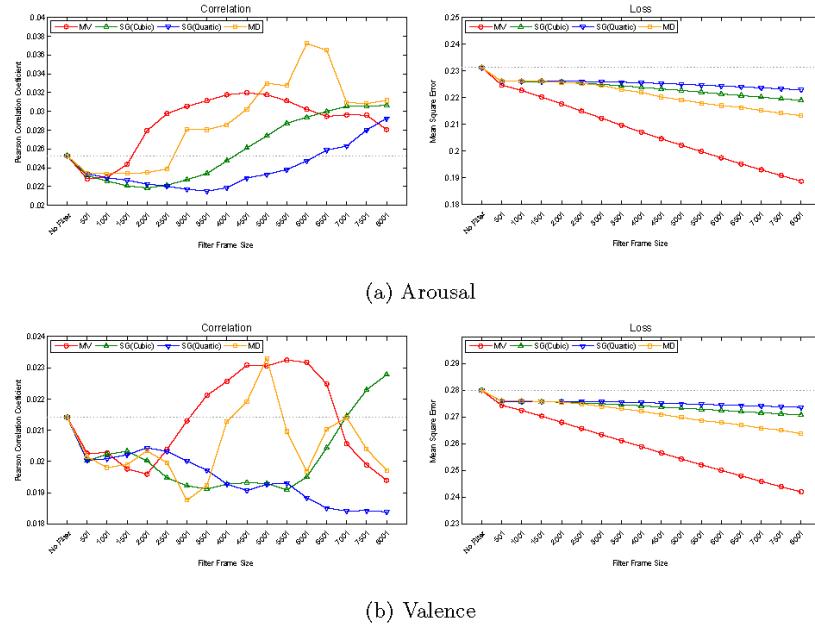


Fig. 4: Performance of subject-independent emotion recognition using regression evaluated on the annotation data smoothed by the moving average (MV), Savitzky-Golay (SG), and median (MD) filters

#### 4 Discussion

The primary goal of this current work is to improve the performance of emotion recognition, and we propose techniques of smoothing annotation data as an approach to manipulate the annotation data prior to training emotion recognizer. The empirical results suggested that our proposed methods could slightly enhance the performance of emotion recognition either using classification or regression techniques. However, several issues leave room for discussion.

Even though the moving average filter has demonstrated its promise in improving emotion recognition, the considerable distortion of the smoothed annotation data in comparison to the original data (shown in Fig. 2) suggested the necessity of trading off between enhancing the performance of emotion recognition and the disparity between the smoothed annotation curve and the original curve. With the consideration of sufficient accuracy of the smoothed annotation data, a statistical parameter capable of indicating the extent of such distortion is encouraged to be introduced in future works. In addition, future study should also include the reporting of annotation strategy and annotation fatigue of each annotator to enable the feasibility to distinguish the unintentional annotation

noises and the annotation with full intention aiming to investigate the validity of the smoothed annotation data. Also, analyzing the fatigue in annotation could provide insight of the validity of annotation data.

According to the results, we found that the inter-subject variability in EEG signals and/or annotation largely involved in the low performance of subject-independent emotion recognition. Incorporating subjective factor, e.g. gender, music familiarity, or music preference, into the emotion recognition model is expected to increase the performance of emotion recognition, and this is considered as our possible future work.

As the familiarity was the main constraint in the song selection phase, we barely found the commonly selected set of songs by the majority of the subjects; each song was selected by the population of subjects for 6.00 times on average ( $SD = 2.76$ ). This resulted in the difficulty to investigate the variance between subjects of the EEG signals and the annotation when listening to the same song. Conducting research by using the same musical stimuli for every subject is worthwhile. The possible agreement of annotation across subjects in a particular song would suggest the possibility to use a general model to recognize emotion in subject-independent fashion. On the other hand, the annotation disagreement would suggest that it could be preferable to construct specific emotion recognition model for a group of subjects having similar annotation rather than creating a generalized model for all subjects.

Though we have presented the application of data smoothing technique to manipulate annotation data, the obtained results were limited to the study using mouse-clicking annotation tool. To generalize the practicability of our proposed methods, conducting further experiments with another dataset in affective computing research that used different annotation tools (e.g., a dataset using joystick annotation (Soleymani et al., 2012)) is also encouraging. In addition, with the aim to generalize the applicability of our emotion recognition approach, our future works include conducting experiments with broader groups of subjects; for instance, recruiting female subjects or aging population to participate in our research could be done in the future.

## 5 Conclusion

In this work, we introduce the methods to manipulate the emotion annotation data prior to proceeding into emotion recognition phase. We applied signal smoothing techniques directly to the acquired annotation data. Based on the empirical results, evaluated in subject-independent fashion, of EEG-based arousal and valence recognition, the moving average filter demonstrated its promise in enhancing the performance of emotion classification and tracking. However, the issue of annotation curve distortion owing to smoothing approaches is a subject to be studied in the future work.

## Acknowledgment

This research is partially supported by the Center of Innovation Program from Japan Science and Technology Agency (JST), JSPS KAKENHI Grant Number 25540101, and the Management Expenses Grants for National Universities Corporations from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT).

## References

- Candra, H., Yuwono, M., Chai, R., Handojoseno, A., Elamvazuthi, I., Nguyen, H., Su, S.: Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In: Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 7250–7253 (2015)
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. ISCA Workshop on Speech and Emotion. pp. 19–24 (2000)
- Delorme, A., Mullen, T., Kothe, C., Acar, Z.A., Bigdely-Shamlo, N., Vankov, A., Makeig, S.: EEGLAB, SIFT, NFT, BCILAB, and ERICA: New tools for advanced EEG processing. Comp. Intell. Neurosci. 2011 (2011)
- Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image. Vision. Comput. 31(2), 120–136 (2013)
- Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D 31(2), 277–283 (1988)
- Kim, M.K., Kim, M., Oh, E., Kim, S.P.: A review on the computational methods for emotional state estimation from the human EEG. Comp. Math. Methods in Medicine 2013 (2013)
- Koelsch, S.: Brain correlates of music-evoked emotions. Nat. Rev. Neurosci. 15(3), 170–180 (2014)
- Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: DEAP: A database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. 3(1), 18–31 (2012)
- Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalogram dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. 8(94) (2014)
- Mariooryad, S., Busso, C.: Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: Proc. 5th Humaine Association Conference on Affective Computing and Intelligent Interaction. pp. 85–90 (2013)
- Metallinou, A., Narayanan, S.: Annotation and processing of continuous emotional attributes: Challenges and opportunities. In: Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. pp. 1–8 (2013)

- Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* 39(6), 1161–1178 (1980)
- Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8), 1627–1639 (1964)
- Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Trans. Affect. Comput.* 7(1), 17–28 (2016)
- Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* 3(1), 42–55 (2012)
- Sourina, O., Liu, Y., Nguyen, M.K.: Real-time EEG-based emotion recognition for music therapy. *J. Multimodal. User.* In. 5(1-2), 27–35 (2012)
- Sourina, O., Wang, Q., Liu, Y., Nguyen, M.K.: A real-time fractal-based brain state recognition from eeg and its applications. In: Babiloni, F., Fred, A.L.N., Filipe, J., Gamboa, H. (eds.) *Proc. BIOSIGNALS*. pp. 82–90 (2011)
- Thammasan, N., Moriyama, K., Fukui, K., Numao, M.: Continuous music-emotion recognition based on electroencephalogram. *IEICE Trans. Inform. Syst.* E99-D(4), 1234–1241 (2016)

## Modeling Negative Affect Detector of Novice Programming Students using Keyboard Dynamics and Mouse Behavior

Larry A. Vea<sup>1,2</sup>, Ma. Mercedes T. Rodrigo<sup>2</sup>

<sup>1</sup> Mapua Institute of Technology, Makati City, Philippines  
 (lavea@mapua.edu.ph)

<sup>2</sup> Ateneo de Manila University, Quezon City, Philippines  
 (mrodrigo@ateneo.edu)

**Abstract.** We developed affective models for detecting negative affective states, particularly boredom, confusion, and frustration, among novice programming students learning C++, using keyboard dynamics and/or mouse behavior. The keystroke dynamics are already sufficient to model negative affect detector. However, adding mouse behavior, specifically the distance it travelled along the x-axis, slightly improved the model's performance. The idle time and typing error are the most notable features that predominantly influence the detection of negative affect. The idle time has the greatest influence in detecting high and fair boredom, while typing error comes before the idle time for low boredom. Conversely, typing error has the highest influence in detecting high and fair confusion, while idle time comes before typing error for low confusion. Though typing error is also the primary indicator of high and fair frustrations, other features are still needed before it is acknowledged as such. Lastly, there is a very slim chance to detect low frustration.

**Keywords:** Affect • model • novice programmer • keyboard dynamics • mouse behavior.

### 1 Introduction

Affect is an observable expression of some emotional state [1,2,3]. It influences the ability of an individual to process information, to accurately understand and to absorb new knowledge [4].

In novice programmer studies, the negative affective states, particularly boredom and confusion, are negatively correlated with the student achievement while positive affect such as flow is positively correlated with achievement [5].

Affect detectors are built based on data acquired by sensors, human observations, or other peripherals. Several studies make use of keyboard dynamics as a source of affective data. These data include: typing speed, number of keystrokes, total time taken for typing, typing errors (the number of hits on the backspace key, delete key, or other unrelated keys), keyboard idleness [6,7], keystroke latency time (dwell time) and

keystroke duration time (flight time) between two-key (digraph or 2G) or three-key (trigraph or 3G) combinations [8,9]. These studies examined further how these keystroke data are related to a generally described as positive and negative affective states.

A few studies also examined how mouse movements are related to irritation, annoyance, reflectiveness [10], and boredom [11]. Other studies also make use of the combined keyboard and mouse data to examine how these are related to affective states in terms of valence and arousal [7].

There are only a few studies that detect the affective states of novice programmers [e.g. 12,13]. Also, there is no literature yet that uses the combined keyboard and mouse data to detect such states. This study hopes to contribute to the literature by building and validating a detector for negative affect of novice programming students using both the keyboard and the mouse data. We also attempt to answer the following research questions: (1) what are the notable features from keyboard dynamics and/or mouse behavior that help out in the recognition of negative affective states of novice programming students?; (2) how is student's affect related to keyboard dynamics and/or mouse behavior; (3) are the notable features "stable" or "consistent" over student's programming time period?; (4) how do these features differ or similar among high / medium / low incidences of boredom, confusion, and frustration?; and (5) what is the effect of combining mouse behavior with the keystroke dynamic features in predicting student's affect compared when using keystroke features alone or mouse features alone?.

This study hopes to contribute to the development of formal models of recognizing affective states of novice programmers, using the most common, low cost, non-intrusive computer devices such as the keyboard and the mouse. The discovered models or patterns to recognize negative affective states in this study may be used by computer scientists in developing computational systems that may automatically provide feedback to both teachers and students.

## 2 Related Works

Though there are different devices for affective states detection when using a computer, the keyboard and the mouse are the most commonly available, low-cost, and non-intrusive devices that could obtain affect indicators.

There were several studies that use only the keyboard as data source for affect detection. For example, Khanna et al [6] extracted keystroke features: typing speed, four statistics (mode, standard deviation, variance and range) from the number of typed characters for a defined time interval, total time taken for typing, number of backspace hits and idle times from recorded key logs to detect positive, negative, and neutral state of a computer user. These keystroke data were gathered from participants who were asked to retype some fixed texts in different time in order to acquire keystroke information under different affect states. The corresponding affect is collected by asking the participants to describe and report their affective state while doing the task. The resulting dataset was then analyzed through some data mining algorithms such as

SMO, MLP, and J48. They found out that the increase in the user typing speed relative to neutral state is an indicator of positive affect state while the decrease in the typing speed relative to neutral state is an indicator of negative affect.

An attempt to detect confusion and boredom states of novice programming students, Felipe et al [12] extracted the same keystroke features used by Khanna et al [6]. They also wanted to determine which of the extracted features could be indicators of the said affective states. The authors were permitted to collect video and key logs from students having programming activities. They reviewed every 20-second segment of the collected video logs and observe the student's behavior. They label affect by matching the corresponding observations from a checklist that describes affective states in terms of student's behavior. Results show that in a 20-second interval, keyboard inactivity in that time interval is the indicator of boredom state while confusion state was observed when the number of backspaces is greater than the idle time.

Tsui et al [9] also used key duration time (key press to key release) and key latency time (from one key release event to the next key press) features to examine the difference between positive and negative affect states. The keystroke data were collected by asking each participant to type a fixed number sequence with a pen on the mouth. The affect is labeled based on the teeth condition (positive) and the lip condition (negative) of the participant while typing. They found out that the duration time significantly show the difference between the two opposite states.

The features used by Bixler and D'mello [16] to discriminate between natural occurrences of boredom, engagement, and neutral states are divided into four keystroke and timing features: relative timing (session and essay timings), keystroke verbosity (number of keys and backspaces), keystroke timing (latency measures) and pausing behaviors. These features were extracted from the key logs of participants who were asked to write an essay about some selected topics using a computer. Likewise, the affect was labeled by asking the participant to view every 15-second segment of his video log and has to make self-judgment on what affective state was present in him during each time segment. Results show that when the identified keystroke and timing features were combined with task appraisal and stable traits features, it yields to a higher accuracy rate in classifying emotions, specifically, between boredom and engagement.

There were also studies that explored mouse as data sources in affect detection. For example, Tsoulouhas et al [11] extracted seven mouse movement features to detect emotional state, specifically boredom, of students who attend a lesson online. The said features are: total average movement speed, latest average movement speed, mouse inactivity occurrences, average duration of mouse inactivity, horizontal movements to total movements ratio, vertical movements to total movements ratio, diagonal movements to total movements' ratio, and the average movement speed per movement direction. They found out that the primary indicators of boredom are the average movement speed per movement direction and the mouse inactivity occurrences.

A more comprehensive study on affect detection in terms of its two dimensions was presented by Salmeron-Majadas et al [7]. They evaluated the keyboard and mouse affective data to identify participant's affective states in terms of valence and arousal. They combined some previously presented keyboard indicators such as the keystroke indicators used by Khanna [6] and Bixler and D'Mello [16], and the digraph and

trigraph used by Epp et al [8]. Their mouse indicators were generated from the participant's mouse clicks, cursor movements and scroll movements. These include: the number of button presses (left, right and both), overall distance, distance the cursor has been moved (covered distance) between two button press events, between a button press and the following button release event, between two button release events and between a button release and the following button press events, the Euclidean distance in the previous described cases, the difference between the covered and the Euclidean distance between the events described before, and the time elapsed between the mentioned events. After the participants finished the given task, they were asked to evaluate and score their affective state using the SAM scale. They computed the correlation between the extracted mouse/keyboard indicators and the reported affective states and found out that the mouse indicators that are correlated to the valence dimension of affect are: the mean time between two consecutive mouse button press events; the mean time between two consecutive mouse button release events; the standard deviation of the difference between the covered and the Euclidean distance between two consecutive mouse button press events; the standard deviation of the difference between the covered and the Euclidean distance between a mouse button release and the following mouse button press events; and the mean time between a mouse button release and the following mouse press button event; while the keyboard indicators are: the standard deviation of the time between two key press events; the mean duration of the digraph; the mean duration between the first key up and the next key down of the digraph; the duration between two key press events when grouped in digraphs; and the mean time between two key press events. On the other hand, the mouse indicators that identify the arousal dimension of affect are: the mean of the difference between the covered and the Euclidean distance between a mouse button release and the following mouse button press events; the mean of the difference between the covered; and the Euclidean distance between two consecutive mouse button press events; while the keyboard indicators are: number of keys pressed; the numbers of alphabetical characters pressed; the mean of the duration of the second key of the digraphs; the duration of the third key of the trigraphs; and the standard deviation of the duration of the digraph. Finally, they used these mouse and/or keyboard indicators in training some classifiers in order for them to know the prediction rates in recognizing positive and negative valence dimension of the participants. Results show that for some well-known classifiers such as C4.5 and Naïve Bayes, keyboard indicators alone provided the higher prediction rates than the mouse data alone, and even the combination of the data sources. However, for some more complex classifiers such as Random Forest and AdaBoost, the combined mouse and keyboard indicators provided the highest prediction rates among all the results.

Though there are some few studies on the affective states of novice programmers (e.g. [3],[5],[12],[13]), to date, there is no literature yet that uses the combined keyboard and mouse data to detect some negative affective states of these novices.

### 3 Methodology

#### 3.1 Participants

The participants in this study were 55 volunteers from first year students of a higher educational institution in Makati City. All of them were given waivers to parents or guardians, asking permission to let their child participate in the study. Hence, only those students with parent's/guardian's consent were allowed to participate.

At the time of the study, the students were enrolled in CS126 - Programming 1 with no or minimal background in C++. CS126 is a first year introduction to programming course using structured programming approach. Topics include: simple C++ syntax; program flow description; variables and data types; C++ operators; C++ control structures such as sequential, selection, and iterative structures; and functions.

#### 3.2 Data Collection Methods and Instruments

With the consent of the school, we used a customized mouse-key logger, web cam, the MS Movie Maker, and the Dev-C++ Integrated Development Environment.

Before the student works on its programming activity, the web cam is already properly in place and turned-on. The mouse-key logger and the Movie Maker were set and running in the background and hidden from the student in order not to bother him/her while he/she is doing the programming activity.

The mouse-key logger captured the mouse motion, mouse clicks, and mouse scrolls and the key event logs while the web cam captured the facial expressions and body movements of the student (video logs). The Dev-C++ was used as the programming environment in doing the programming activities.

Data was collected from the participants where the problem is about selection constructs and loop constructs, respectively. Data recording took almost three (3) hours.

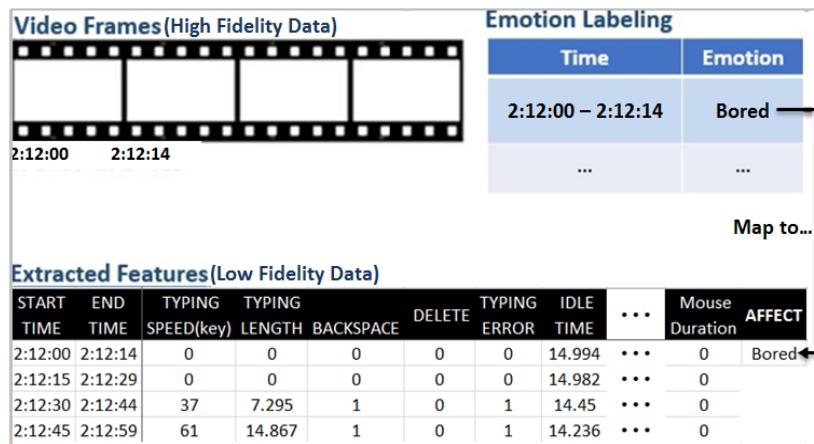
#### 3.3 Data Processing

We mapped the mouse-key logs with the video logs in several steps: We first cleaned the data by removing segments in the mouse-key logs that had no corresponding video logs; then we extracted potential keystroke and mouse dynamic features identified in some previous works, plus other features that may influence affect detection, from the mouse-key logs. The result was a comma separated value (csv) file containing the keyboard and mouse dynamic features at every 15-second interval. This file was called the "incomplete dataset" since the affect labels were not yet attached. We also divided the video logs into 15-second video time segments that corresponded to mouse-key time segments in the incomplete dataset. Then, affect labeling on each video segment was done by three trained labelers, one was a graduate student serving as lead and the other two were college seniors with strong background in computer programming. They watched the video together and came to a consensus regarding the student's affective state based on the coding scheme in Table 1. If there were disagreements, they played the segment until they agreed. Video segments where the participants showed

curiosities about being monitored through the camera or not seen in the video were marked “X”. Finally, we mapped each label of the video segment in the incomplete dataset (Fig. 1) , and the instances labeled with “X” were deleted.

**Table 1.** Affective state criteria

Affective States	Description
Boredom	Slouching and resting the chin in his/her palm; Yawning; Zoned out within the software; Looks uninterested/ unfocused; Barely uses the mouse /keyboard; Slouching; Eyes wandering
Confusion	Scratching his/her head; Repeatedly looking at the same interface elements consulting with a classmate or a teacher; Flipping through lecture slides or note; Statements such as “Why didn’t it work?”; Still engage with the software; Cannot grasp/experiencing difficulty with the material; On-task conversation; Pouts /Frowns / wrinkles brows/forehead; Nail biting; Lip biting; Lip slightly ajar
Frustration	Banging on the keyboard or pulling at his/her hair; cursing; statements such as “What’s going on?!”; Scratching the back of his head.; Rubbing his neck from behind; Scratching any part from his body; Changing his sitting position; Lips pulled inward; Raising the arms lifts sometimes up (or two arms- like throwing something in the air); Deep breath.



**Fig. 1.** Mapping of high fidelity data with the low fidelity data.

Determining of student’s affect from the video segment was based on the modified coding scheme adopted from [3],[5],[14] and is presented in Table 1. The scheme was modified to find the state of confusion (negative valence, positive arousal), boredom

(negative valence, negative arousal), frustrated, and a special affective state labeled as “others” [3],[6] in which the emotion with respect to the time frame was found to be neither confused, bored, nor frustrated.

The resulting complete dataset was then further divided into training and test set. Every fifth participant from the list was chosen as part of the test set while the rest were part of the training set.

### 3.4 Model Development and Data Analysis Methods

We used these datasets to develop several affective models for detecting confusion, frustration and boredom by training some well-known tree classifiers that could handle datasets with nominal class such J48, Decision Tree, and Random Forest using RapidMiner. Each classifier were trained and validated using different feature set, such as: keystroke verbosity features alone (KV); keystroke time duration and latency features of the digraph and trigraph alone (KT); all keystroke features - the combined verbosity, time duration and latency features (KF); mouse features alone (MF); and, the combined all keystroke and mouse features (KM). The gini index attribute criterion was used for feature selection and batch-X-validation to validate the model. The depth of the tree in each tree classifier was also explored in order to determine the model that has the highest performance in terms of accuracy rate and/or kappa statistic.

It was observed during the experiment that using keystroke time duration and latency features on the digraph and trigraph alone (KT), as well as mouse features alone (MF) do not provide a good model to detect negative affect since the kappa statistic is very low (less than 0.2) which implies a slight agreement [15]. It was also observed that the decision tree classifier consistently provide the highest kappa statistic and accuracy rate. It also implies that decision tree classifier gave the most acceptable model.

**Table 2.** Model performance using Decision tree classifier

<b>Feature Set</b>	<b>Training Phase</b>		<b>Testing Phase</b>	
	<b>Kappa</b>	<b>Accuracy (%)</b>	<b>Kappa</b>	<b>Accuracy (%)</b>
KV	0.493	70.80	0.564	74.08
KF	0.489	71.03	0.568	74.28
KM	0.490	71.06	0.567	74.23

Lastly, the kappa and accuracy of the other feature sets are statistically tied (Table 2, columns 2 and 3). And since the kappa is in moderate agreement [15], it implies that these feature sets can be used to model negative affect detector. The models generated by the decision tree classifier for the said feature sets were tested using a pre-labeled test set for further investigation. The result of the tests is also presented in Table 2, columns 4 and 5. The table shows that the kappa and the accuracy significantly increased but are still statistically tied. This confirms that the three (3) feature sets can be used to model negative affect detectors of novice programming students.

The tree models were further analyzed to find the significant features that help out in the recognition of negative affective states of novice programming students and how these features are related to student's affect. This was done by listing the unique inner nodes of the decision tree models generated by the classifier.

Using correlations in RapidMiner, it was observed that some of the notable features in the tree are strongly correlated. For example: typing error is highly correlated with backspace; total keyevents is also highly correlated with typing speeds and total time for typing; the sum of all time durations the student acted on the 1st key of the digraph (SUM\_2G\_1Dur) is fairly correlated with the maximum value in the set of the durations of the 1st key of the trigraph (MAX\_3G\_1Dur); and the total distance travelled by the mouse along the x-axis (MM\_Total\_X) is highly correlated with mouse activity duration. Thus, to achieve a more parsimonious model, we tried iteratively removing some features that are highly correlated to other features. Results show that the kappa and accuracy slightly improved (see Table 3). The table shows that kappa and accuracy in all the feature sets are almost equal. It implies that the notable features from the keystroke verbosity feature set alone (KV) or the combined verbosity, duration, and latency keystroke features (KF) are already enough to model a negative affect detector of novice C++ programming students. However, adding MM\_Total\_X (total distance travelled by the mouse along the x-axis) mouse feature with the keyboard features (KF) slightly improved the recognition rate of the model (Table 3).

**Table 3.** Model performance when some features correlated to other features were removed.

Feature Set	Kappa		Accuracy (%)		Remaining Significant Features
	Before Removal	After Removal	Before Removal	After Removal	
KV	0.564	0.569	74.08	74.23	Typing Error, Typing Variance, Idle Time, Total Key Events, F9
KF	0.568	0.568	74.28	74.28	Typing Error, Typing Variance, Idle Time, Total Key Events, F9, MAX_3G_1Dur, AVE_3G_2D3D, and SUM_2G_1Dur
KM	0.567	<b>0.572</b>	74.23	<b>74.37</b>	Typing Error, Typing Variance, Idle Time, Total Key Events, F9, SUM_2G_1Dur, AVE_3G_2D3D, and MM_Total_X

To specifically determine how student's affect related to keyboard and mouse dynamics, the unique paths from the root of the decision tree of the KM feature set, to the its leaves were analyzed and then transformed into rules. The result is shown in Tables 4.

**Table 4.** How student affect related to keystroke dynamics and mouse behaviors.

Affect (most likely)	Pattern (based on a 15-second observation)
Boredom	IF ((Typing error $\leq 0.50$ ) and (Idle time $> 14.98$ ) and (Total keyevents $\leq 0.50$ ) and (MM_Total_X $\leq 243.50$ ));
Frustration	IF ((Typing error $\leq 0.50$ ) and (Idle time $> 14.98$ ) and (Total Keyevents $> 4.50$ )).
Confusion	IF (Typing error $> 3.50$ ); OR IF ((1.50 $<$ Typing error $\leq 3.50$ ) and (Typing variance $> 0.815$ ) (MM_Total_X $> 17383.50$ )); OR IF ((Typing variance $> 0.815$ ) and (Typing error $\leq 1.50$ ) and (SUM_2G_1Dur $\leq 0.83$ )); OR IF((1.5 $\leq$ Typing error $\leq 3.50$ ) and (Typing variance $\leq 0.815$ )); OR IF ((Typing variance $\leq 0.815$ ) and (Typing error $\leq 1.50$ ) and (AVE_3G_2D3D $\leq 5.170$ )); OR IF ((Typing error $\leq 0.50$ ) and (Idle time $> 14.98$ ) and (Total Keyevents $\leq 0.50$ ) and (MM_Total_X $> 850$ )); OR IF ((Typing error $\leq 0.50$ ) and (Idle time $\leq 14.98$ ) and (Total Keyevents $> 9.50$ ) and (SUM_2G_1Dur $> 0.924$ ) and (F9 $> 27$ )); OR IF ((Typing error $\leq 0.50$ ) and (Idle time $\leq 13.636$ ) and (Total Keyevents $\leq 9.50$ ) and (MM_Total_X $> 872956.5$ )).

We examined if the features were stable over time since it is possible that student keyboard and mouse dynamics change as the student develops and completes a program. Also, a student may type more in the beginning of the development process, when he is still writing code, and less so when he is debugging. We therefore divided the dataset into the first 1/3, the second 1/3, and the last 1/3 of the observation period and re-processed each subsets.

Results show that when students are just starting with their programming activity (first 1/3 of the period), the most notable features that determines student's negative affect in all feature sets are the typing error, and idle time. Though typing error and the idle time are also the dominant features on the second 1/3 of the period, other keystroke verbosity features such as typing variance, total keyevents, and the number of times the student presses F9 (shortcut to compile and run the program) are included. Also, adding the average duration time of the first key in the trigraph (AVE\_3G\_1Dur) or the total distance travelled by the mouse along the x-axis (MM\_Total\_X) improves the recognition rate. Lastly, at the time the programming period is almost toward its end (last 1/3 of the period), the typing error and idle time are still the dominant features, but adding the typing variance and total mouse movement along the x-axis increases student's negative affect detection. It was also observed that the total keyevents and the

average duration time of the first key in the trigraph (AVE\_3G\_1Dur) that represent the movements of the keys, including F9 which represents running the program were gone towards the end of the programming period. This may indicate that there were only few monitored keyboard activities. Probably, some of the students may have stopped working; either they are already finished with the activity or they have abandoned their work.

Finally, to determine how the notable features differ or similar among high/medium/low incidences of boredom, confusion, and frustration, the original dataset was divided into other subsets by computing the percentage of the time each student was observed to be bored, confused or frustrated and then segregate the data into the top 1/3 of those who are bored, confused or frustrated, the middle 1/3, and the lowest 1/3, and then re-process each subsets. The result is shown in Table 5.

**Table 5.** Differences or similarities among high/medium/low incidences of boredom, confusion, and frustration.

	Affect	Keystroke Dynamic Features	Combined Mouse and Keystroke Dynamics Features
Boredom	High	Idle time, Total keyevents	Idle time, MouseActivityDuration
	Fair	Idle time, Total keyevents	Idle time, Total keyevents
	Low	Typing error, Idle time	Typing error, Idle time, Numbers, MouseActivityDuration
Confusion	High	Typing error	Typing error
	Fair	Typing error, Typing speed(char)	Typing error, Typing speed(char)
	Low	Idle time, Typing error	Idle time, Typing error
Frustration	High	Typing error, Typing variance, Idle time, Control	Typing error, Typing variance, Idle time, Control
	Fair	Typing error, Idle time, Keypress_DeltaX, F9, etc.	Typing error, Idle time, CMM_time, Keypress, etc.
	Low	(no sign of low frustration)	(the tree is very deep where there is a very minimal sign of frustration)

#### 4 Conclusion

This study was conducted to address the following research questions: (1) what are the notable features from keyboard dynamics and/or mouse behaviors that help out in the recognition of negative affective states of novice programming students?; (2) how is student's affect related to keyboard dynamics and/or mouse behaviors; (3) are the notable features "stable" or "consistent" over student's programming time period?; (4) how do these features differ or similar among high / medium / low incidences of boredom, confusion, and frustration?; and (5) what is the effect of combining mouse

features with the keystroke features in predicting student's affect compared when using keystroke dynamic features alone or mouse behaviors alone?. These questions are answered as follows:

- (1) The notable features from keyboard dynamics and/or mouse behaviors that help out in the recognition of negative affective states of novice programming students are presented in Table 3. These include: the student's typing errors incurred (the number times the backspace and delete keys were pressed); the length of time the student is idle (not pressing any key in the keyboard); the student's typing variance (his/her typing varies with time); the number of key events (keydown+keypress+keyup) he/she executed in the keyboard; total distance the student moved the mouse along the x-axis (MM\_Total\_X); the sum of all time durations the student acted on the 1st key of the digraph (SUM\_2G\_1Dur); the average time duration between the 2nd and 3rd keydown of the trigraph (AVE\_3G\_2D3D); and, the number of times F9 key (shortcut to compile and run the program) was pressed.
- (2) As shown in Table 4, student's boredom is related to both keystroke dynamics and mouse behavior. The keyboard has almost no activity while the mouse has a very minimal movement along the x-axis. On the other hand, student's frustration is similar to boredom, except for the mouse features, since for this affect, students tend to release the mouse and scratch their head or do some other hand gestures. There is almost no keyboard activity too since when a student get frustrated, he/she usually pause for a while and do nothing. Lastly, student's confusion is both related to keystroke dynamics and mouse behavior. The table shows that there are several indicators when a student is confused.
- (3) After analyzing the data at first 1/3, the second 1/3, and the last 1/3 of the observation period, it was observed that the features are not stable since there are many features needed in detecting negative affect at the middle (second 1/3) of the observation period. It was also observed that the typing error has the greatest influence during the first 1/3 and second 1/3 of the observation period followed by the idle time, while the idle time has the greatest influence during the last 1/3 followed by the typing error.
- (4) Table 5 shows that idle time has the greatest influence in detecting high and fair boredom but it is just secondary with the typing error for low boredom. On the other hand, typing error has the greatest influence in detecting high and fair confusion but it is just secondary with the idle time for low confusion. Though typing error is also the primary indicator of high and fair frustrations, it requires other features before it is acknowledged as such.

As shown in the last row of Table 3, adding a mouse feature, particularly with the distance it travelled along the x-axis, with the keystroke features improve the detection of student's affect compared when using keystroke dynamic features alone or mouse behaviors alone.

## 5 References

1. Affect. Encyclopedia of Mental Disorders. Retrieved from <http://www.minddisorders.com/A-Br/Affect.html>.
2. Psychiatry Clerkship. Retrieved from <http://depts.washington.edu/psychclerk/glossary.html>.
3. Carlos, C. M., Delos Santos, J. E., Fournier, G. and Vea, L. (2013, March). Towards the Development of an Intelligent Agent for Novice Programmers through Face Expression Recognition. In *Proceedings of the 13th Philippine Computing Science Congress*, 101-106.
4. Picard R.W. and the Medial Lab – Affective Computing Group. Affective computing. Retrieved from <http://affect.media.mit.edu/>.
5. Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., ... & Tabanao, E. S. (2009, July). Affective and behavioral predictors of novice programmer achievement. In *ITiCSE '09 Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education* 41(3), 156-160. doi: <http://doi.acm.org/10.1145/1562877.1562929>.
6. Khanna, P., & Sasikumar, M. (2010). Recognising emotions from keyboard stroke pattern. *International journal of computer applications*, 11(9), 1-5.
7. Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2014, July). Exploring indicators from keyboard and mouse interactions to predict the user affective state. In *Educational Data Mining 2014*.
8. Epp, C., Lippold, M., & Mandryk, R. L. (2011, May). Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 715-724. doi: <http://doi.acm.org/10.1145/1978942.1979046>.
9. Tsui, W. H., Lee, P., & Hsiao, T. C. (2013, July). The effect of emotion on keystroke: an experimental study using facial feedback hypothesis. *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2870-2873. <http://dx.doi.org/10.1109/EMBC.2013.6610139>.
10. Schuller, B., Rigoll, G., & Lang, M. (2004, June). Emotion recognition in the manual interaction with graphical user interfaces. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME'04)*, 2, 1215-1218. doi: <http://dx.doi.org/10.1109/ICME.2004.1394440>.
11. Tsoulouhas, G., Georgiou, D., & Karakos, A. (2011). Detection of learner's affective state based on mouse movements. *Journal of Computing* 3(11), 9-18.
12. Felipe, D. A. M., Gutierrez, K. I. N., Quiros, E. C. M., & Vea, L. A. (2012). Towards the Development of Intelligent Agent for Novice C/C++ Programmers through Affective Analysis of Event Logs. In *Proceedings of the International MultiConference of Engineers and Computer Scientists 1*, 511-518.
13. Lee, D. (2011). Detecting Confusion Among Novice Programmers Using BlueJ Compile Logs. *Master's thesis, Ateneo de Manila University, Quezon City, Philippines*.
14. Dragon, T., Arroyo, I., Woolf, B. P., Burleson, W., El Kalioubi, R., & Eydgahi, H. (2008, January). Viewing student affect and learning through classroom observation and physical sensors. In *Proceedings of the 9th international conference on Intelligent tutoring systems*, 29-39. Springer Berlin Heidelberg.
15. Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5), 360-363.
16. Bixler, R. & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. *Proceedings of the International Conference on Intelligent User Interfaces* (2013), 225-234.

## Multimodal Latent Feature Learning for Psycho-Physiological Stress Modelling and Detection

Juan Lorenzo Hagad, Kenichi Fukui, and Masayuki Numao

Department of Architecture for Intelligence,  
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan  
[{hagad,fukui,numao}@ai.sanken.osaka-u.ac.jp](mailto:{hagad,fukui,numao}@ai.sanken.osaka-u.ac.jp)

**Abstract.** Together with mobile computing and wearable medical devices, artificial intelligence is poised to play an important role in the field of mental health and stress management. Along with a number of other advancements, using data from different data modalities has been shown to be an effective way of building accurate and flexible stress models. However, recent findings indicate that traditional machine learning techniques may lack the ability to effectively identify salient inter-modal correlations, especially when seemingly unrelated modalities are used concurrently. To examine this effect, in this work we investigated the efficacy of building multimodal stress models using a combination of psychological and physiological data. A monitoring platform was built and unobtrusive wearable sensors were used to gather data from subjects engaged in authentic work activities. The final models were created by combining psychological data from stress coping profiles and physiological signals from the wearable sensors. Finally, self-annotated stress annotations to establish the ground truths used for model training. A performance comparison was made between standard machine learning approaches and unsupervised latent feature learning, including deep learning architectures. The results indicate that significant improvements can be achieved by applying deep multimodal feature learning to construct mental stress models.

**Keywords:** Stress Detection, Wearable Physiological Sensors, Stress Coping, Latent Feature Learning, Multimodal Deep Learning

### 1 Introduction

The past decade has witnessed the emergence of a number of technologies for the automated analysis of stress. Formally, stress can be defined as a biochemical or physiological change in response to internal and external stressors. It is recognized in clinical literature [3] as a risk factor for a number of cardio-vascular diseases and is one of the leading causes of work disabilities worldwide. Due to its pervasiveness, the field of automated mental stress monitoring and diagnosis has gained widespread interest and is poised to introduce key technologies for addressing severe mental health issues such as depression.

### 1.1 Multimodal Stress Models

One of the most effective means of improving model performance is the use of multimodal data. This multimodal modelling involves merging data from different data channels into a single model. Technologies for stress monitoring in particular have experienced a boost due to the emergence of modern sensing technologies. This study focuses primarily on using wearable sensors for physiological signal measuring. Among the commonly used physiological signals for stress detection, the most effectively applied are galvanic skin response (GSR)[7, 6, 10] and heart rate (HR)[7, 6, 10, 4].

In [10] they used galvanic skin response, blood volume pulse, pupil diameter and skin temperature to detect changes in stress levels. Models were trained to distinguish between two states: stressed and relaxed. The aforementioned work achieved over 90% accuracy using support vector machines (SVM) tweaked for personalization. Other studies such as [7] followed a similar data gathering framework. In [7] they performed out-of-laboratory experiments using the Intel Shimmer platform, a sensor platform which includes ECG, GSR and accelerometers. In [7] stress was induced using the Stroop test and mental arithmetic tests. With the aid of accelerometer data, it was shown that activity context could be extracted and could lead to overall improvements in detection performance.

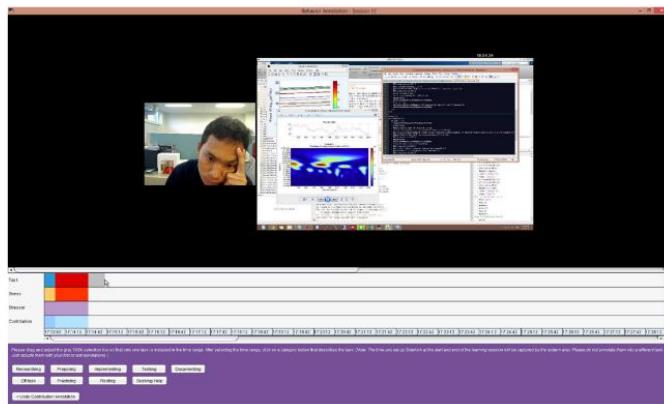
A common feature among these works is the use of traditional machine learning methods such as SVM and neural networks (NN). Furthermore, they are usually trained using shallow concatenations of multimodal data. In this work, we aim to test whether this simplistic approach has any major shortcomings, and then we propose a more appropriate method. In addition, we address observations that existing studies tend to rely on data from artificially induced stress, employing methods such as the Stroop test. While these are designed to simulate authentic stress conditions, they have a tendency towards exaggerated stress responses. As a result, the resulting models may not properly represent the full spectrum of stress responses that may be encountered in real-world scenarios. On the other hand, using naturalistic data poses its own challenges since samples from natural environments are susceptible to noise, inconsistencies, and have a tendency to feature more ambiguous expressions of stress. Furthermore, there is the challenge of gathering consistent ground truth labels from different subjects. Yet, it is necessary to investigate such models in order to build an appropriate model of real-world stress features.

To offset the effects of noise and other user variabilities, we employed latent feature learning over multiple modalities. Furthermore, we investigated various deep learning structures to discover the best structure for learning effective multimodal features. Finally, we approach all of this while attempting to merge data from psychology and physiology. Specifically, we combined coping profiles with wearable sensor signals. This was inspired by evidence from medicine and psychology that show that the stress response is not simply a function of the severity of the stressor, but is also a function of the ability of the organism to cope [9]. Thus, individual coping characteristics may be used as factors that affect the stress response.

## 2 Methods

### 2.1 Data and Annotations

All experiments were performed on computers equipped with our purpose-built monitoring and annotation software, as shown in Figure 1. Each subject conducted self-regulated work activities on their personal PCs while using the software. This allowed the experimenter to record the subjects personal profiles and work activities without providing direct supervision. Since most experimental procedures involved minimal interaction between the experimenter and the subjects, transmission of experimenter biases was also minimized.



**Fig. 1.** A screenshot of the annotation module.

(See Figure 1)

For the ground truths, self-reported stress annotations were made immediately following the hour-long work sessions. Using webcam and desktop recordings, subjects reviewed and selected time segments in the recorded video based on their performed tasks. For each task, they identified the amount of stress felt at the time. A selection of work task categories and stressors was provided through the UI, as well as a stress annotation slider using a 4-point Likert-scale (1=very low, 4=very high).

### 2.2 Physiological Signals

Participant physiological signals were measured using wearable ECG sensors and wrist sensors. These wireless, wearable devices allowed continuous measurement of heart rate (HR) and skin conductance (SC), respectively.

Heart rate variability (HRV) features were extracted from the HR data. These are HR features that have been shown to have strong correlations with

autonomic stress responses [4]. Specifically, the following features were used: Average of NN intervals (AVNN), Standard deviation of NN intervals (SDNN), root-mean-squared differences between adjacent NN intervals (rMSSD), percentage of differences between adjacent NN intervals greater than 50ms (pNN50), spectral power measures of NN intervals of varying frequencies (ULF, VLF, LF, HF) and the ratio of low to high frequency power (LF/HF).

For GSR, two major components of the conductance signal were analyzed: skin conductance level (SCL) and the skin conductance response (SCR) [1]. These features cover different aspects of sympathetic neuronal activity. SCL reflects the *tonic* level or the slowly changing component of the GSR signal, while SCR or rapid *phasic* refers to the faster changing elements of the signal. These measurements were obtained by using software included with the wearable devices. The values were then cleaned and time-dependent statistics were extracted such as the mean, variance, and difference from the baseline levels.

### 2.3 Psychological Profiles

To build the personal profile, subjects answered the COPE Inventory [2], a questionnaire designed to assess coping responses in response to stressful situations. It determines a person's inclination towards exhibiting responses that are expected to be either functional or dysfunctional. Those with dysfunctional coping mechanisms are expected to be more prone to the negative effects of stress. By including these factors into our machine-learned models, we hope to be able to reduce the ambiguity of stress features relative to each subject.

### 2.4 Baseline Machine-Learners

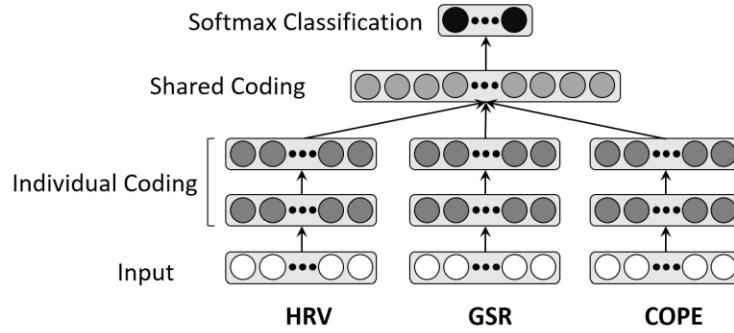
Supervised and unsupervised machine learning techniques were used to build baseline stress models to compare with the deep learning models. These were selected from machine learning techniques most commonly featured in related works. For the supervised models, we used support vector machines (SVM) and multilayer perceptrons (MLP), and for the unsupervised model we used k-means clustering. The SVMs featured used radial basis function (RBF) kernels since these have been shown to be highly flexible for classification tasks. The MLPs featured a single hidden layer with a number of nodes equal to half of the sum of the number of input attributes plus the number of output classes. Finally, the number of k-means clusters were adjusted to match the number of output classes.

### 2.5 Deep Feature Learning

In this work we implemented Deep Learning using Autoencoders [8]. These artificial neural network structures are similar to traditional feed-forward networks. In its most basic form it includes input, hidden, and output layers. Unlike typical neural networks, autoencoders learn the ideal parameters to generate an output

that is a reconstruction of the inputs. Through this process, the hidden layers are able to discover latent features that can efficiently represent the training data. Specifically, we applied Denoising Autoencoders (DA) a form of autoencoder that is trained by reconstructing using stochastically corrupted versions of the input.

The deep learning structures in this work used autoencoders stacked in a greedy layerwise fashion to form a deep network similar to stacked Restricted Boltzmann Machines (RBM) in deep belief networks [8]. The different levels of the stack allow learning multiple levels of abstraction and can be used to learn inter-modal features. However, strong intra-modal feature correlations may still prevent the discovery of some important inter-modal features.



**Fig. 2.** Multimodal Deep Learning Structure

In [5], a study was performed to measure the saliency of multimodal features discovered by different deep RBM structures. It was discovered that models trained on shallow concatenations of audio and video features were not able to capture effective correlations across the modalities. Bimodal deep RBMs, those that featured pretraining at different modal levels, were much better able to capture these correlations. In this work, we apply the same concept, however using stacked autoencoders to build the structure shown in Figure 2.

### 3 Experiments and Results

Data was collected from 4 healthy male participants aged 20-32. All subjects were graduate school students from Osaka University. After agreeing to participate in the study, they answered a number of psychology-related questionnaires, including the COPE Inventory. The individual questionnaire answers from COPE inventory were used as psychological feature data. Each subject performed at least 5 work sessions with each session lasting 1 hour. Participants were urged to

perform work activities during this time, however no explicit activity restrictions were applied.

These annotations resulted in 184 usable work task segments labelled with stress levels from 1 (very low stress) to 4 (very high stress). Each task segment lasted around 5 to 30 minutes. HR was recorded at a sampling frequency of 256Hz while SC was recorded at 128Hz. All frequency-domain and time-domain features were extracted over task segments. These segments were aligned based on the participants' task segment annotations. Data synchronization was performed using time-stamp data marked at the start of each session.

### 3.1 Baseline Results

The baseline performance for the standard machine models are shown in Table 1. Classification performance was measured using stratified 10-fold cross-validation accuracy. Based on these results, all models performed better than random classification with MLPs showing the best performance.

**Table 1.** Baseline Performance Results

SVM	MLP	K-means	Random
42.069%	47.414%	40.702%	29.501%

### 3.2 Denoising Autoencoder Results

For the first round of experiments, we built and tested models using Denoising Autoencoders (DA). For the following experiments we used DAs with 5x overcomplete hidden units for the combination of physiological features (105 units) connected to a logistic regression layer. Pretraining was performed over 100 epochs and with a 0.001 learning rate. The per-fold results are listed in Table 2. When comparing the performance of DAs to MLPs, we noted a rise in mean accuracy from 47.90% to 50.53%. However, statistical analysis via a paired t-test indicated that this was not sufficiently significant ( $p=0.16$ ).

### 3.3 Stacked Autoencoder Results

In the next investigation, we attempted to discover features through a deep structure. We built and tested Stacked Denoising Autoencoders (SDA) and compared their performance with the previous DAs on the 4-class dataset. For the SDA we used 3 hidden layers based on results cited in [8]. Once again, we used 5x overcomplete hidden units (105 units) for each of the hidden layers. Pretraining was performed using 100 epochs and at a 0.001 learning rate.

**Table 2.** Comparison of Performance Results

<b>Model Fold</b>	<b>MLP</b>	<b>DA</b>	<b>SDA</b>	<b>MDL</b>
1	47.37%	47.37%	47.37%	57.89%
2	47.37%	52.63%	47.37%	57.89%
3	47.37%	52.63%	47.37%	52.63%
4	47.37%	47.37%	47.37%	52.63%
5	47.37%	52.63%	47.37%	57.89%
6	47.37%	47.37%	47.37%	52.63%
7	47.37%	47.37%	57.89%	52.63%
8	52.63%	52.63%	47.37%	52.63%
9	47.37%	52.63%	47.37%	52.63%
10	47.37%	52.63%	52.63%	52.63%
<b>Mean</b>	<b>47.89%</b>	<b>50.53%</b>	<b>48.95%</b>	<b>54.21%</b>
<b>Variance</b>	2.77%	7.39%	12.62%	6.46%

Referring to the results in Table 2 and comparing the results of the single layer DA and the 3-layer SDA, there is a slight reduction in performance. Intuition states that adding more layers may eventually lead to improvements, so to confirm that these were optimal results we also tested models with additional layers. As shown by the pattern of performance in Figure 3, deviating from 3 layers actually leads to similar or worse performance. Error bars indicate standard error over 10-fold cross-validation accuracy. Basically, results indicate that using 1 or 2 layers leads to a high variation in performance, while adding more layers beyond 3 leads to a dramatic decrease in performance. This coincides with previous results found by [8] in their investigation with deep RBM's and could be an indication of the increased training data requirements of deeper models.

### 3.4 Multimodal Deep Learning Results

For the final experiment we modified our approach by applying multimodal deep learning (MDL) [5]. Each modality was pre-trained as its own isolated denoising autoencoder. This allowed the discovery of unimodal latent features. On top of these, we placed a fully-connected autoencoder layer intended to learn a shared coding for the inter-modal features. Finally, the last layer was a softmax logistic regression layer used for supervised learning and classification.

When comparing the performance of the final model with the previous attempts, there is a noticeable improvement. Based on Table 2, the MDL model achieved a 54% accuracy, which is an additional 5.26% compared to the SDA

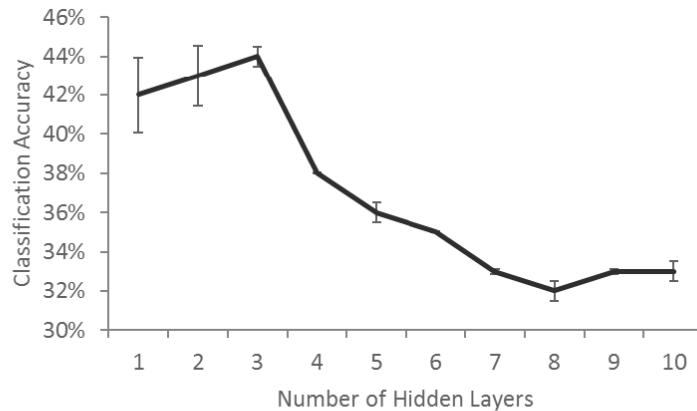


Fig. 3. Stack Depth Test Results

using only concatenation of features (48.95%). This time results were statistically significant with  $p=0.008$  at  $\alpha=0.05$ . Furthermore, we noticed a reduced variability with regards to how the model performed on the different folds.

#### 4 Discussion

In the first experiment we compared single-layer DAs with single-layer MLPs in order to assess the effects of latent feature discovery. Focusing on the mean accuracy, we noted an increase of 2.63% after using DAs. However, statistical analysis revealed that the improvements were not significant. A possible explanation is that although latent features were discovered, they lacked sufficient discriminative ability due to the model not being able to learn inter-modal features due to the shallow structure. For the next investigation we attempted to use a deeper structure.

The outcome for the SDA experiment was surprising since deep learning is usually expected to lead to improvements, however based on these results it was instead detrimental. Statistical testing shows that this was not significant ( $p=0.19$ ), although it still meant that there was no observable advantage to simply applying a deeper model. It was apparent that the problem was with how the data fusion was handled. In the succeeding experiment, we corrected the approach by applying MDL.

Based on the results in Table 2, significant performance were made by applying a MDL strategy. The final MDL model achieved a 54% accuracy, an additional 5.26% compared to the SDA, and a statistically significant improvement with  $p=0.008$  at  $\alpha=0.05$ . In addition, there was less variability with regards to how the model performed on the different folds. These results, indicate that a

deep multimodal learning approach is an effective modelling strategy for classifying multimodal stress data.

## 5 Summary and Conclusion

In summary, this work presented an improved method for building mental stress models using multimodal data and multimodal deep learning. By using a monitoring platform and unobtrusive wearable sensors, data was gathered from subjects engaged in authentic work activities. Psychology-based annotation tools collected stress-related context while wearable sensors tracked physiological signals of heart rate and skin conductance. Then, different structural combinations of autoencoders were tested to discover which could most effectively discover latent features in the physiological and psychological data. Specifically, single-layer denoising autoencoders (DA), a 3-layer stacked denoising autoencoders(SDA), and an SDA with a multimodal deep learning scheme(MDL) were tested. Based on the results, all models performed better than the baseline traditional machine learning approaches. The most significant gains were achieved by applying the MDL on the 3-channel data. On the other hand, even deep models using simple feature-level concatenation (i.e., the SDA) experienced a slight performance loss even when compared to single layer DAs using the same concatenated data. These results support findings from previous works in multimodal learning that showed that combinations of certain modalities require separate feature learning phases to discover unimodal features and multimodal features. Furthermore, all models performed better than random despite using naturalistic work activity data without artificially injected stressors. The most significant performance gains were achieved by applying a multimodal deep learning strategy compared to all other tested approaches. These results show that deep multimodal learning is an effective method of building psycho-physiological stress models. On the other hand, the performance is not yet sufficient for practical applications. This is likely a result of the relatively small dataset that was used for this investigation. Further data gathering and dataset validation may be necessary to build more complete and balanced models featuring the full spectrum of stress responses. Furthermore, explicit noise reduction methods need to be explored, as well as a more exhaustive investigation of the possible machine learning structures. In future work, we also plan to evaluate the generalization performance of the models.

## References

1. Braithwaite, J.J., Watson, D.G., Jones, R., Rowe, M.: A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology* 49, 1017–1034 (2013)
2. Carver, C.S., Scheier, M.F., Weintraub, J.K.: Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology* 56(2), 267–83 (1989)

3. Iso, H., Date, C., Yamamoto, A., Toyoshima, H., Tanabe, N., Kikuchi, S., Kondo, T., Watanabe, Y., Wada, Y., Ishibashi, T., Suzuki, H., Koizumi, A., Inaba, Y., Tamakoshi, A., Ohno, Y.: Perceived mental stress and mortality from cardiovascular disease among Japanese men and women: the Japan Collaborative Cohort Study for Evaluation of Cancer Risk Sponsored by Monbusho (JACC Study). *Circulation* 106(10), 1229–1236 (Sep 2002)
4. Melillo, P., Bracale, M., Pecchia, L.: Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *BioMedical Engineering OnLine* 10, 96+ (Nov 2011), <http://dx.doi.org/10.1186/1475-925X-10-96>
5. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: International Conference on Machine Learning (ICML). Bellevue, USA (June 2011)
6. Shi, Y., Nguyen, M.H., Blitz, P., French, B., Fisk, S., De la Torre, F., Smailagic, A., Siewiorek, D.P., alAbsi, M., Ertin, E., et al.: Personalized stress detection from physiological measurements. International symposium on quality of life technology pp. 28–29 (2010)
7. Sun, F.T., Kuo, C., Cheng, H.T., Buthpitiya, S., Collins, P., Griss, M.: Activity-aware mental stress detection using physiological sensors. In: Mobile computing, applications, and services, pp. 211–230. Springer Berlin Heidelberg (2012)
8. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, 3371–3408 (Dec 2010), <http://dl.acm.org/citation.cfm?id=1756006.1953039>
9. Vogel, W.H.: Coping, stress, stressors and health consequences. *Neuropsychobiology* 13(3), 129–135 (1985)
10. Zhai, J., Barreto, A.: Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE* pp. 1355–1358 (Aug 2006), <http://dx.doi.org/10.1109/embc.2006.259421>

## Affective Laughter Expressions from Body Movements

Ma. Beatrice L. Luz, McAnjelo D. Nocum, Timothy Jasper T. Purganan,  
Wing San T. Wong, Jocelynn W. Cu

Center for Human Computing Innovations, De La Salle University, Philippines  
{ma\_beatrice\_luz, mcanjelo\_nocum, timothy\_purganan,  
wing\_wong, jocelynn\_cu}@dlsu.edu.ph}

**Abstract.** The main goal of this study is to classify affective laughter expressions from body movements. Using a non-intrusive Kinect sensor, body movement data from laughing participants were collected, annotated and segmented. A set of features that include the head, torso, shoulder movements, as well as the positions of the right and left hands, were used by a decision tree classifier to determine the type of emotions expressed in the laughter. The decision tree classifier performed with an accuracy of 71.02% using a minimum set of body movement features.

**Keywords:** affective laughter, laughter expression, analysis of body movement, gestures

### 1 Introduction

Laughter is an innate human emotional expression more commonly associated with positive feelings. People often laugh when they feel happy, or excited. But they also laugh when they feel embarrassed, or sad. Recent studies have shown that there is a wider range of emotions that are expressed through laughter. In particular, the study of [5] has identified 25 laughter types associated with both positive and negative emotions, and 6 laughter types not necessarily used to express emotions.

The work of [2] classified affective laughter into five types, i.e., happiness, excitement, giddiness, embarrassment, and hurtful, by analysing facial and vocal cues. Their study, however, did not include body movement and gestures in characterizing different types of laughter.

Griffin and his colleagues [3] identified perceptible body movements that can be used to distinguish different types of natural laughter. Then from these body movements, they applied supervised machine learning techniques to automatically classify different types of laughter. Motion data were recorded from 9 participants (3 males and 6 females) wearing motion capture suits in a standing and sitting positions. Samples were taken when the participants are performing specific tasks, such as playing word games, and during conversations between tasks. Participants were also request-

ed to produce fake laughter. Samples are initially labelled as laughter or non-laughter, with laughter types further classified as hilarious, social, awkward, and fake. Based on the body movement analyses of 32 observers, the study was able to determine that the hands, shoulders, spine and neck movements are useful discriminating features. Supervised learning models, which include k-Nearest Neighbor (kNN), Multi Layer Perceptron (MLP), Random Forest (RF), Linear and Kernel Ridge Regression (RR, KRR), Linear and Kernel Support Vector Regression (SVR, KSVR), were built to automatically classify laughter types. Recognition results show that the RF model outperforms the other models in terms of accuracy in classification. The model was also able to distinctly classify hilarious laughter, social laughter, and non-laughter.

The work of Niewiadomski and his colleagues on [6], on the other hand, is focused on showing the robustness of the body as a cue in discriminating laughter expressions from non-laughter expressions. This study developed a real-time system prototype to demonstrate this potential. For the first part, data from ten participants (8 males and 2 females) wearing motion capture suits were collected. Participants are allowed to converse in their native language while performing specific tasks, such as watching comedies or playing games. Two raters were engaged to annotate the data in the absence of audio information. A feature vector that captures full body movement, consisting of 13 features (F1-F13) describing head movements (left, right, front, back), weight shifting, knee bending, abdomen, trunk, arm, and shoulder movements, was computed from the motion capture data. Five learning models were built and tested to evaluate how well the discriminative algorithms (SVM, kNN, RF) and probabilistic algorithms (Naive Bayes (NB), Logistic Regression (LR)) work on the data. Based on experiment results, it was confirmed that discriminative classifiers outperforms probabilistic classifiers. The second part of this study focused on building a real-time system prototype, in which a Kinect sensor was used to body movement, 9 features (K1-K9) were computed. These features track the head, torso, and shoulder movements. Features that tracks legs and arms movement were not computed since the participants in this set-up are in a sitting position. SVM was used classifier for the prototype.

Similar to [6], this study also built a real-time system prototype that captures body movement for laughter type classification. We also used Kinect sensor to capture motion data. However, in addition to the head, torso, and shoulders, we included the hand position in reference to the head position, as part of our feature set. With this, we attempted to classify more than two types of affective laughter. This paper describes the methodology we employed and the results of our work.

This paper is organized as follows: Section 2 describes the methodology we used to build the real-time system prototype, which includes the creation and annotation of our training and test data, the computation of our feature set, and the modelling tasks; Section 3 presents the results of our experiments and the discussion of these results; and, our conclusion in Section 4.

## 2 Methodology

In this study, our goal is to investigate the possibility of using an accessible device like the Kinect sensor to capture full body movement, and with these less precise features, try to differentiate affective laughter expressions into one of the following types [7]: happiness, giddiness, excitement, embarrassment, and hurtful.

### 2.1 Data collection and annotation

Data was collected using Microsoft Kinect 360, to capture full body movement, and a Sony HD video recorder, to capture the entire session for annotation purposes. Groups of 2 to 3 friends, aged 17 to 20 years old, a mix of males and females, were invited to the data collection sessions. Recording sessions last from 5 minutes to 20 minutes, depending on how comfortable the participants are talking about a particular topic. To make sure that full body movement is captured, the target participant is asked to stand 1.5 meters away facing the other person, where the Kinect sensor is also positioned. The only instruction given to the participants is that they choose and talk freely about a topic from a given list. They were not given any time limit. The actual list of topics given to the participants include their personal interests or hobbies, embarrassing moments in high school or college, inside jokes of their group, secret crush or love life. Among these topics, talking about their crush and love life generated a lot of laughter segments. Incidentally, so are telling obscene jokes, insulting each other, or insulting a common friend.

Manual segmentation resulted to 245 laughter segments (around 72 minutes of recording) collected from 9 participants (6 males, 3 females). Non-laughter segments, speechLaughs, and fake laughs were removed. Four raters who got an above average score ( $> 49$ ) in the Baron-Cohen Empathic Quotient (EQ) Test were asked to review and annotate the laughter segments. These raters are not acquainted with the participants and they have no idea what the participants are talking about. The annotation process requires the raters to watch the video, without sound, and to identify which of the five types of affective laughter is evident in the clips. If the laughter segment cannot fit into one of the five categories, it will be marked for removal. To obtain a unified result, the inter-rater agreement was computed using Fleiss' Kappa, which is 0.57 and is considered "moderate agreement". The process resulted to 217 laughter segments distributed as follows: 71.4% happiness, 23% embarrassment, 3.2% hurtful, 1.8% giddiness, and 0.5% excitement. Table 1 shows the distribution.

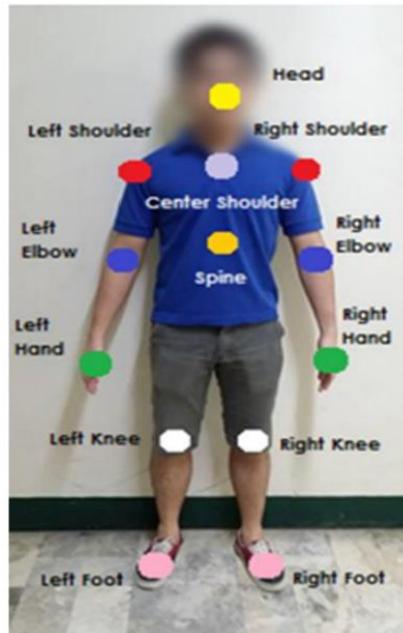
Brief interview with the raters revealed that aside from the head, torso, and shoulder, they also rely on the hand position to help them determine the laughter types. Since the participants are not moving around while laughing, points on the knee and the foot were ignored and not included in the analysis.

### 2.2 Feature extraction

The Kinect sensor collects 13 points from the body, as shown in Figure 1. These points are located at the head, left shoulder, center shoulder, right shoulder, left elbow, left hand, spine, right elbow, right hand, left knee, right knee, left foot, and right foot.

**Table 1.** Laughter segment distribution based on types of affective laughter.

<i>Laughter type</i>	<i># of segments</i>
Happiness	155
Embarrassment	50
Hurtful	7
Giddiness	4
Excitement	1
TOTAL	217

**Fig. 1.** Placement of body points as determined by the Kinect sensor.

Based on the recommendations of the raters, we computed the features from movements of the head, torso, shoulders and position of the hands. The formulas were taken from the works of [1] on real-time academic affective states recognition from body movements and gestures, and of [4] on control methods in a 3D space using Kinect. Table 2 compares the features used in this study and the features used by Niewiadomski et al in [6].

**Table 2.** Summary of body movement features used in this study in comparison with that of Niewiadomski et al in [6].

<i>Body part</i>	<i>Movement/Position</i>	<i>Niewiadomski et al [6]</i>
Head	Head tilt (left or right)	F1 – Head side movement
	Head nod (downward or upward)	F2 – Head front and back movement
Torso	Moving (forward or backward)	F6 – Trunk straightening movement
Shoulders	Leaning (forward or backward)	F7 – Trunk rocking movement
	Shrugging (shrug or no shrug)	F13 – Shoulder shaking movement
Hand	Left and Right Hand Positions (above head, on head/face, below head)	none

### 2.3 Classification

Three classification algorithms were tested on the data prior to implementation on the prototype. These are the k-Nearest Neighbor (kNN), Decision Trees (J48), and Naive Bayes (NB). Weka was used to build the models for comparison and to do the 10-fold cross validation on the results.

The Bi-Directional Best Fit First Feature Selection algorithm was applied on the feature set to determine which features are useful and should be included in the prototype.

## 3 Results and Discussion

Due to the limited number of segments for the other laughter types, the automatic classification task was reduced to binary classification between happiness and embarrassment laughter. A dataset with equal number of happiness and embarrassment segments was built using the complete feature set. The result is shown in Table 3.

The complete feature set includes the head tilt, head nod, torso movement, leaning, shrugging, right hand position, and left hand position. The reduced feature set was generated by applying the feature selection algorithm on the complete set to determine minimal set of features useful in discriminating laughter types. This resulted to a set of features that considers only the head tilt, torso movement, and the positions of the right and left hands. In terms of accuracy, the reduced feature set did not perform better than the complete feature set. However, in a system that is expected to do a classification in real-time, a reduced feature set is attractive because it implies reduced computational load.

Intuitively, the features that were retained for classification were also evident in the general observations of the raters. Head tilting movements (i.e., either to the left, or to the right) combined with hands on the head (i.e., either covering one's face, or touching one's hair) can differentiate when one is happy or embarrassed, as shown in Figure 2.

**Table 3.** Summary of classification tests using 10-fold cross validation.

Algorithm	Complete Feature Set			Reduced Feature Set		
	Accuracy	Kappa	RMSE	Accuracy	Kappa	RMSE
NB	0.5284	0.4105	0.3485	0.5272	0.4090	0.3486
kNN (n=5)	0.6628	0.5785	0.3092	0.6670	0.5838	0.3090
J48	0.7156	0.6445	0.2978	0.7102	0.6377	0.2997

Majority of the participants shrug their shoulders while laughing. However, in the data, it seems that the movement is minute which makes it undetectable by the machine. Head nodding is another movement that is uncommon for either genders. Leaning the body forward while laughing is observed in only 1 female participant. This is also the same participant who produced the most number of laugh segments.

Regardless of the number of features used in the classification task, kNN and J48 performed better than NB, which is consistent with the findings of [3] and [6].

#### 4 Conclusion

We set out to investigate the possibility of using a non-intrusive but less precise sensor like the Kinect to capture full body movement, and use these to differentiate affective laughter expressions into five classes.

Unfortunately, we were unable to collect enough laughter segments that express hurt, giddiness and excitement to appropriately explore this problem. This lack of data reduces the problem into a binary classification task. However, the classification results support the work of [6] in that it is indeed possible to use body movements as cue to differentiate laughter types, in the absence of other modalities like facial expressions or vocalizations.

Moreover, of the 217 laughter segments, 48.57% were produced by the three female participants, while the remaining 51.43% were produced by the six male participants. This may have also affected the laughter models of the classifiers.

The Kinect sensor was able to collect enough points on the body to extract useful body movement features. Based on our feature selection task, it seems useful to also include the position of the left and right hands in differentiating laughter types.

With a wide range of emotions associated with laughter, a multidimensional description of affective laughter, in conjunction with contextual information at which the laughter was expressed, will complicate the classification task but improve the accuracy of laughter type classification.



**Fig. 2.** (a) and (b) show male and female participants expressing happy laughter type. (c) and (d) show male and female participants expressing embarrassed laughter type.

**References:**

1. Cheung, O. H.: Real time academic emotion recognition using body gestures. Masters Thesis, De La Salle University (2012).
2. Galvan, C., Manangan, D., Sanchez, M., Wong, J., and Cu, J.: Audiovisual affect recognition in spontaneous Filipino laughter. In Knowledge and Systems Engineering (KSE) 2011 Third International Conference on, IEEE, 266-271 (2011).
3. Griffin, H. J., Aung, M. S. H., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., and Bianchi-Berthouze, N.: Laughter type recognition from whole body motion. In Affective Computing and Intelligent Interaction 2013 Humaine Assoc. Conf on, IEEE, 349-355 (2013).
4. Kang, J. W., Seo, D. J., and Jung, D. S.: A study on the control method of 3-dimensional space application using Kinect system. In Computer Science and Network Security Int'l Journal of, 11(9), 55-59 (2011).
5. McKeown, G., Cowie, R., Curran, W., Ruch, W., and Douglas-Cowie, E.: ILHAIRE laughter database. In Workshop on Emotion Sentiment and Social Signals LREC, Istanbul, Turkey, 32-35 (2012).
6. Niewiadomski, R., Mancini, M., Varni, G., Volpe, G., and Camurri, A.: Automated laughter detection from full-body movements. In Human Machine-Systems IEEE Trans, (2015).
7. Suarez, M. T., Cu, J., and Sta. Maria, M.: Building a multimodal laughter database. In LREC 2012, Istanbul, Turkey, (2012).

## Computational Model for Affect Detection in Learning

Najlaa Sadiq Mokhtar and Syaheerah Lebai Lutfi

School of Computer Sciences,  
Universiti Sains Malaysia,  
11800, Pulau Pinang, Malaysia

**Abstract.** Generally, computer-based tutoring system is less engaging compared to human tutoring due to its “insensitiveness” to learners’ affect. A truly intelligent tutoring system (ITS) will take into consideration the learner’s state, and adapt to that state before providing an appropriate learning content or feedback to improve learning. However, an ITS needs to firstly understand the learner’s state and his/her environment. This step is crucial to give appropriate feedback. This study focused on the four (4) frequent emotions attached to learning, which are frustration, boredom, uncertainty and neutral. Using several existing task-based features from previous studies centered in the west, combined with new features from our study, we constructed a four-class localized computational model of affect detection in learning through machine learning approach. The features collected were evaluated with several standard classifiers. Results revealed that the J48 classifier learned best when evaluated using the task-based features using a host e-tutorial that was especially developed for evaluation purposes.

**Keywords:** Intelligent tutoring system; computational emotion model; affect detection; learning; culture-sensitive

### 1 INTRODUCTION

Intelligent Tutoring Systems (ITS) should consider the needs and preferences of a learner before providing appropriate learning content. According to [10], ITS should capture all interactions between the students and ITS, such as the time taken to response to questions in the ITS log file. However, an ITS needs to firstly understand the learner’s state before providing suitable needs. This step is crucial to give appropriate feedback. It has been suggested that ITS should have the ability to adapt learners’ states in order to improve learning [2].

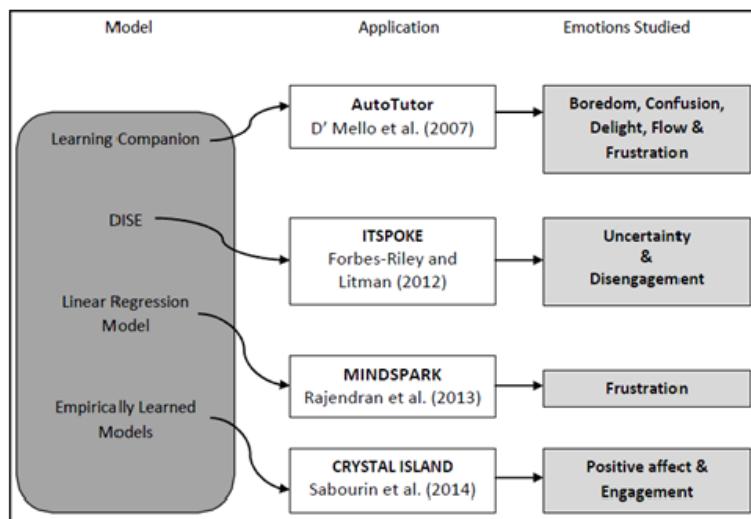
Boredom, frustration, confusion, delight, engaged concentration and surprise are the relevant affective states suggested by [1] when a student is interacting with an ITS. D’mello and colleagues [2] have found that boredom and frustration were reported at a higher rate in the relationship between affect and learning. A further study by [3] showed that uncertainty plays an important role in the

learning process because it is an affective state in the main interest of the domain of dialogue-based tutoring. In addition, affective experiences such as attitude, motivation, creativity and problem solving skills have been observed to become the factors facilitating other changes by positive emotions [9].

In this paper, we show a continued effort of our previous work [8] towards constructing a four-class computational model of affect detection in learning. By using task-based features that had been identified in the previous study [8], this paper focuses in the automatic classification of four frequent emotions that accompany learning, namely, frustration, boredom, uncertainty and neutral. A user study is conducted using a web-based e-tutorial platform that was developed especially for this study, with 33 undergraduates of the School of Computer Sciences from Universiti Sains Malaysia (USM). The following section presents the literature review in this study. Section 3 presents the methodology framework. User study and the results obtained are in Section 4. Finally, in Section 5, we conclude the findings and share the future directions of this study.

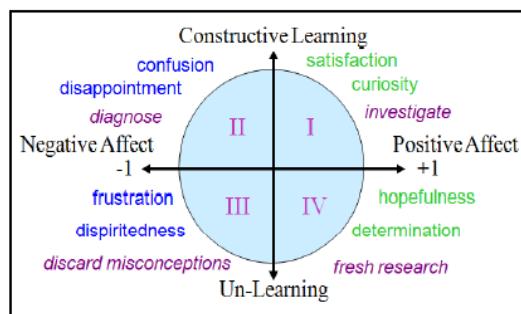
## 2 LITERATURE REVIEW

There are many models that have been developed to be used in computer-based learning environment for affect detection [10, 2, 12, 5]. Fig. 1 below summarizes the state-of-art computational model that has been integrated into various tutoring applications.



**Fig. 1.** Computational Model of affect in learning from the literature.

A previous research that was widely cited used a comprehensive four-quadrant model that clearly links learning and affective states [2]. D'Mello and colleagues (2007) developed a tutoring system, called the AutoTutor, and integrated a model called Learning Companion by [7], which modelled learning behaviour and detects the students' affective state while in the learning session. This is done by monitoring the student's posture patterns, facial features, and on-screen keyboard/mouse behaviour [2]. Fig. 2 shows the Learning Companion model.



**Fig. 2.** Learning Companion Model [7].

This model is divided into 4 quadrants, with the horizontal axis as an Emotion axis, and the vertical axis as the Learning axis. Quadrant I in the model works for investigating the student's feelings. In quadrant II, the model diagnoses specific emotions that a student actually feels. If they failed to build the simulation, they will move down to the lower space. At this point, the student's emotion may be negative and the cognitive focus changes to eliminate some misconception (quadrant III). As they consolidate their knowledge with the awareness of a sense of making progress, the student may move to quadrant IV. With a fresh idea, the student will return into the upper half of the space. Naturally, it is a counter-clockwise movement direction.

On a similar vein, [5] introduced the DISE model, which detects user's disengagement through sensing uncertainty, during spoken dialogue computer tutoring. They utilized an uncertainty-adaptive spoken dialogue tutoring system originally adopted from [4], called ITSPROKE (Intelligent Tutoring SPOKEn dialogue system). Student's speech is digitized from head-mounted microphone input after each tutor's question. It was then sent to the Sphinx2 recognizer, which yields an automatic transcript. The answer will be automatically classified based on the transcript.

On the other hand, MINDSPARK application by [10] is an approach to predict frustration in ITS. MINDSPARK is a computer-based self learning system that for mathematics. Their model works by identifying the frustration from the

ITS log file. From the linear regression model, it informs the most factors and determines which features that contribute to frustration. Thus, it is also helpful to systematically address frustration and thereby help students to avoid it.

Furthermore, a study by [12] encourages positive affect and engagement during the students' learning period. The authors used CRYSTAL ISLAND, a game-based learning platform with an underlying empirically learned model based on Bayesian Network to predict students' affect and engagement. Specifically, CRYSTAL ISLAND investigates students' affect while they are trying to solve problems on the platform.

However, it is still a challenge to design a model that underlies agent affective behavior as there are no uniformly accepted models of "learning" emotions [11]. In addition, there are other difference a learning model should take into consideration. For example, the learning environment between Malaysia and western countries is different. Based on [12], individual differences such as gender, academic and cultural background will affect their learning style. A suitable computational model that is context-specific (localized) is therefore required for this reason.

### 3 METHODOLOGY

Fig. 3 shows the high level methodological framework that encompasses this study. As mentioned earlier, using task-based features, this study focused on constructing a computational model that could predict students' emotions that occur frequently in learning in a Malaysian context; boredom, frustration, uncertainty and neutral. Further details on the methodology is reported in [8].

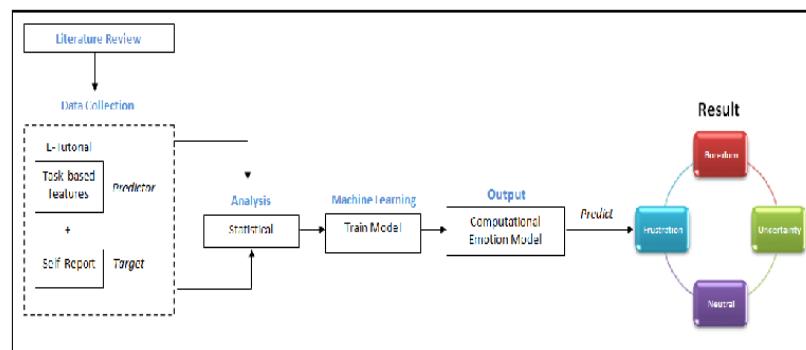


Fig. 3. High level research framework.

The data collection phase involved a user study with 33 students of Year 1 and Year 2 undergraduates in Computer Science from USM. Based on race,

there were 13 Malay and 20 Chinese students who participated. There were no Indian students (The Computer Sciences School in USM has quite a low number of Indian students in comparison to other races). Each student was asked to use the E-Tutorial Programming system that was developed especially for this study. There are two tutorials, Tutorial 1 and 2, and both the tutorials contain basic to advanced programming questions that are randomly mixed up. Specifically, each tutorial contains 30 questions – which are: basic (10 questions), intermediate (10 questions) and advanced (10 questions) – and it is split into 10 questions for each session that must be answered. For each tutorial, students get 30 minutes, totaling 1 hour for both tutorial sessions.

As shown in Table 1, there are 11 objective features that were grouped into three parts - student, question and tutorial-based features. These features were automatically collected and kept in a log during the student-system interaction. The significant objective features mined from the interaction will be fed as predictors in training the model in a supervised learning.

**Table 1.** Task-based features and their descriptions

Type	Features	Description
Student Features	1. Gender 2. Races 3. Experience 4. Interest 5. Personality 6. Tutor Chosen	Gender of the student Race of the student ( <i>Malay/Chinese/Indian</i> ) Student experience using computer (not familiar/familiar/very familiar) Student interest with the subject in the tutorial ( <i>Strongly agree/ Agree/ Neutral/ Disagree/ Strongly disagree</i> ) Personality of the students ( <i>Openness/ Conscientiousness/ Extravert/ Agreeableness/ Neurotic</i> ) The students preferred synthetic mix (only for Tutorial 2)
Tutorial Features	7. Time Taken (TimeTaken_Sec) 8. Performance (Answer) 9. Clicking	Time taken (in seconds) to answer each question in the tutorial. The number of correct answers The number of clicks before submitting the final answer
Question Features	10. Difficulty (QuesLevelDiff) 11. Type (Ques_type)	Level of difficulty of each question (easy/medium/hard) Type of the question asked (theory/codintg)

From a total of 11 task-based features, only 8 features showed correlation to emotion [8]. Table 1 shows the selected task-based features that had been used in order to construct the model. All the task-based features were ranked. This is to see which features are more important among all features. All the features were ranked using WEKA [6]. Therefore, Table 2 shows the features after ranking that was done for Tutorial 1 and for Tutorial 2.

**Table 2.** Selected task-based features

Tutorial 1 Ranked attributes:		Tutorial 2 Ranked attributes:	
1	Answer	1	Races
2	Interest	2	Answer
3	QuesType	3	Clicking
4	Clicking	4	QuestionLevelDiff
5	TimeTaken_Sec	5	TimeTaken_Sec
6	Races	6	Interest
7	Gender	7	Gender
8	QuestionLevelDiff	8	QuestionType

As mentioned previously, there were two (2) tutorials that the students are required to complete. The main difference between Tutorial 1 and Tutorial 2 is that Tutorial 1 has no e-tutor while in Tutorial 2 there are three built-in non-adaptive standard e-tutors, and students can choose their preferred e-tutor. Fig. 4 showed the synthetic tutors that were featured in Tutorial 2.



**Fig. 4.** Synthetic Tutors. From left to right: perceived Malay, Chinese and Indian tutors.

These synthetic tutors are specially designed to reflect Malaysian races. Students were not told which tutor represents which race and were not conditioned to choose their preferred tutors.

## 4 RESULTS AND DISCUSSION

This section summarizes the results obtained from the experiment mentioned above.

### 4.1 Training the Detection Model using WEKA

As for the result, in order to validate the features, the model should be trained first. The model is trained using WEKA and 10-fold cross validation. We compared five (5) types of classifiers in the experiment, which are zeroR, Naive Bayes, SMO function, Simple Logistic function and J48 tree. The experiments were run separately for Tutorial 1 and Tutorial 2.

Table 3 shows the results obtained for Tutorial 1, for all the above mentioned classifiers. The J48 tree based classifier yielded the best accuracy using 8 selected optimum features (refer to Table 1) at 99.16% (F-measure = 0.99).

**Table 3.** The comparison of classifiers for Tutorial 1

Features	Classifier	ZeroR <sup>1</sup>	Naive Bayes	J48	SMO	Simple logistic
Optimum Features	Accuracy	42.73	79.90	99.16	84.57	86.23
	F-measure	0.26	0.79	0.99	0.84	0.86

1. All classifiers showed statistically significantly improved results above and beyond the baseline at  $p = 0.01$  (two-tailed t-test)

Table 4 below shows the confusion matrix for Tutorial 1 obtained using J48 tree. From the result, it showed that correct prediction for boredom is 178 and incorrect prediction only in 3 cases that confused with frustration and uncertainty. While, for frustration, correct prediction is 204 and incorrect prediction 4 (confused with neutral) and 1 (confused with uncertainty). In addition, neutral gets 421 correct prediction and the other 2 confused with frustration. Uncertainty shows 100% correct prediction with 181.

Table 5 shows the percentage accuracies obtained by all classifiers in Tutorial 2. Again, J48 classifier learned best with a 97.59% of accuracy (F-measure = 0.98).

Table 6 below shows the confusion matrix for Tutorial 2 when learned by the J48 tree-based classifier. The confusion matrix revealed boredom, frustration and uncertainty are learned near perfection with very little confusion while neutral

**Table 4.** The confusion matrix for Tutorial 1

a	b	c	d	Classifier
178	1	0	2	= a)Boredom
0	204	4	1	= b)Frustration
0	2	421	0	= c)Neutral
0	0	0	181	= d)Uncertainty

**Table 5.** The comparison of classifiers for Tutorial 2

Features	Classifier	ZeroR <sup>1</sup>	Naive Bayes	J48	SMO	Simple logistic
Optimum Features	Accuracy	39.39	87.11	97.59	88.22	90.22
	F-measure	0.22	0.86	0.98	0.88	0.90

1. All results are statistically significant at  $p = 0.01$  (two-tailed)

is mostly confused with boredom. This is expected, as when the learners show no emotion, it could be taken as being bored.

Therefore, we conclude that J48 classifier is the best classifier to predict student's emotion using task-based features in e-tutorial. This is because results obtained showed high accuracies from both tutorials. Table 7 below shows the accuracy using J48 tree classifier for both Tutorial 1 and Tutorial 2.

#### 4.2 Computational Model

In order to achieve the main objective of this paper, a computational model of affect detection is constructed using J48 tree classifier to predict emotions; boredom, frustration, uncertainty and neutral in learning. Fig. 5 below shows the model for Tutorial 1 produced by J48 tree classifier and Fig. 6 the model for Tutorial 2.

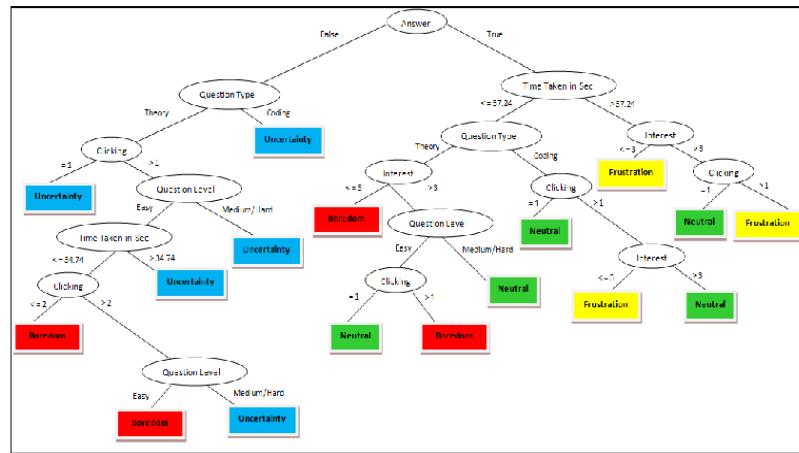
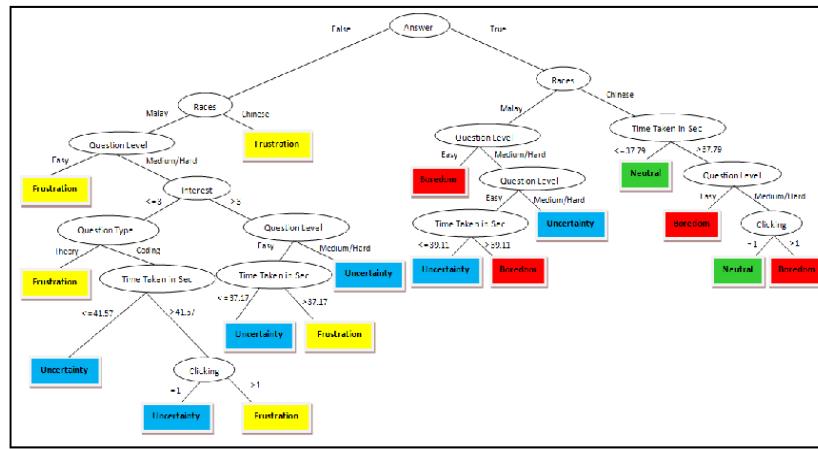
The J48 tree classifier produced a set of rules. Those rules are extracted from the optimum task-based features that had been trained. From the tree (Fig. 5

**Table 6.** Confusion matrix for Tutorial 2

a	b	c	d	Classifier
199	0	2	1	= a)Boredom
0	196	0	4	= b)Frustration
5	3	379	3	= c)Neutral
2	2	1	193	= d)Uncertainty

**Table 7.** Percentage accuracy when evaluated by J48 classifier.

Tutorial	Correctly Instances	Incorrectly Instances
1	99.16%	0.84%
2	97.59%	2.41%

**Fig. 5.** Computational model for Tutorial 1.**Fig. 6.** Computational model for Tutorial 2.

and Fig. 6) we can gain insight as to what leads to the emotions (boredom, frustration, uncertainty and neutral).

However, based on both models that had been constructed, we agreed that it should have some modification (in the future) to make it flexible with any other situation, especially in the real-time learning session since this model is one part of the computational model which is only to detect the emotion of students during learning.

We also agreed that affect synthetic virtual tutor is important (in the future) in order to give appropriate feed-back towards students' emotion. In the future, with some fine-tuning, these rules can be implemented in a host synthetic tutor to make it more adaptive towards students' states.

## 5 CONCLUSION

In this study, a 4-class localized computational model of emotion that would be the underlying model of an affect-sensitive synthetic tutor in learning was successfully constructed using J48 tree classifier. The model yielded high recognition accuracy statistically significantly above and beyond the base rate for both Tutorial 1 and Tutorial 2 to predict boredom, frustration, uncertainty and neutral using a host e-tutoring platform. However, this is an offline prediction, meaning that training and prediction were done using previously collected data and not in a real-time environment. We foresee a slight mismatch between offline and online prediction. It is vital for a synthetic tutor to have the sense-and-express ability to provide a proper intervention in learning in view of the learner's emotion.

## 6 ACKNOWLEDGEMENT

The authors would like to thank Universiti Sains Malaysia for the funding of this work from the grant no. 304/PKOMP/6312153.

## References

1. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68(4), 223–241 (2010)
2. D'Mello, S., Picard, R.W., Graesser, A.: Toward an affect-sensitive autotutor. *IEEE Intelligent Systems* (4), 53–61 (2007)
3. D'mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B., Graesser, A.: Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction* 18(1-2), 45–80 (2008)
4. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53(9), 1115–1136 (2011)

5. Forbes-Riley, K., Litman, D., Friedberg, H., Drummond, J.: Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 91–102. Association for Computational Linguistics (2012)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
7. Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In: icalt. p. 0043. IEEE (2001)
8. Mokhtar, N.S., Lutfi, S.L.: Identifying significant task-based predictors of emotion in learning
9. Plass, J.L., Heidig, S., Hayward, E.O., Homer, B.D., Um, E.: Emotional design in multimedia learning: Effects of shape and color on affect and learning. Learning and Instruction 29, 128–140 (2014)
10. Rajendran, R., Iyer, S., Murthy, S., Wilson, C., Sheard, J.: A theory-driven approach to predict frustration in an ITS. IEEE Transactions on Learning Technologies 6(4), 378–388 (2013)
11. Sabourin, J., Mott, B., Lester, J.: Computational models of affect and empathy for pedagogical virtual agents. In: Standards in emotion modeling, Lorentz Center International Center for workshops in the Sciences (2011)
12. Sabourin, J.L., Lester, J.C.: Affect and engagement in game-basedlearning environments. IEEE Transactions on Affective Computing 5(1), 45–56 (2014)

## Verifying Properties of Multi-agent Systems via Bounded Model Checking <sup>\*</sup>

Agnieszka M. Zbrzezny

IMCS, Jan Dlugosz University, Al. Armii Krajowej 13/15, 42-200 Czestochowa, Poland.  
agnieszka.zbrzezny@ajd.czest.pl

**Abstract.** The objectives of this research are to further investigate the foundations for novel SMT and SAT-based bounded model checking (BMC) algorithms for multi-agent systems. A major part of the research will involve the development of SMT-based BMC methods for different kinds of interpreted systems.

### 1 Introduction

Model checking problem has been proposed independently by Quielle and Sifakis [6], and by Clarke and Emerson [3] as a method for automatic and algorithmic verification of finite state concurrent systems. Model checking of multi-agent systems [5] is a very active field for both theoretical research and practical applications, but it is known to be a difficult problem and its practical applicability is strongly limited by the state explosion problem. There is still a lack of efficient methods for MAS. In view of this, there is an obvious need to develop efficient SMT/SAT-based verification methods which could be used in practice.

The main aim is to compare the existing SAT-based bounded model checking algorithms for different extensions of interpreted systems with SMT-based bounded model checking techniques for the same models.

### 2 Extensions of Interpreted Systems

In the paper we have modelled MAS using two extensions of interpreted systems: weighted interpreted systems and timed weighted interpreted systems. We expressed properties of MAS in WELTLK and WECTLK [7] logics. The formalism of weighted interpreted systems (WIS) [7] extends ISs to make the reasoning possible about not only temporal and epistemic properties, but also agents's quantitative properties. Timed weighted interpreted systems (TWIS) were proposed in [10] to extend interpreted systems in order to make possible reasoning about real-time aspects of MASs and agents's quantitative properties.

Let  $\mathcal{A} = \{1, \dots, n\}$  denotes a non-empty and finite set of agents, and  $\mathcal{E}$  be a special agent that is used to model the environment in which the agents operate, and  $\mathcal{AP} = \bigcup_{i \in \mathcal{A} \cup \{\mathcal{E}\}} \mathcal{AP}_i$  be a set of atomic formulae, such that  $\mathcal{AP}_{i_1} \cap \mathcal{AP}_{i_2} = \emptyset$  for all  $i_1, i_2 \in \mathcal{A} \cup \{\mathcal{E}\}$ .

---

\* Partly supported by National Science Centre under the grant No. 2014/15/N/ST6/05079.

**Weighted Interpreted Systems.** The *weighted interpreted system* (WIS) is a tuple  $(\{L_i, Act_i, P_i, \mathcal{V}_i, d_i, \iota_i\}, \{t_i\}_{i \in \mathcal{A}}, \{t_{\mathcal{E}}\})$ , where  $L_i$  is a non-empty set of *local states* of the agent  $i$ ,  $S = (L_1 \times \dots \times L_n \times L_{\mathcal{E}})$  is the set of all the *global states*  $\iota \subseteq L_i$  is a non-empty set of initial states,  $Act_i$  is a non-empty set of *possible actions* of the agent  $i$ ,  $Act = Act_1 \times \dots \times Act_n \times Act_{\mathcal{E}}$  is the set of *joint actions*,  $P_i : L_i \rightarrow 2^{Act_i}$  is a *protocol function* that define rules according to which actions may be performed in each local state,  $t_i : L_i \times Act \rightarrow L_i$  is a (partial) *evolution function*,  $\mathcal{V}_i : L_i \rightarrow 2^{\mathcal{AP}}$  is a *valuation function* assigning to each local state a set of propositional variables that are assumed to be true at that state, and  $d_i : Act_i \rightarrow \mathbb{N}$  is a *weight function*.

For a given WIS we define: a set of all *possible global states*  $S = L_1 \times \dots \times L_n \times L_{\mathcal{E}}$ ; by  $\ell_i(s)$  we denote the local component of agent  $i \in \mathcal{A} \cup \{\mathcal{E}\}$  in a global state  $s = (\ell_1, \dots, \ell_n, \ell_{\mathcal{E}})$ ; and a *global evolution function*  $t : S \times Act \rightarrow S$  as follows:  $t(s, \tilde{a}) = s'$  iff for all  $i \in \mathcal{A}$ ,  $t_i(\ell_i(s), \tilde{a}) = \ell_i(s')$  and  $t_{\mathcal{E}}(\ell_{\mathcal{E}}(s), \tilde{a}) = \ell_{\mathcal{E}}(s')$ , where  $\tilde{a} \in Act$  is a joint action. In brief we write the above as  $s \xrightarrow{\tilde{a}} s'$ .

Now, for a given weighted interpreted system we define a *weighted model* (or simply a *model*) as a tuple  $\mathcal{M} = (Act, S, \iota, \mathcal{V}, d)$ , where:  $\iota = (\iota_1 \times \dots \times \iota_n \times \iota_{\mathcal{E}})$  is the set of all the *initial* global states,  $\mathcal{V} : S \rightarrow 2^{\mathcal{AP}}$  is the valuation function defined as  $\mathcal{V}(s) = \bigcup_{i \in \mathcal{A} \cup \{\mathcal{E}\}} \mathcal{V}_i(l_i(s))$ ,  $T \subseteq S \times Act \times S$  is a transition relation defined by the global evolution function as follows:  $(s, \tilde{a}, s') \in T$  iff  $s \xrightarrow{\tilde{a}} s'$ , and  $d : Act \rightarrow \mathbb{N}$  is the “joint” weight function defined as follows:  $d((a_1, \dots, a_n, a_{\mathcal{E}})) = d_1(a_1) + \dots + d_n(a_n) + d_{\mathcal{E}}(a_{\mathcal{E}})$ . Given a WIS one can define the indistinguishability relation  $\sim_i \subseteq S \times S$  for agent  $i$  as follows:  $s \sim_i s'$  iff  $l_i(s') = l_i(s)$ .

**Timed Weighted Interpreted Systems** Let  $\mathbb{N}$  be a set of natural numbers, and  $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ . We assume a finite set  $\mathbb{X}$  of variables, called *clocks*. Each clock is a variable ranging over a set of non-negative natural numbers. For  $x \in \mathbb{X}$ ,  $\bowtie \in \{<, \leq, =, >, \geq\}$ ,  $c \in \mathbb{N}$  we define a set of clock constraints over  $\mathbb{X}$ , denoted by  $\mathcal{C}(\mathbb{X})$ . The constraints are conjunctions of comparisons of a clock with a time constant  $c$  from the set of natural numbers  $\mathbb{N}$ , generated by the grammar:  $cc := \text{true} \mid x \bowtie c \mid cc \wedge cc$ . A clock valuation  $v$  of  $\mathbb{X}$  is a total function from  $\mathbb{X}$  into the set of natural numbers. The set of all the clock valuations is denoted by  $\mathbb{N}^{|\mathbb{X}|}$ . For  $\mathbb{X}' \subseteq \mathbb{X}$ , the valuation which assigns the value 0 to all clocks is defined as:  $\forall_{x \in \mathbb{X}'} v'(x) = 0$  and  $\forall_{x \in \mathbb{X} \setminus \mathbb{X}'} v'(x) = v(x)$ . A clock valuation  $v$  satisfies a clock constraint  $cc$ , written as  $v \models cc$ , iff  $cc$  evaluates to true using the clock values given by  $v$ .

A *timed weighted interpreted system* is a tuple  $\text{TWIS} = (\{L_i, Act_i, \mathbb{X}_i, P_i, \mathcal{V}_i, \mathcal{I}_i, \iota_i, d_i\}_{i \in \mathcal{A} \cup \{\mathcal{E}\}}, \{t_i\}_{i \in \mathcal{A}}, \{t_{\mathcal{E}}\})$ , where:  $L_i$  is a non-empty set of *locations* of the agent  $i$ ,  $\iota_i \subseteq L_i$  is a non-empty set of initial locations,  $Act_i$  is a non-empty set of *possible actions* of the agent  $i$ ,  $Act = Act_1 \times \dots \times Act_n \times Act_{\mathcal{E}}$  is the set of *joint actions*,  $\mathbb{X}_i$  is a non-empty set of *clocks*,  $P_i : L_i \rightarrow 2^{Act_i}$  is a *protocol function*,  $t_i : L_i \times \mathbb{X}_i \times \mathcal{C}(\mathbb{X}_i) \times 2^{\mathbb{X}_i} \times Act \rightarrow L_i$  is a (partial) *evolution function* for agents,  $t_{\mathcal{E}} : \mathbb{X}_{\mathcal{E}} \times \mathcal{C}(\mathbb{X}_{\mathcal{E}}) \times 2^{\mathbb{X}_{\mathcal{E}}} \times Act \rightarrow L_{\mathcal{E}}$  is a (partial) *evolution function* for environment,  $\mathcal{V}_i : L_i \rightarrow 2^{\mathcal{AP}_i}$  is a *valuation function* assigning to each location a set of atomic formulae that are assumed to be true at that location,  $\mathcal{I}_i : L_i \rightarrow \mathcal{C}(\mathbb{X}_i)$  is an *invariant function*, that specifies the amount of time the agent  $i$  may spend in a given location, and  $d_i : Act_i \rightarrow \mathbb{N}$  is a *weight function*. It is assumed that locations, actions and clocks for an environment are “public”. We also assume that if  $\epsilon_i \in P_i(\ell_i)$ , then  $t_i(\ell_i, \ell_{\mathcal{E}}, cc_i, \mathbb{X}, (a_1, \dots, a_n, a_{\mathcal{E}})) =$

$\ell_i$  for  $a_i = \epsilon_i$ , any  $\text{cc}_i \in \mathcal{C}(\mathbb{X})$ , and any  $\mathbb{X} \in 2^{\mathbb{X}_i}$ . Each element  $t$  of  $t_i$  is denoted by  $\langle \ell_i, \ell_{\mathcal{E}}, \text{cc}_i, \mathbb{X}', a, \ell'_i \rangle$ , where  $\ell_i$  is the source location,  $\ell'_i$  is the target location,  $a$  is an action,  $\text{cc}$  is the enabling condition for  $t_i$ , and  $\mathbb{X}' \subseteq \mathbb{X}$  is the set of clocks reset when performing  $t$ . An invariant condition allows the TWIS to stay at the location  $\ell$  as long only as the constraint  $\mathcal{I}_i(\ell_i)$  is satisfied. The guard  $\text{cc}$  has to be satisfied to enable the transition.

For a given TWIS let the symbol  $S = \prod_{i \in \mathcal{A} \cup \{\mathcal{E}\}} (L_i \times \mathbb{N}^{|\mathbb{X}_i|})$  denotes the non-empty set of all *global states*. Moreover, for a given global state  $s = ((\ell_1, v_1), \dots, (\ell_n, v_n), (\ell_{\mathcal{E}}, v_{\mathcal{E}})) \in S$ , let the symbols  $l_i(s) = \ell_i$  and  $v_i(s) = v_i$  denote, respectively, the local component and the clock valuation of agent  $i \in \mathcal{A} \cup \{\mathcal{E}\}$  in  $s$ . Now, for a given TWIS we define a *timed model* (or a *model*) as a tuple  $\mathcal{M} = (Act, S, \iota, T, \mathcal{V}, d)$ , where:  $\iota = (\iota_1 \times \{0\}^{|\mathbb{X}_1|}) \times \dots \times (\iota_n \times \{0\}^{|\mathbb{X}_n|}) \times (\iota_{\mathcal{E}} \times \{0\}^{|\mathbb{X}_{\mathcal{E}}|})$  is the set of all the *initial* global states,  $\mathcal{V} : S \rightarrow 2^{\mathcal{AP}}$  is the valuation function defined as  $\mathcal{V}(s) = \bigcup_{i \in \mathcal{A} \cup \{\mathcal{E}\}} \mathcal{V}_i(l_i(s))$ ,  $T \subseteq S \times (Act \cup \mathbb{N}) \times S$  is a transition relation defined by action and time transitions. For  $\tilde{a} \in Act$ : action transition:  $(s, \tilde{a}, s') \in T$  (or  $s \xrightarrow{\tilde{a}} s'$ ) iff for all  $i \in \mathcal{A} \cup \{\mathcal{E}\}$ , there exists a local transition  $t_i(l_i(s), \text{cc}_i, \mathbb{X}', \tilde{a}) = l_i(s')$  such that  $v_i(s) \models \text{cc}_i \wedge \mathcal{I}(l_i(s))$  and  $v'_i(s') = v_i(s)[\mathbb{X}' := 0]$  and  $v'_i(s') \models \mathcal{I}(l_i(s'))$ ; time transition  $(s, \delta, s') \in T$  iff for all  $i \in \mathcal{A} \cup \{\mathcal{E}\}$ ,  $l_i(s) = l_i(s')$  and  $v'_i(s') = v_i(s) + \delta$  and  $v'_i(s') \models \mathcal{I}(l_i(s'))$ . The “joint” weight function  $d : Act \rightarrow \mathbb{N}$  is defined as follows:  $d((a_1, \dots, a_n, a_{\mathcal{E}})) = d_1(a_1) + \dots + d_n(a_n) + d_{\mathcal{E}}(a_{\mathcal{E}})$ . Given a TWIS, one can define for any agent  $i$  the indistinguishability relation  $\sim_i \subseteq S \times S$  as follows:  $s \sim_i s'$  iff  $l_i(s') = l_i(s)$  and  $v_i(s') = v_i(s)$ .

### 3 SAT and SMT

SAT-based bounded model checking (BMC) [1] uses a reduction of the problem of truth of a modal formula in a model (transition system) to the problem of satisfiability of formulae of the classical propositional calculus, i.e. SAT-problem. The reduction is achieved by a translation of the transition relation and a translation of a given property to formulae of classical propositional calculus. It should be emphasised that for a given temporal logic, bounded model checking is mainly used to disprove safety properties and to prove liveness properties.

The SMT problem [2] is a generalisation of the SAT problem, where Boolean variables are replaced by predicates from various background theories, such as linear, real, and integer arithmetic. SMT generalises SAT by adding equality reasoning, arithmetic, fixed-size bit-vectors, arrays, quantifiers, and other useful first-order theories. Using SMT to express different problems has important advantages over SAT.

### 4 Experimental results

We have performed the experiments using extensions of benchmark: the *generic pipeline paradigm* (GPP) [4]: the weighted GPP (WGPP) [7], and the timed weighted GPP (TWGPP) [10].

#### 4.1 SAT/SMT-based BMC for WELTLK and WIS [8, 11]

Let  $Min$  denote the minimum cost needed to ensure that Consumer receives the data produced by Producer. Note, that we describe specifications as universal formulae, for which we verify the corresponding counterexample formulae that are interpreted existentially and belong to WELTLK. Moreover, for every specification given, the corresponding WELTLK formula holds in the model of the benchmark.

**WGPP** The specifications we consider for the WGPP are:

- $\varphi_1 = K_P G_{[Min, Min+1]} \text{ConsReceived}$ , which expresses that Producer knows that always the cost of receiving by Consumer the commodity is  $Min$ .
- $\varphi_2 = K_P G(ProdSend \rightarrow F_{[0, Min - d_P(Produce)]} \text{ConsReceived})$ , which states that Producer knows that always if she/he produces a commodity, then Consumer receives the commodity and the cost is less than  $Min - d_P(\text{Produce})$ .

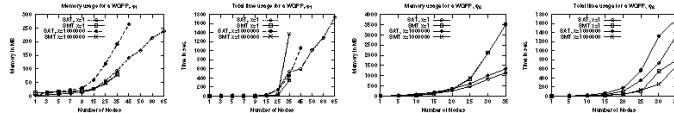


Fig. 1: SMT- and SAT-based BMC for WELTLK and WIS: WGPP with n nodes.

The experimental results show that the SMT-BMC is sensitive to scaling up the size of the benchmarks, but it is not sensitive to scaling up the weights, while the SAT-based BMC is more sensitive to scaling up the weights.

#### 4.2 SAT/SMT-based BMC for WELTLK and TWIS [12]

**TWGPP** The specifications we consider are as follows:

- $\varphi_1 = F_{[0, \infty)}(G_{[0, \infty)}(\neg \text{ConsReady}))$ , which expresses that there exist a computation that always Consumer is ready to ready to consuming the data.
- $\varphi_2 = K_P G_{[Min, Min+1]} \text{ConsFree}$ , which expresses that Producer knows that always the cost of receiving by Consumer the commodity is  $Min$ .

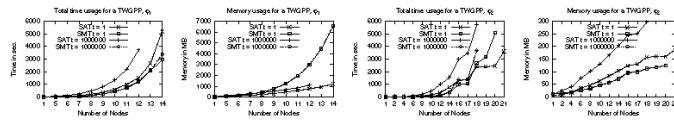


Fig. 2: SAT- and SMT-based BMC for TWIS and WELTLK: TWGPP with n nodes.

As one can see from the line charts (Fig. 2) the SMT-based BMC performs much better in terms of the total time and the memory consumption for both the tested formulae.

#### 4.3 SAT/SMT-based BMC for WECTLK and WIS [9,10]

**WGPP** The specifications we consider are as follows:

- $\varphi_1 = \overline{K}_P \text{EF}_{[Min, Min+1)} \text{ConsReady}$  - it expresses that it is not true that Producer knows that always the cost incurred by Consumer to receive data is Min.
- $\varphi_2 = \overline{K}_P \text{EF}(\text{ProdSend} \wedge \overline{K}_C \overline{K}_P \text{EG}_{[0, Min-p)} \text{ConsReady})$  - it states that it is not true that Producer knows that always if it produces data, then Consumer knows that Producer knows that Consumer has received data and the cost is less than  $Min - p$ .

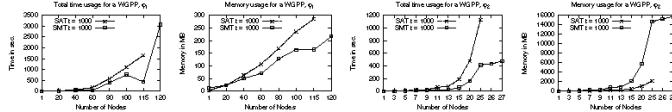


Fig. 3: SAT- and SMT-based BMC for WECTLK and WIS: WGPP with  $n$  nodes.

From Fig. 3 we can notice that for the WGPP system and both considered formulae the SMT-based BMC is faster than the SAT-based BMC, however, the SAT-based BMC consumes less memory.

#### 4.4 SMT-based BMC for WECTLK and TWIS [10]

**TWGPP** The specifications we consider are as follows:

- $\varphi_1 = \text{EF}_{[0, Right]}(\text{ConsFree})$  - it states that there exists a path on which Consumer receives a data and the cost of receiving the data will be less than Right.
- $\varphi_2 = \text{EF}_{[0, Right]}(\text{ConsFree} \wedge \text{EG}(\text{ProdSend} \vee \text{ConsFree}))$  - it states that there exists a path on which Consumer receives a data and the cost of receiving the data is less than Right and from that point there exists a path on which always either the Producer has sent a data or the Consumer has received a data.

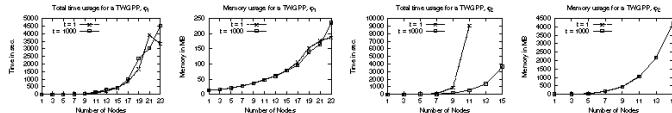


Fig. 4: SMT-based BMC for WECTLK and WIS: WGPP with  $n$  nodes.

One can observed that our SMT-based BMC is not slightly sensitive to scaling up the weights, but it is sensitive to scaling up the size of benchmark.

## 5 Conclusions

We have compared our SMT-based methods with the corresponding SAT-based method. The experimental results show that the approaches are complementary. Also an observation of experimental results leads to the conclusion that the SAT-based BMC uses less memory comparing to the SMT-based BMC. In our future work we would like to define SAT-based bounded model checking method for the properties expressible in WECTLK and interpreted over timed weighted interpreted systems.

## References

1. A. Biere, A. Cimatti, E. Clarke, M. Fujita, and Y. Zhu. Symbolic model checking using SAT procedures instead of BDDs. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC'99)*, pages 317–320, 1999.
2. A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors. volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009.
3. E. M. Clarke, E. A. Emerson, and A. P. Sistla. Automatic verification of finite state concurrent systems using temporal logic specifications: A practical approach. In *Conference Record of the 10 Annual ACM Symp. on Principles of Programming Languages*, pages 117–126. ACM Press, 1983.
4. Edmund M. Clarke, Orna Grumberg, and Doron Peled. *Model Checking*. MIT Press, 2001.
5. R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.
6. J. P. Quielle and J. Sifakis. Specification and verification of concurrent systems in CESAR. In *Proceedings of the 5th International Symp. on Programming*, volume 131 of *LNCS*, pages 337–351. Springer-Verlag, 1981.
7. B. Woźna-Szczęśniak. SAT-based Bounded Model Checking for Weighted Deontic Interpreted Systems. In *EPIA 2013*, volume 8154 of *LNAI*, pages 444–455. Springer-Verlag, 2013.
8. B. Woźna-Szczęśniak, A. M. Zbrzezny, and A. Zbrzezny. SAT-based Bounded Model Checking for Weighted Interpreted Systems and Weighted Linear Temporal Logic. In *PRIMA 2013*, volume 8291 of *LNAI*, pages 355–371. Springer-Verlag, 2013.
9. B. Woźna-Szczęśniak, A. M. Zbrzezny, and A. Zbrzezny. Bounded model checking for weighted interpreted systems and for flat weighted epistemic computation tree logic. In *PRIMA'2014*, volume 8861 of *LNAI*, pages 107–115. Springer-Verlag, 2014.
10. A. M. Zbrzezny and A. Zbrzezny. Checking WECTLK Properties of Timed Real-Weighted Interpreted Systems via SMT-Based Bounded Model Checking. In *EPIA 2015*, volume 9273 of *LNCS*, pages 638–650. Springer, 2015.
11. A. M. Zbrzezny and A. Zbrzezny. Checking WELTLK properties of weighted interpreted systems via smt-based bounded model checking. In *PRIMA 2015*, LNAI, pages 660–669, 2015.
12. A. M. Zbrzezny, A. Zbrzezny, and F. Raimondi. Efficient model checking timed and weighted interpreted systems using SMT and SAT solvers. In *Smart Innovation, Systems and Technologies. 2016, KES AMSTA*, pages 45–55, 2016.

## **MOEPSO for Multi-Objective Optimization**

Ittikon Thammachantuek<sup>1</sup> and Mahasak Ketcham<sup>2</sup>

Faculty of Information Technology King Mongkut's University of Technology North Bangkok  
*i.thammachantuek@gmail.com<sup>1</sup>, maoquee@hotmail.com<sup>2</sup>*

**Abstract.** This paper presents an Multi-Objective Evolutionary Particle Swarm Optimization (MOEPSO) which each iteration particles are improved by Evolutionary Algorithms. To solves Multi-Objective Optimization problem the proposed algorithm has been compare with Multi-Objective Particle Swarm Optimization (MOPSO) using a Test Problem BNH results show that the proposed algorithm can provide more efficient solution.

**Keywords:** Evolutionary Particle Swarm Optimization , Multi-Objective Optimization , Multi-Objective Evolutionary Particle Swarm Optimization

### **1 INTRODUCTION**

Optimization problem is a common problem in the field of engineering, science, management and more. The answer has to be under condition which is a mathematical function. One way to get the answer is Multi-Objective Particle Swarm Optimization (MOPSO) which apply Particle Swarm Optimization (PSO) to solve Multi-Objective Optimization problem. However, there are still some problems in the algorithm. For example, PSO has a convergence to local optimum too early [1].The main objective of this paper is apply Evolutionary Particle Swarm optimization (EPSO) to solve Multi-Objective Optimization problem. The new algorithm called Multi-Objective Evolutionary Particle Swarm Optimization (MOEPSO)

### **2 LITERATURE REVIEW**

#### **2.1 Optimization problem**

Optimization problem plays a key role in helping to bring solutions in engineering, science, management and more [2]. Optimization Mathematical Programming is a method that finding the value of  $x$  which makes  $f(x)$  is a minimum or maximum [3, 4]. Finding Global optimum is divided into two types. First is Global maximum and another is Global minimum. The difficult of this problem is how to split the really best (Global optimum) from Local optimum. In the last few years, there are some studies that use computer algorithm to solve this problem such as Genetic Algorithm,

Ant Colony Optimization, Particle Swarm Optimization, etc. Optimization problem has two types considered by Objective function. First, Single-Objective Optimization problem that has one Objective function. Moreover, Multi-Objective Optimization that has at least two Objective functions.

## 2.2 Particle Swarm Optimization (PSO)

The Particle Swarm Optimization is an optimization algorithm that was introduced in 1995 by Kennedy [5]. PSO has a population of particles looking around in a given search space for the global optimum. Each particle has fitness value evaluated by fitness function and velocity which is a trajectory of particle. In each iteration a particle will update its two values. First, is its best position (pbest) and other is the best position of population (gbest).

## 2.3 Multi-Objective Particle Swarm Optimization (MOPSO)

[6] Discussed about Multi-Objective Particle Swarm Optimization. The algorithm of MOPSO is the following.

1. Initialize the population POP
  - FOR i = 0 to MAX /\* MAX = number of particle \*/
  - Initialize POP[i]
2. Initialize the velocity (speed) of each particle
  - FOR I = 0 TO MAX
  - VEL[i] = 0
3. Evaluate each of the particles in POP
4. Store the positions of the particles that represent nondominated vectors in the repository REP.
5. Generate hypercube of the search space explored so far, and locate the particles using these hypercube as a coordinate system where each particle's coordinates are defined according to the values of its objective functions.
6. Initialize the memory of each particle (this memory serves as a guide to travel through the search space. This memory is also stored in the repository)
  - FOR i = 0 TO MAX
  - PBESTS[i] = POP[i]
7. WHILE maximum number of cycles has not been reached
 

DO

  - Compute the speed of each particle using the following expression

$$v_i(t) = w * v_i(t-1) + c_1 r_1 [P_{pbest_i} - p_i(t)] + c_2 r_2 [P_{gbest} - p_i(t)] \quad (1)$$

Where  $v_i(t)$  is a current speed of particle

$w$  is an inertia weight

$c_1, c_2$  is a constant value in this paper  $c_1 = 2.8$  and  $c_2 = 1.3$

$r_1, r_2$  is a random value in the range  $[0,1]$

$P_{pbest_i}$  is the best position that the particle  $i$  has had.

$P_{gbest}$  is the best position of the population.

- Update the particle's position using the following expression.

$$p_i(t) = p_i(t-1) + v_i(t) \quad (2)$$

Where  $p_i(t)$  is a current position of the particle

$p_i(t-1)$  is a previous position of the particle

$v_i(t)$  is a current speed of the particle

- Update pbest and gbest
- Increment the loop counter

END WHILE

#### 2.4 Multi-Objective Optimization Problem (MOOP)

[3],[7] Multi-Objective Optimization problem has a number of objective functions which are to be minimized or maximized. As in the Single-Objective Optimization problem. The difference between them is the Multi-Objective Optimization problem has at least two objective functions in a problem. We can state the Multi-Objective Optimization problem in the following form

$$\text{Minimize (or maximize)} : \{f_1(x), f_2(x), \dots, f_m(x)\} \quad (3)$$

Where  $x$  is a vector of  $n$  decision variables

$f_i(x)$  is an objective function,  $i = 1, 2, \dots, m$

Finding answer in MOOP can be classified in 3 ways

- To find answers to all of the objective function is minimize.

- To find answers to all of the objective function is maximize.

- To find answers to some objective function is minimize and other is maximize

The answer in MOOP is a set of solutions. Solutions in this set aren't dominated from other solutions. The set of solutions called Pareto Optimal, Non-dominated Set

## 2.5 Benchmark of the Pareto optimal

[3] The answer from MOOP is a set of solutions. A good answer is an answer that closet the real answer (True Pareto Optimal). There are some benchmark use to measure the performance of the answer such as Convergence Distance Measurement, Pareto Spread Measurement etc.

## 2.6 Evolutionary Particle Swarm Optimization (EPSO)

[1] Although, PSO will be able to solve optimization problem as well. However, there are some problems in its algorithm. That is PSO has a convergence to local optimum too early. Because of the inertia weight in PSO movement is a constant value. In the last few years, there are some studies use an evolutionary strategy to improve this drawback. The new algorithm was introduced call Evolutionary Particle Swarm Optimization (EPSO). In EPSO each particle is defined by the following characteristics:

- $x_i^k$  is a position i of k particle
- $v_i^k$  is a velocity i of k particle
- $x_i^{best}$  is a position I of the best particle (gbest)
- $x_i^{k,mem}$  is the best previous position of k particle

The particles will reproduce and evolve along a number of generations, according to the following steps:

- Replication : each particle is replicated a number r of times, giving place to identical particles(in this paper we take r = 1).
- Mutation : the strategic parameters of the replicated particles undergo mutation according to

$$* w_{i,j}^k = w_{i,j}^k + \tau N(0, \sigma^2) \quad (4)$$

Where  $\tau$  is learning dispersion parameter and  $N(0,1)$  is a random number following a the Gaussian distribution with zero mean and variance  $\sigma^2$ .

The strategic parameters are randomly set between 0 and 1 at the beginning of the algorithm. In each iteration, the strategic parameters of the replicated particles are mutated according to equation (1). In this equation, j can be the inertia, memory or the coordination factor.

- Reproduction (movement) : each particle generates as offspring a new particle according to the transformation process, similar to the Classic PSO basic equation :

$$* v_i^k = w_{i,inertia}^k v_i^k + w_{i,mem}^k (x_i^{k,mem} - x_i^k) + w_{i,coop}^k (x_i^{best*} - x_i^k) \quad (5)$$

$$* x_i^k = x_i^k + * v_i^k \quad (6)$$

The offspring is held separately for the original particles and for the mutated particles. Furthermore, instead of defining a crisp best-so-far point as a target the particles are attracted to a sort of “foggy best-so-far region”(another change relative to Classic

PSO). This is done by introducing random noise in the definition of the best-so-far point :

$$x_i^{best*} = x_i^{best} + \tau N(0,1) \quad (7)$$

$\tau$  is a noise dispersion parameter, usually small, and  $N(0,1)$  is a random number following a the normalized Gaussian distribution with zero mean and variance 1.

- Evaluation : each offspring particle plus the originals are evaluated according to their current position.
  - Selection : among the offspring of particle, with and without mutated parameters , a stochastic tournament is played to select the particle that will survive to the next generation.
- The problem of PSO as discussed above and improvement with EPSO, we interested to solve Multi-Objective Optimization with EPSO and discuss it in the next section.

## 2.7 Multi-Objective Evolutionary Particle Swarm Optimization (MOEPSO)

The main objective of this paper is introducing a new algorithm called Multi-Objective Evolutionary Particle Swarm Optimization (MOEPSO) to solve Multi-Objective Optimization Problem. The algorithm of MOEPSO is the following[5,6].

Initialize the population POP

- FOR i = 0 to MAX /\* MAX = number of particle \*/
- Initialize POP[i]

1. Initialize the velocity (speed) of each particle

- FOR I = 0 TO MAX
- VEL[i] = 0

2. Evaluate each of the particles in POP

- Calculate fitness value of each particle from objective function.
- Compare a fitness value and select a particle that has a good fitness value (this paper select minimize).

3. Store the positions of the particles that represent nondominated vectors in the repository REP.

4. Generate hypercube of the search space explored so far, and locate the particles using these hypercube as a coordinate system where each particle's coordinates are defined according to the values of its objective functions.

5. Initialize the memory of each particle (this memory serves as a guide to travel through the search space. This memory is also stored in the repository)

- FOR I = 0 TO MAX

- PBESTS[i] = POP[i]
6. Mutation : the strategic parameters of the replicated particles undergo mutation according to (4)
  7. WHILE maximum number of cycles has not been reached
- DO
- Reproduction (movement) : each particle generates as offspring a new particle according to the transformation process, similar to the Classic PSO basic equation (5),(6)
  - Evaluation: each offspring particle plus the originals are evaluated according to their current position.
  - Selection: among the offspring of particle, with and without mutated parameters , a stochastic tournament is played to select the particle that will survive to the next generation.
- END WHILE

### 3 EXPERIMENT

The proposed algorithm has been compare with Multi-Objective Particle Swarm Optimization (MOPSO) using a Test Problem BNH [7] detail is in the following

Test Problem BNH

Minimize these Objective Functions

$$f_1(x, y) = 4x^2 + 4y^2 \quad (8)$$

$$f_2(x, y) = (x - 5)^2 + (y - 5)^2 \quad (9)$$

Constraint

Subject to

$$g_1(x, y) = (x - 5)^2 + y^2 \leq 25 \quad (10)$$

$$g_2(x, y) = (x - 8)^2 + (y + 3)^2 \geq 7.7 \quad (11)$$

**Table 1.** Experiment's parameter

Parameter	Value	
	MOPSO	MOEPSO
Number of particle	10	10
C1	2.8	2.8
C2	1.3	1.3
w	1 / (2 * 0.301)	Random value between [0,1]
Number of itera-	30	30

tion		
------	--	--

#### 4 RESULT

**Table 2.** Solutions from Algorithms

Iter no.	MOPSO		MOEPSO	
	x	y	x	y
-	5	3	6	4
1	4	4	6	4
2	4	4	6	4
3	4	4	5	5
4	4	4	5	5
5	4	4	5	5
...	4	4	5	5
30	4	4	5	5

It was found that, MOPSO returns the same solution( $x = 4, y = 4$ ) when run into the second iteration while, MOEPSO return the same solution ( $x = 5, y = 5$ ) when run into the fourth iteration. It shows that, MOPSO convergence to local optimum earlier than MOEPSO. Because of some value in MOPSO's movement equation shown as (1) are constant. Then in each iteration a speed of particle is the same value. In contrast, values in MOEPSO's movement equation are random value. As a result, MOEPSO returns diversity value.

#### 5 CONCLUSION

This paper proposes a Multi-Objective Evolutionary Particle Swarm Optimization (MOEPSO) which improved by Evolutionary Algorithm. To solves Multi-Objective Optimization problem the proposed algorithm has been compare with Multi-Objective Particle Swarm Optimization (MOPSO) using a Test Problem BNH results show that MOPSO convergence to local optimum earlier than MOEPSO. In the future, we will improve the performance of MOEPSO and study in a large-scale problem (more than 2 objective functions).

#### 6 REFERENCES

1. Vladimiro Miranda., Nuno Fonseca."NEW EVOLUTIONARY PARTICLE SWARM ALGORITHM (EPSO) APPLIED TO VOLTAGE/VAR CONTROL"
2. Pengsiri Prudtipong.,Sodsee Sunantha.,Meesad Phayung."A Modification of Multi-Objective
3. Ammaruekarat Paranya.,Meesad Phayung."Multi-Objective Optimization using Evolutionary Algorithms",Information Technology Journal,2012.
4. Pengsiri Prudtipong.,Sodsee Sunantha.,Meesad Phayung."A Modification of Multi-Objective Optimization Genetic Algorithm with Initial Population Partition",NCCIT,2015.

5. Thammachantuek Ittikon.,Kongkachandra Rachada."University Course Timetable Planning using Hybrid Evolutionary Particle Swarm Optimization and Constraints-Based Reasoning",ie-network,2011.
6. Reyes-Sierra, Magarita., & Coello Coello, Carlos A."Multi-Objective Particle Swarm Optimizers: A Survey of the State-of-the-Art",International Journal of Computational Intelligence Research,2006.
7. Deb, K.(2008).Multi-Objective Optimization using Evolutionary Algorithms.JOHN WILEY & SONS, LTD.

## Enhancement of Palm-Leaf Manuscript for Segmentation

(Siriya Phattarachairawee)<sup>1</sup>, (Montean Rattanasiriwongwut)<sup>2</sup> (Mahasak Ketcham)<sup>3</sup>

Department of Information Technology Faculty of Information Technology King Mongkut's University of Technology North Bangkok, Thailand

E-mail: william.siriya@gmail.com

E-mail: montean@it.kmutnb.ac.th

E-mail: mahasak.k@it.kmutnb.ac.th

*Abstract* – This paper presents Enhancement of Palm-Leaf Manuscript as noise occurred in images It's suitable for Segmentation. The proposed method consists of image adjustment, contrast stretching and histogram equalization technique. All these techniques improve the quality of images for human viewing. The method we proposed, it made more effective images of palm-leaf manuscripts. The color image of palm-leaf manuscripts became pale and the visible images are not beautiful and hardly colorless because of image enhancement in image processing. In the other hand, the proposed method enhanced clarity of black alphabets. Thus, the color images were transformed into grayscale images affects the clearest alphabets on palm-leaf manuscripts. It also reduces noise. Next, the output images are suitable for carrying out a recognition system.

*Keyword* – Palm-Leaf Manuscript, Adjustment, Stretching, Histogram

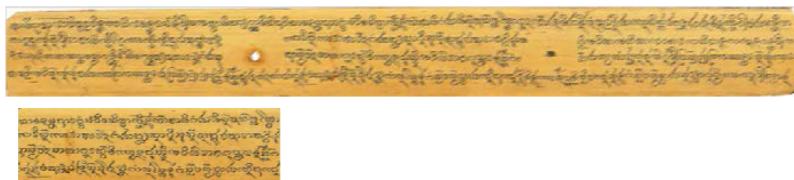
### 1. Introduction

Palm-leaf manuscripts were used as writing materials for recording of events and stories, such as the principles of Buddhism, historical events, astronomy, literatures, dead chronicle, ancient legal histories, pharmacopeia etc. Writing or also called as “inscribe”, the inscribed alphabets on palm-leaf manuscripts are different style based on current nationality and languages.



Fig. 1 Deterioration of Palm-Leaf Manuscripts

When the time has passed, palm-leaf manuscripts might be deterioration and outdated paper are varying contrast. Up to now, image processing offers a selection of approaches to preserve the manuscripts readable. Those manuscripts have gone through the preservation process using computer technology. [6]



**Fig. 2.** Palm-Leaf Manuscript is unclear.

The quality of images from preservation process is often poor, it makes a lot of mistakes in palm-leaf manuscript recognition. There are several reason that cause the faults, such as state on image of palm-leaf manuscripts, unequal visual images, smudges on Isan alphabets of palm-leaf manuscripts, the paleness of the ink, quality of flatbed scanner, quality of image etc.

In this paper proposed 3 models consist of normalization techniques for palm-leaf manuscript is model 1 [2] - [4] Model 2 is background enhancement of palm-leaf manuscript [5] - [6] Model 3 is contrast enhancement technique for palm-leaf manuscript [2]

As noted above, we have proposed the methods to improve the quality of palm-leaf manuscripts image, [18-19] before input the image of manuscripts into recognition process and to increase accuracy percentage scores of recognition on palm-leaf manuscripts. Previous studies have proposed image enhancement using several techniques. We applied the techniques from them to develop image enhancement of palm-leaf manuscripts using image adjustment, contrast stretching and histogram. The expected results is to enhance images on palm-leaf manuscripts better than original images and to recognize the information from palm-leaf manuscripts.

## 2. Theories

In this paper, we proposed the methods for enhancement of palm-leaf manuscript, including adjustment, contrast stretching and grayscale images.

**2.1 Adjustment** [8] [10] is a color saturation adjustment method for color adjustment which can effectively prevent color-saturation and make the image look natural. It indicates that the color saturation has certain regularity relations with brightness and also reveals the efficacy of image contrast enhancement method.

**2.2 Contrast Stretching** [11] is an image enhancement technique that attempts to improve the contrast in images by stretching the range of intensity values

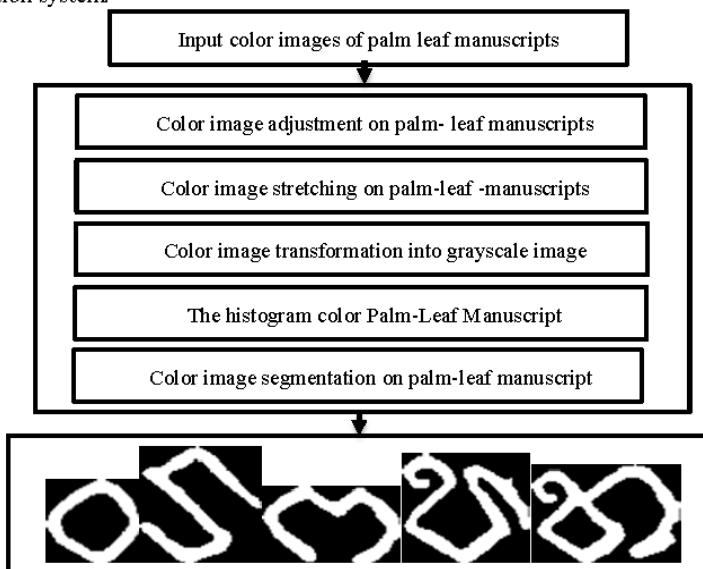
and it also replaced the values of original intensity. By the way, Adaptive Transfer Function will provide the values of new intensity that is designed by using an input image with basic statistics.

**2.3 Gray Scale Image** [20-21] is images transformations into gray scale, due to grayscale images of luminous intensity is between 0-255. It is easy to create characters and background segmentation. The calculation is done by separation of the color intensity with the pixels in RGB mode.

**2.4 Histogram Equalization** [12]-[14] is one of image enhancement method. It is contrast image enhancement using histogram to distribute the intensity values that appeared in images. Intensity values of images are from computation, which affects uniform distribution of histogram values. [15] [11] Output of image processing is a complete images without the mistake such as colorless of images. [4] HE technique was used for distribution the contrast of images on palm-leaf manuscripts.

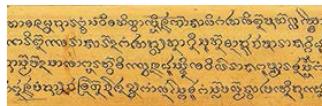
### 3. The Proposed Method

Enhancement of palm-leaf manuscript is the preparation of a set of images to eliminate unwanted features and only preserves the quality features. Initially, we input the original images using flatbed scanner. All these images were set a display resolution of 360 pixels with the end in JPG. Size of image segmentation is 4×2 inches. Next, adjustment, contrast stretching, grayscale image transformations and histogram techniques were carried out as image processing. Noises also were eliminated by image processing. Finally, the output images are suitable for carrying out a recognition system.



**Fig. 3.** The proposed method

**3.1 The images that used in the experiment** were set the resolution values as 300 dpi with ends in .JPG. Initially, we input color images of palm leaf manuscripts as Figure 4.



**Fig. 4.** Input color images of palm leaf manuscripts

### 3.2 Enhancement of palm-leaf manuscripts

3.2.1 Image adjustment on palm-leaf manuscripts were adjusted for high resolution effect more clarity of image as Figure 5.



**Fig. 5.** Color image adjustment on palm- leaf manuscripts

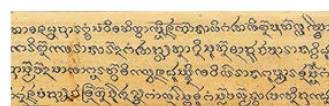
3.2.2 Contrast stretching is information overlay that is designed to overlay the original intensity values. This method manages the replacing of new intensity values onto all pixels and it also eliminate noises in image enhancement as equation 1.

$$b(m,n) = 255 \frac{a(m,n) - \text{Min}(m,n)}{\text{Max}(m,n) - \text{Min}(m,n)} \quad (1)$$

When

- |                    |  |
|--------------------|--|
| $b(m, n)$          | = Luminance values after image enhancement |
| $a(m, n)$          | = Luminance values                         |
| $\text{Max}(m, n)$ | = Maximum pixel values                     |
| $\text{Min}(m, n)$ | = Minimum pixel value                      |

From the equation above, this method manages the replacing of new intensity values onto all pixels and it also eliminate noises in image enhancement as Figure 6



**Fig. 6.** Color image stretching on palm-leaf manuscripts

3.2.3 Color image transformation into grayscale image, the scanning image is color (RGB). Firstly, it have to transform into grayscale image before leading to the next process by segmentation each pixel of palm-leaf manuscript apart in the color scheme RGB Into the equation to calculate the gray and bring value to the original pixels. Finally, the result is the equation 2-3.

$$f_{gr} = 0.3R + 0.59G + 0.11B \quad (2)$$

or

$$f_{gr} = \frac{R+G+B}{3} \quad (3)$$

When

$f_{gr}$  = The gray values

R = The red values

G = The green values

B = The blue values

Then,  $f_{gr}$  values are substituted in original pixels.

The results of color image transformation into grayscale image **Figure 7.**

**Fig. 7.** Color image transformation into grayscale image

3.2.4 Histogram equalization is image enhancement using the method of creating a neighbor pixel or transformed image. The method is a determinant of pixels in each of pixel values. Palm-leaf manuscripts which are grayscale images as the fourth equation, effected to **Fig.8**

$$S_k = T(r_k) = \sum_{i=0}^k \frac{n_i}{N} \quad (4)$$

When

$S$  = Output intensity

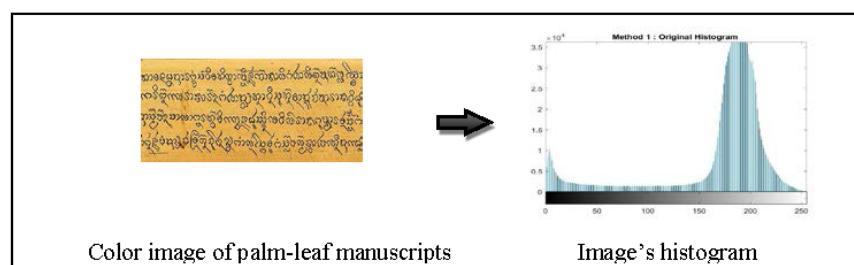
$n_i$  = The number of pixels intensity values as  $i$

$N$  = The total number of pixels



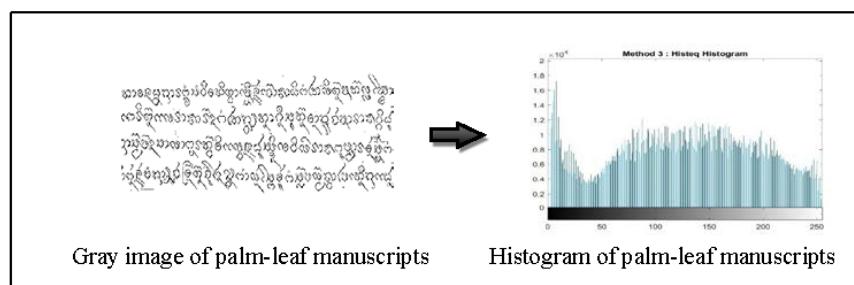
**Fig. 8** Palm-leaf manuscript for creating histogram

Input the original image into histogram technique **Figure 9.**



**Fig. 9** shows histogram of palm-leaf manuscripts image

Color image of palm-leaf manuscript using histogram equalization depend on statistical calculations with probability distribution of gray level as **Figure 10**



**Fig. 10** shows image histogram from histogram equalization technique.

**3.3 Segmentation** [15-17] is one of important methods that isolate one word from another and separate the various letters of a word. It affects the latter phase of the character segmentation and recognition directly. During image enhancement technique is applied to solve the problems such as noise, low contrast, shadow etc. A number of 30 images are from scanning image onto computer with they end in JPG and a set of image is a display resolution of 360 pixels with ends in JPG.

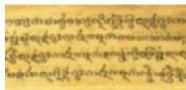
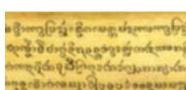
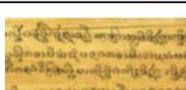
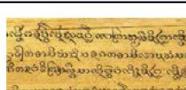
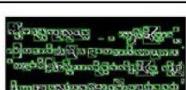
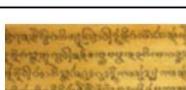
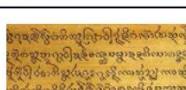
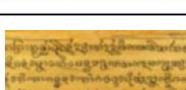
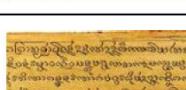
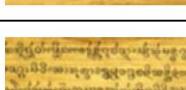
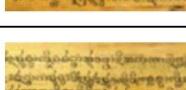
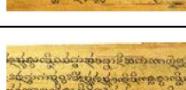
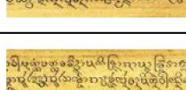
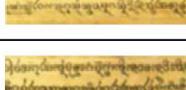
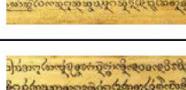
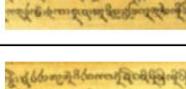
#### 4. The Experimental Results

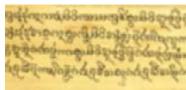
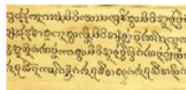
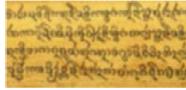
We have applied our proposed methods on a set of palm-leaf manuscripts which includes 30 original images.

**Table 1** The following table shows the performance of image enhancement.

No.	Original	Enhancement	Segmentation
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			

No.	Original	Enhancement	Segmentation
9.			
10.			
11.			
12.			
13.			
14.			
15.			
16.			
17.			
18.			

No.	Original	Enhancement	Segmentation
19.			
20.			
21.			
22.			
23.			
24.			
25.			
26.			
27.			
28.			

No.	Original	Enhancement	Segmentation
29.			
30.			

From Table 1, it revealed that the alphabets in palm-leaf manuscripts image is unclear, dhamma character on palm-leaf manuscripts fade away and unreadable. For these reasons, we have presented the method for image enhancement on palm-leaf manuscripts to increase the quality of image. Palm-leaf manuscript image became clarity and readability. Therefore, image enhancement made us found that dhamma character on palm-leaf manuscripts can be made more clearly by segmentation technique.

The method we proposed, it made more effective images of palm-leaf manuscripts. The color image of palm-leaf manuscripts became pale and the visible images are not beautiful and hardly colorless because of image enhancement in image processing. In the other hand, the proposed method enhanced clarity of black alphabets. Thus, the color images were trans-formed into grayscale images affects the clearest alphabets on palm-leaf manuscripts. It also reduces noise. As the experimental result, the proposed method is suitable for characters segmentation.

## 5. Conclusion

In conclusion, a set of images that used in image processing were color image of Isan dhamma character on palm-leaf manuscripts with size of  $4 \times 2$  inches, amount 30 images. Only the alphabets without black color of background have been required to recognition systems.

As recognition system, the palm-leaf manuscripts are the most important, due to a quality of palm-leaf manuscripts affects to efficacy of recognition. In this case, if manuscripts have a lot of noise, varying contrast, dirty background or outdated paper, they affects to low level recognition. On the other hand, if noises were eliminated by proposed techniques and image enhancement on palm leaf manuscripts are complete, they affect to high level recognition. At first, we should enhance the images using adjustment, contrast stretching and histogram technique. It is hardly seems to look difference in good original image. If original images are poorly output, our proposed method can enhance the image as best as possible. However, improved images were difficult to transform into color images with clear alphabets. When the improved images were adjusted into color images, they were hardly colorless. Thus, the color images were transformed into grayscale images affects the clearest alphabets

on palm-leaf manuscripts. It also reduces noise. Next, the output images are suitable for carrying out a recognition system. Other than image processing using three proposed methods, speeding up the image enhancement on palm-leaf manuscripts makes more accuracy in process of image recognition. Our proposed methods can be applied and improved to business. It might also be guideline for further studies in a field of image enhancement on palm-leaf manuscripts.

## 6. References

- [1] Chamchong R. and Surinta O. (2007). Text Segmentation from Palm Leaf Manuscripts. Doctoral dissertation, Computer science, Management Information System, Graduate School, Mahasarakham University.
- [2] Shi Zh., Setlur S. and Govindaraju V. (2007). Digital Enhancement of Palm Leaf Manuscript Image using Normalization Techniques, Center of Excellence for Document Analysis and Recognition (CEDAR), State University of New York at Buffalo, Amherst, USA.
- [3] Cherala S. and Rege P.P. "Palm leaf manuscript/color document image enhancement by using improved adaptive binarization method." in 6th Indian Conference on Computer Vision, Graphics & Image Processing. 2008: 687-692.
- [4] Medithi R., Prasad N. V. G., N. Rao V. R. "Palm Leaf Manu Script Document Enhancement by Combined Binarization and Normalization Method." International Journal of Engineering Research & Technology (IJERT). Vol. 2 No.1 (2013) : 1-8.
- [5] Rege P. P. and Chiddarwar A. S. "Enhancement of Palm-leaf Manuscript and color document images with synthetic background generation." Journal of Advances in Engineering Science. Vol.2 No.2 (2008) : 25-34.
- [6] Yahya S. R., et al. (2010). "Review on Image Enhancement Methods of Old Manuscript with Damaged Background." International Journal on Electrical Engineering and Informatics. Vol. 2 No.1 (2009) : 1-13.
- [7] Chiddarwar, S. A. and Rege, P. P. "Contrast Based Enhancement of Palm-leaf Manuscript Images." in Second International Conference on Computer Engineering and Applications (ICCEA). Vol.1, 2010: 219-223.
- [8] Ke W.-M., Chen C.-R., and Chiu C.-T. "BiTA/SWCE: Image Enhancement with Bilateral Tone Adjustment and Saliency Weighted Contrast Enhancement" Journal of Transactions on Circuits and Systems for Video Technology. Vol. 21 No. 3 (2011): 360-364
- [9] Chiang J-H., et al. "Color Image Enhancement with Saturation Adjustment Method." Journal of Applied Science and Engineering, Vol. 17, No. 4 (2011) : 341-352
- [10] Kwok N., et al. "Adaptive Scale Adjustment Design of Unsharp Masking Filters for Image Contrast Enhancement." in International Conference on Machine Learning and Cybernetics. Kensington: The University of New South Wales Press, Vol.2, 2013: 884-889.

- [11] Ye Y., et al. "On Linear and Nonlinear Processing of Underwater, Ground, Aerial and Satellite Images." in 2005 IEEE International Conference on Systems, Man and Cybernetics. Shenyang, Liaoning province: Shenyang Press, Vol. 4, 2005: 3364 – 3368
- [12] Yeganeh H., Ziae A., and Rezaie A. "A novel approach for contrast enhancement based on Histogram Equalization." in International Conference on Computer and Communication Engineering. Tehran: Amirkabir University of Technology Press, 2008 : 256 – 260
- [13] Yun S.-H., Kim J. H., and Kim S. "Image Enhancement using a Fusion Framework of Histogram Equalization and Laplacian Pyramid." Journal of Transactions on Consumer Electronics. Vol.56 No. 4 (2010) : 2763-2771.
- [14] Al-Wadud M. A., et al. "A Dynamic Histogram Equalization for Image Contrast Enhancement." Journal of Transactions on Consumer Electronics. Vol. 53 No. 2 (2007) : 593-600.
- [15] K. Arulmozhi, et al. "Image Enhancement Techniques on Indian License Plate Localized Image for Improved Character Segmentation." in International Conference on Computational Intelligence and Computing Research. India : M.S. University Press, 2012: 1-6.
- [16] Chamchong R and Fung C.C. "A Combined Method of Segmentation for Connected Handwritten on Palm Leaf Manuscripts" in International Conference on Systems, Man, and Cybernetics (SMC), 2012: 4158-4161.
- [17] Chamchong R and Fung C.C. "Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand" in International Conference on Frontiers in Handwriting Recognition, 2012: 588 – 593.
- [18] Inkeaw P. et al. "Lanna Dharma Handwritten Character Recognition on Palm Leaves Manuscript based on Wavelet Transform" in International Conference on Signal and Image Processing Applications (ICSIPA), 2015: 253 – 258.
- [19] Kumar S. N., et al. "Ancient Indian Document Analysis using Cognitive Memory Network" in International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014: 2665 – 2668.
- [20] Chamchong R and Fung C.C. "Comparing Background Elimination Approaches for Processing of Ancient Thai Manuscripts on Palm Leaves" in Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, 2009: 3436 – 3441.
- [21] Chamchong R , et al. "Generation of optimal binarisation output from ancient Thai manuscripts on palm leaves" Proceedings of International Conference on Machine Learning and Cybernetics,2013: 1643 - 1648

## Comparison of Edge Detection Algorithms for Coastline Detection in Satellite Imagerys

Chutiwan Boonarchatong<sup>1</sup>, Sucha Smanchat<sup>2</sup>, Mahasak Ketcham<sup>2</sup>, and Nawaporn Wisitpongphan<sup>2</sup>

<sup>1</sup> Department of Information Technology, Suan Dusit University, Thailand  
chutiwan\_boo@dusit.ac.th

<sup>2</sup> Department of Information Technology, King Mongkut's University of Technology  
North Bangkok, Thailand  
ssmanchat@acm.org, {mahasak.k, nawaporn.w}@it.kmutnb.ac.th

**Abstract.** Finding the edge of a coastline is one of the crucial tasks in monitoring the coastline erosion which is an indicator of ecological change. The goal of this work is to find a suitable edge detection operator. In this work, the dataset was derived from the raw satellite imageries, namely THEOS. The result shows that Canny and Laplacian of Gaussian are the best detector in both less noise and Medium noise when compared with Robert, Sobel, Prewitt, and Laplacian of Gaussian detection algorithms.

**Keywords:** edge detection • coastline • satellite imagery

### 1 Introduction

Thailand coasts govern several land cover forms such as a sandy beach, and mangroves which house marine breeding, sea grass beds, and coral reefs. On the contrary, the coastal erosion affects the national economy and also causes ecological imbalance. Therefore, the coast should be monitored for any sign of erosion which can later be extended to formulate coastal master plans to protect the coast [1]. One popular approach to monitor the coastal erosion features is the use of satellite imageries, where the imageries' snapshot can be repeatedly accumulated at the same point over the period of time.

The classical edge detection algorithms are divided into 2 main groups; first: gradient based, Sobel, Canny, Prewitt and Robert, and second: zero-crossing based, and Laplacian of Gaussian [2]. In order to acquire the optimize result, the most suitable edge detection algorithms which can best extract the coastal imageries in the presence of noise shall be determined. Thus in this research, we aim to test the satellite imageries of coastline of various landscapes using different edge detection algorithms in order to detect Thailand's coastline with THEOS imageries.

## 2 Method

The imageries obtained from THEOS satellite are processed with different edge detection algorithms. They are tested performance against multiple metrics; processing time, PSNR and MSE were evaluated to locate the best edge detection algorithms.

### 2.1 Edge Detection Algorithms

The gradient-based methods, Sobel, Canny, Prewitt, Robert, and Laplacian of Gaussian algorithms are based on computing local gradient, brightness pixel, in pair of orthogonal direction.

- **Roberts Algorithm**; Robert is similar to Sobel with an exception that the kernel is rotated by 90 degree [3]. It is sensitive to noise, since it is computed in different direction [4, 5]. Moreover if the pixel quality is not sharp, the response will be fairly low [5].
- **Sobel Algorithm**; Sobel performs better than the others gradient based algorithms unless the image data contains noise [6]. The trouble can be seen when there is noise in data such as medical image, ultra sound image of tumour or bone [4].
- **Canny Algorithm**; Canny is the most popular and dominant algorithm in image edge detection [6]. Canny's process is more complicate than Sobel. It is suitable for image containing noise [7-10]. Canny can cope well with noisy image, however, its operating time is longer than Roberts, Prewitt and Sobel when processing image with noise [6].
- **Prewitt Algorithm**; Prewitt is similar to Sobel [11]. It was developed from Sobel. Noise detected is smoothly illustrated in grayscale where the accurate of edge detection algorithm can easily be increased [12]. Though noise can easily be removed by this method, the algorithm is still sensitive to noise, especially images that contain random noises [13].
- **Laplacian of Gaussian**; Laplacian is one of the most popular algorithm in image edge detection although it is sensitive to noise. However, such weakness can be fulfilled by Gaussian transformation [14]. Using Laplacian of Gaussian, the noise can be filtered out, thus, sensitivity of its algorithm is possibly protected [3]. The process of Gaussian works in such a way that noise is filtered with Gaussian Smoothing filter and then Laplacian can correctly detect edge [14].

### 2.2 Performance analysis

The performance of imagery edge detections is measured from image quantitative methods consisting of operating time, Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE). Both of PSNR and MSE are used to measure the noise which is removed from the images. Since PSNR generally represents a peak error measurement, as a result, if MSE is low, the error will also be low.

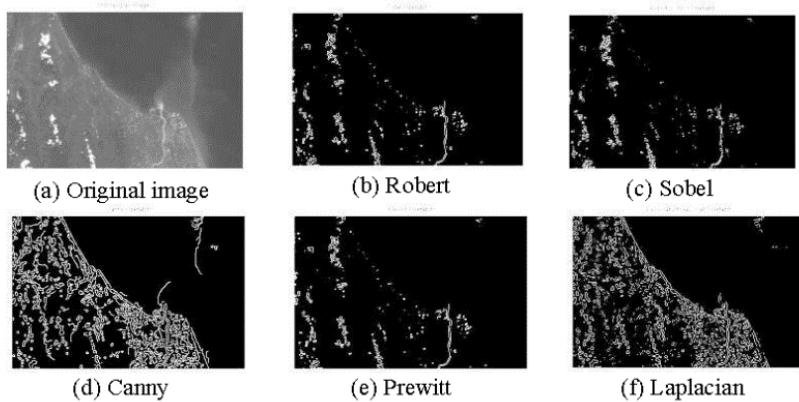
### 3. Dataset

Prior to conducting the performance comparison, we have to prepare a reliable dataset and find out a suitable tool which can run different type of edge detecting algorithms. These imageries are taken from THEOS satellite with TOP2 sensor on MS sensor mode. They are acquired from GISTDA [15], from 2014 to 2015. The imageries are separated into 3 main groups by criteria noise, percentage of cloud coverage.

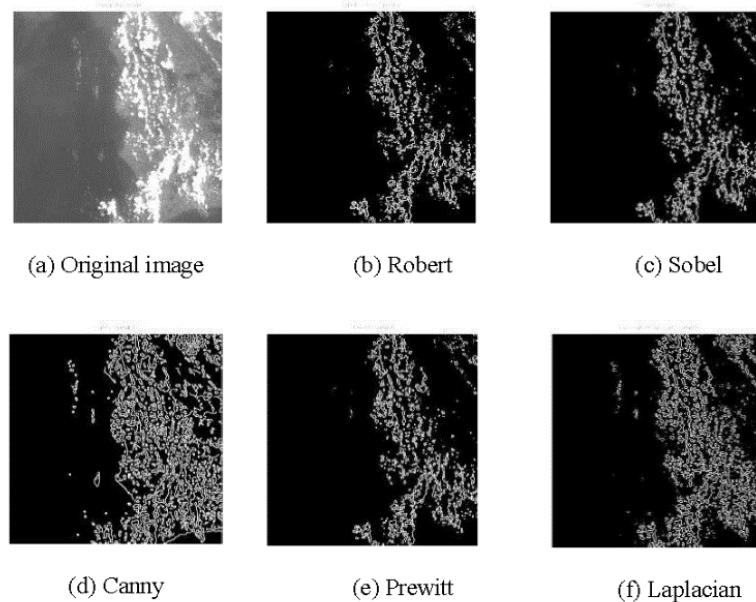
- **Less noise (sparsely patched shape clouds);** Overall clouds coverages are less than 10%. The appearances of cloud look like tiny sheet or patch. Clouds are in high level in the sky called Cirrus, Cirrostratus, and Cirrocumulus.
- **Medium noise (patch shape clouds);** Overall cloud coverage is around 40%. Their features are thick sheet or moderate size patch. As a result, raw satellite imageries are not clear.
- **Medium noise (round shape clouds);** Overall cloud coverage is around than 40% with round masses. These clouds float at low level: Cumulus and Cumulonimbus. They are naturally situated above the land rather than the sea so the coastal imageries are still obvious.

### 4. Result

Canny illustrated the best performance among all, following by Laplacian in the scenario where there was less noise coast edge, but their performances are quite poor with medium noise. Partial results are shown in fig. 1 and fig. 2. Both also show maximum PSNR and minimum of MSE. In another word, they represent a minimum failure to detect the original image as shows the comparison in table 1. Although Canny and Laplacian acquire the best result, they are the most time-consuming algorithm for operating, according to the experimental results stated in [6-10].



**Fig. 1.** Less noise coastal imagery (Narathiwat)

**Fig. 2.** Medium coastal imagery (Chon Buri)**Table 1.** Comparison of algorithms performance the best results

Noise Types & Imageries name	PSNR		MSE		Elapsed Time (s)	
	1 <sup>st</sup> Best	2 <sup>nd</sup> Best	1 <sup>st</sup> Best	2 <sup>nd</sup> Best		
<b>Less Noise</b>						
Narathiwat	Canny 26.6633	Laplacian 26.5477	Canny 141.3009	Laplacian 145.115	Canny 3.5516	Laplacian 2.4532
Nakhon	Canny 26.7226	Laplacian 26.4737	Canny 139.3847	Laplacian 147.6071	Canny 0.6821	Laplacian 0.4594
<b>Medium (Patch shape)</b>						
Samuth Prakan	Canny 26.8212	Laplacian 26.495	Canny 136.2579	Laplacian 146.8849	Canny 0.2099	Laplacian 0.1514
Pet Buri	Canny 27.0087	Laplacian 26.759	Canny 143.4297	Laplacian 138.2222	Canny 0.4641	Laplacian 0.3451
<b>Medium (Round)</b>						
Chon Buri	Canny 26.3838	Laplacian 26.243	Canny 150.6969	Laplacian 155.6594	Canny 0.2099	Laplacian 0.1514
Trad	Canny 26.4187	Laplacian 26.3069	Canny 149.4905	Laplacian 153.3889	Canny 2.2030	Laplacian 0.6130

## 5. Conclusion and Discussion

This research was investigated to determine the best algorithm to detect the coastline erosion by comparing the performance of various edge detection algorithms using satellite imageries. The imageries are delivered from THEOS satellite. Canny and Laplacian of Gaussian algorithms are recommended since they produce the best result with the less MSE value and more PSNR value. However, the performance was traded-off with process time since they are the most time consuming processed. Despite Canny and Laplacian of Gaussian produce the best result to detect the costal imageries, they do not work well with satellite imageries with medium and thick noise. As a result, Canny and Laplacian of Gaussian algorithms for detecting edge or boundary of medium noise imageries remains an interesting gap for improvement.

## References

1. Marine and Coastal Resources Research Center, L.G.o.T., Ko Sanea Cheng Na Yo By Naw Tang Kan Phun Phu Kan Chai Pra Yot Had Thray Lae Kan A Nu Rak (Policy Recommendations Measures for Rehabilitation and Conservation of Beach Uses). 2013.
2. Patel, J., et al., Fuzzy inference based edge detection system using Sobel and Laplacian of Gaussian operators, in Proceedings of the International Conference &#38; Workshop on Emerging Trends in Technology. 2011, ACM: Mumbai, Maharashtra, India. p. 694-697.
3. Raman Maini, H.A., Study and Comparison of Various Image Edge Detection Techniques. International Journal of Image Processing (IJIP), 2009. 9(3): p. 1-11.
4. Caixia, D., M. Weifeng, and Y. Yin. An edge detection approach of image fusion based on improved Sobel operator. in Image and Signal Processing (CISP), 2011 4th International Congress on. 2011.
5. Jamil, N. and T. Sembok, Gradient-Based Edge Detection of Songket Motifs, in Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access, T. Sembok, et al., Editors. 2003, Springer Berlin Heidelberg. p. 456-467.
6. Kelefouras, V., A. Kritikakou, and C. Goutis, A methodology for speeding up edge and line detection algorithms focusing on memory architecture utilization. The Journal of Supercomputing, 2014. 68(1): p. 459-487.
7. Bing, W. and F. Shaosheng. An Improved CANNY Edge Detection Algorithm. in Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on. 2009.
8. Geng, X., K. Chen, and X. Hu. An improved Canny edge detection algorithm for color image. in Industrial Informatics (INDIN), 2012 10th IEEE International Conference on. 2012.
9. Cai-Xia, D., W. Gui-Bin, and Y. Xin-Rui. Image edge detection algorithm based on improved Canny operator. in Wavelet Analysis and Pattern Recognition (ICWAPR), 2013 International Conference on. 2013.
10. Guo, L. and J. Nan. Canny edge detection algorithm based on wavelet transform and RAMF. in Computational Problem-Solving (ICCP), 2010 International Conference on. 2010.
11. Jose, A., et al. Performance study of edge detection operators. in Embedded Systems (ICES), 2014 International Conference on. 2014.

12. Wenshuo, G., et al. Based on soft-threshold wavelet de-noising combining with Prewitt operator edge detection algorithm. in Education Technology and Computer (ICETC), 2010 2nd International Conference on. 2010.
13. Lei, Y., et al. An improved Prewitt algorithm for edge detection based on noised image. in Image and Signal Processing (CISP), 2011 4th International Congress on. 2011.
14. Xin, W., Laplacian Operator-Based Edge Detectors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007. 29(5): p. 886-890.
15. GISTDA. [cited 2015 30 April]; Available from: <http://www.gistda.or.th>.

## Real-time Snoring Sound Detecting U Shape Pillow System using Data Analysis Algorithm

Patiyuth Pramkeaw<sup>1</sup>, Penpitchaya Lertritchai<sup>2</sup>, Nipaporn Klangsakulpoontawee<sup>3</sup>

Department of Media Technology, King Mongkut's University of Technology  
Thonburi, Thailand

patiyuth.pra@kmutt.ac.th, bewwylert@gmail.com, mmayo643@gmail.com

**Abstract.** This paper aims to design and build snoring sound detecting u-shape pillow. Research operating includes four steps as (1) to study the problem of snoring, (2) to analyze the related information for develop the snoring sound detecting u shape pillow with designing the structure and sensory circuit inside the pillow. The u-shape has been designed as a neck supporter for user. The main part of project is the module having microphones that receive a sound of snoring. When a snoring sound was detected, the module will command the vibrating motor to work and alert the user to change his/her body posture. This change will help user stopping a snoring, which controlling by the C language programs, (3) to assess the quality of pillow detect snoring by five experts, (4) as result shown, the proposed pillow can detect snoring sound at 80% of accuracy based on testing with three different users.

**Keywords:** Snoring sound, Neck pillow, Detect snoring sound

### 1 Introduction

Snoring is a problem that is commonly found in people between 30-35 years old, often occurs to males rather than females, and basically increases with ages. It can be simple snoring which has no harms but social effects as well as impacts on others' quality of life, especially bedfellows', owing to its disturbing sounds. The other type of snoring that can happen is snoring with obstructive sleep apnea. Patients of this type normally face dogsleep and waking up frightened intermittently, resulting in poor sleep. These patients, therefore, tend to work inefficiently and may have traffic accidents on account of drowsy driving [1]. Or even those who have this state while working with machines in factories can also have a high risk of dangers.

However, most people usually overlook snoring as they think it is a normal symptom which causes no severe perils. In fact, if they let snoring continue without treatment [2], it will lead to unpleasant situations with their bedfellows or people around them. To make matters worse, it might bring about jeopardies to the daily life, work performance, health, and risks of any other diseases to death [3].

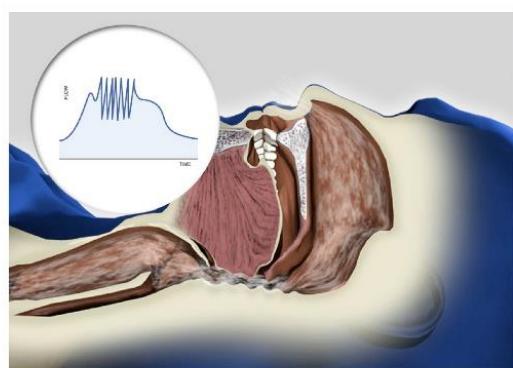
According to the background and the significance of the mentioned problems, the research were interested in the technology to solve these problems. Relevant principles as well as theories were explored in order to create a “sensor pillow” or a “snoring sensor pillow.” The main objective was to relieve a patient with snoring symptom by applying the microphone module to detect his/her snoring sound. After the sound is detected, the vibrating motor will vibrate the pillow so as to motivate the consciousness of the sleeper and stop his/her snoring. The pillow is easy to use without electricity that regularly possesses high risks and hazards. It is also small and portable, and thus suitable for a patient with snoring problem and has to travel a lot.

## 2 Snoring

Snoring is a state or a symptom with abnormal volumes of breathing sounds during sleep. It is one of the symptoms that normally happen to people at all ages from children to the elderly. The older they are, the more symptom increases. Several studies revealed that the symptom was found in males at 24-30% whereas approximately 15% in females. And when they reach the ages of 60-65, the symptom raises up to 60% in males and 40% in females [2].

### 2.1 Primary Snoring

Primary Snoring do not badly affect health. It merely causes annoyance to people nearby and partial airway obstruction. That is because while you are sleeping, it is the time for muscle relaxation, including pharynx. Your tongue and uvula fall behind, particularly when you lie on your back [2]. The airway or intrarespiratory at this part becomes narrower. So, when you breathe in through this narrow position, your uvula and soft palate or root of tongue are vibrated. This is how snoring occurs.



**Fig. 1.** Shows Primary Snoring

## 2.2 Obstructive Sleep Apnea: OSA

The symptom takes place due to the extremely narrow airway, possibly because of very narrow pharynx. For example, there is soft palate tissue, uvula, or huge flabby root of tongue; or enlarged tonsil gland that obstructs pharynx. Some people have little face bones or molars, so the back of their airways are narrower than usual, including those who have shorter chins, of which tongues fall behind deeper than normal people. These group of patient basically produce inconsistent snoring sounds. To clarify, they project both loud and soft snoring sounds intermittently, and they will keep snoring more loudly. Then, they will stop snoring for a moment or it is call "obstructive sleep apnea." It is regarded as a hazardous moment because oxygen levels drop down, acting upon malfunctions of some of their organs such as lungs, hearts, and brains. Their bodies will respond to this state automatically afterwards, that is, their brains are stimulated and thus subjected patients are awakened from sleep to regain their breathing [2]. They usually wake up frightened like being taken aback or chocking on their own saliva. Or they might breathe hard to get oxygen as if to recover from suffocation. Soon after, their brains will fall asleep again, and their breathing is obstructed again. Then, the brains need to be aroused once more, and their sleep is impeded again, too. Such incident keeps circulating repeatedly. As a result, the sound sleep of patients with inconsistent snoring is relatively deficient [5]. They, therefore, always wake up with the feeling of insufficient sleep despite fine numbers of sleeping hours, and it also leads to damages of health, especially their hearts, blood circulation system, lungs, and brains.

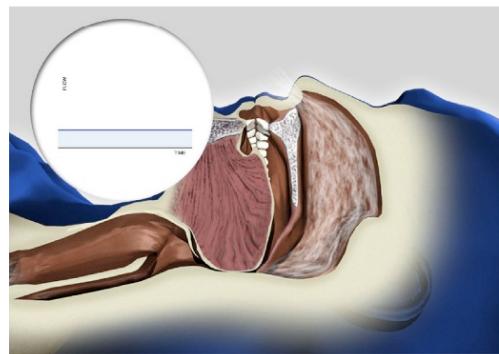


Fig. 2. Obstructive Sleep Apnea

## 2.3 A kind of snoring sound

When concentrating on recording with tools or recording equipment from various reports, the researcher found that snoring sounds can be divided into some major characteristics, i.e., sound properties (e.g., loudness and frequencies); simple snoring sounds and obstructive sleep apnea sounds; and snoring sounds from different sources

--- from soft palate and root of tongue. The collected information helps noticing 8 characteristics of snoring sounds as per below:

- 1) Frequencies of snoring sounds are between 2 up to 50 Hz, and the loudness can be 20 dB up to 85 dB.
- 2) Snoring sounds of patients with simple snoring are with fundamental frequencies and harmonic pattern, and their snoring sounds are higher than 150 Hz with wide bandwidths.
- 3) Snoring sounds of patients with obstructive sleep apnea are not harmonic pattern. The sounds are with low frequencies (< 150 Hz) with narrow bandwidths.
- 4) Snoring sounds from soft palate are in the well-organized wave form. Waves are repeated every 10-30 milliseconds. Their peak frequencies are low (285 Hz on average), and often found more than from root of tongue.
- 5) Snoring sounds from root of tongue initiate disorganized wave form, with high peak frequencies (885 Hz on average).
- 6) Snoring sounds from soft palate are like flapping noises whereas those from root of tongue are like stridor.
- 7) Flexible nasopharyngoscopes can be manipulated to investigate any specific spots of the vibrating upper airway that causes snoring sounds while patients are sleeping or under sleeping pills. This is why the diagnosis is called “sleep nasendoscopy.”
- 8) Very loud snoring sounds and constant pauses relate to sleep apnea. And when patients breathe again, the sounds emerge again as well. This is mostly found in patients with obstructive sleep apnea.

**Table 1.** A kind of snoring sound and Frequency (Hertz)

A kind of snoring sound	Frequency (Hertz)
Primary snoring	> 150 Hertz
Obstructive Sleep Apnea	< 150 Hertz
Snoring sounds from soft palate	285 Hertz
Snoring sounds from root of tongue	885 Hertz

### 3 Literature Review

#### 3.1 Sensor Pillow System [12]

“Sensor pillow system” embraces the development of diagnosis system of sleep disorders among paralyzed patients. The system basically examines hearts, respiratory system, and reactions during sleep. They are all up to polysomnography and intimate

care from physicians. Zigbee or GSM polysomnography will record brain waves, oxygen levels in blood, and pulse rates from body movements. They can be measured by using sensor pillows which comprise of the management of FSR sensors, Zigbee or GSM which emits analog signal. The entire management of FSR sensors, respiratory system sensors, Zigbee, or GSM with Visual Basic (VB) software can be utilized to check/diagnose sleep disorders, pulse rates, and blood pressures.

### **3.2 Short-term outcomes of transoral radiofrequency somnoplasty treatment for snoring [9]**

For the results of short-term treatment of snoring symptom through oral radiofrequency (RF), it will be elaborated next. This research gathered the data/information of patients with snoring who were treated by oral radiofrequency from Department of Otolaryngology, Maharaj Nakorn Chiang Mai Hospital. The results of polysomnography were also embraced. After that, the data was analyzed and compare between pre-treatment and post-treatment (after 3 months of treatment). The results indicated that there were 34 patients, 24 of them were males. Mean of the violence of snoring sounds before treatment was 7.0, and down to 3.6 after treatment (reduced by 48.7%). Mean of daytime drowsiness before treatment was 8.4, and dropped down to 5.4 after treatment. To conclude, the application of oral radiofrequency could efficiently diminish the violence of snoring sounds as well as daytime drowsiness. It also generated less pain to wounds, with only minor mild complications. Thus, this can be one of the great alternatives to cure patients with this problem.

## **4 Methodology**

### **4.1 Sensor and System**

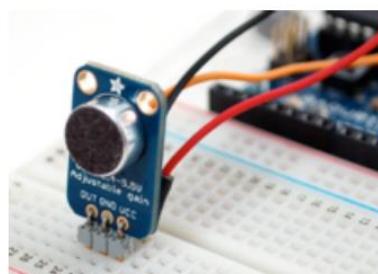
Fig.3, presented the sound sensor module as an electret microphone amplifier with frequencies between 20 Hz to 20 KHz. The module also had the extended circuit with IC MAX4466, which was a low-noise microphone amplifier that could cut off all noises [6]. A tiny trimmer pot was included and could adjust the gain from 25x to 125x which produced the value around 200 MVpp, and possessed an intrinsic ability to cut noises. It was used to detect snoring sounds, with rail-to-rail outputs, and was connected to a 5-volt DC power supply [7]. The noise was projected out of output's legs with DC Bias at VCC2. The motor was vibrated to stimulate the consciousness of snoring patients so that they would stop snoring as shown in Fig.4, and the frequency response was set at 0.8Hz-250Hz Vibration Motor ERMS as show in Fig.5.

Fig.6. Shows the location of two sensors and vibration motors inside the pillow. Vibration motor ERMS 1 is located under the neck in order to detect vibration caused by snoring sound [4]. Sound sensor module are located in below position of the pillow to detect the sounds [9, 11].

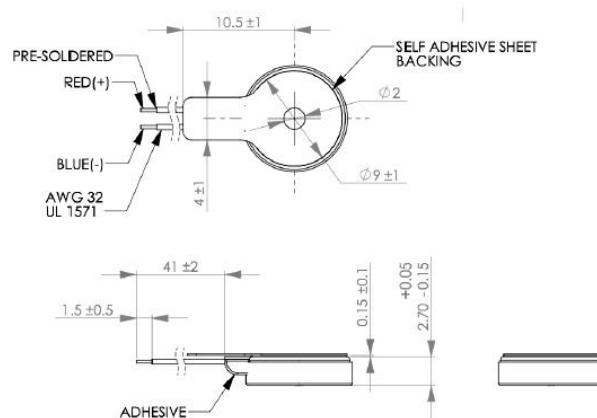
According to Fig.7, the researcher studied the obtained data for research development. The data was collected from the interviews with physiotherapists, nurses, and expert physicians for the design of the portable snoring sensor pillow. The operations of the equipment began with the vibration of the motor when snoring sounds were

over the critical level as determined so that users were warned to change their sleeping postures (positions) and stop snoring [5]. The researcher designed and developed a set of equipment for the snoring sensor pillow. It consisted of a sound sensor module, a vibrating motor, and Arduino control panel. The procedures of the design together with the development were imparted as follows:

- 1) Designed an electronic circuit to control and monitor the operation of the motor. The command was written to Arduino control panel in C programming language. Then, the operation was tested.
- 2) Connected the sound sensor module circuit to Arduino control panel. Wrote the command so that the sound sensor module could adjust values of voices. Next, tested the operation. If snoring sounds were detected, the motor would vibrate to activate snoring sleepers and to stop their snoring [8].
- 3) The last step was to insert the set of the equipment inside the designed pillow.



**Fig. 3.** Sound Sensor Module



**Fig. 4.** Vibration Motor ERMS

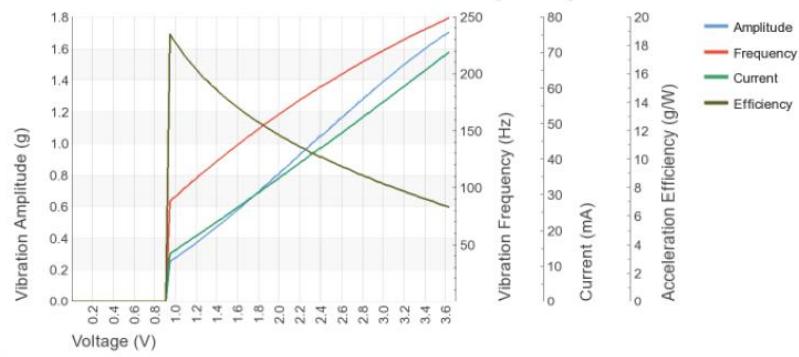


Fig. 5. Typical frequency response of vibration motor ERMS

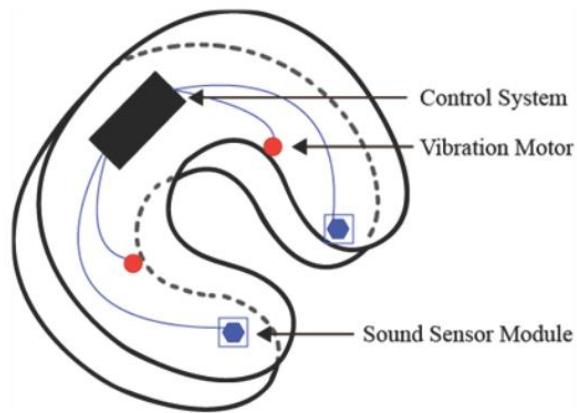
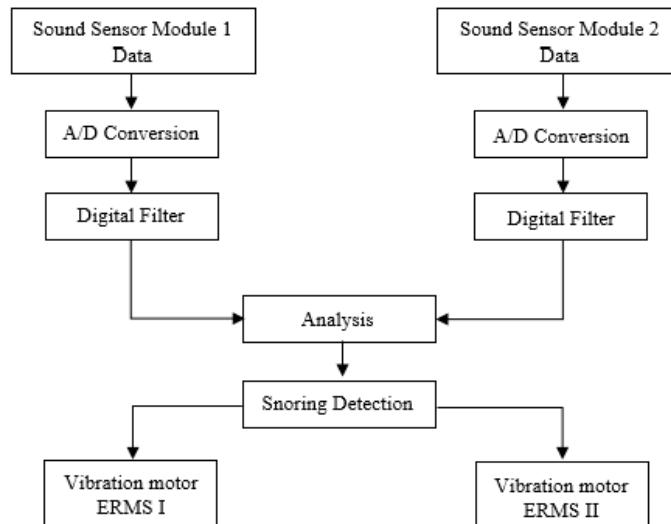


Fig. 6. Location of the vibration motor ERMS and Sound Sensor Module inside the pillow



**Fig. 7.** Block diagram of the system operation

#### 4.2 Data Acquisition

There were 3 samples who met the selection criteria of the research and treatment follow up. All of them were males and were tested by employing the snoring sensor pillow.

**Table 2.** Summary of volunteer's physical information at the first stage

Parameters	Volunteer 1	Volunteer 2	Volunteer 3
Age	27	30	34
Sex	female	male	male
Height(cm)	162	175	172
Weight(Kg)	48	76	72
BMI	18.29	24.82	24.34

At the first stage, three subjects (males) summarized in Table 2. Data were participated in the laboratory environment to adjust the threshold value in order to separate the snoring only from the subject and ratio of snoring time [10]. Table 3. Shows envi-

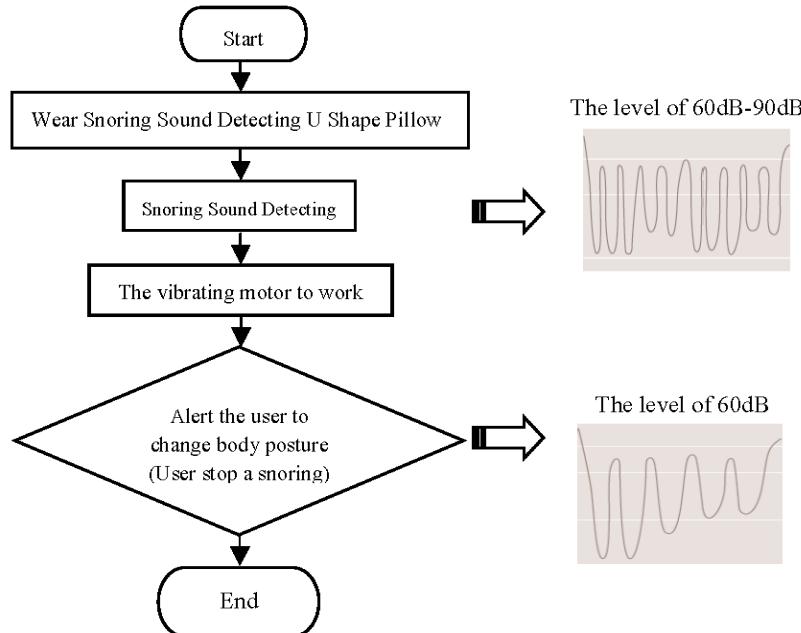
ronment of stored data. Since it is known that snoring sounds measurements from each subject were recorded by the data acquisition system, biopac [11].

**Table 3.** Environment of stored data

Noise	Definitions
Case 1	Snoring only from the subject
Case 2	Snoring with ambient music

#### 4.3 Data Analysis Algorithm

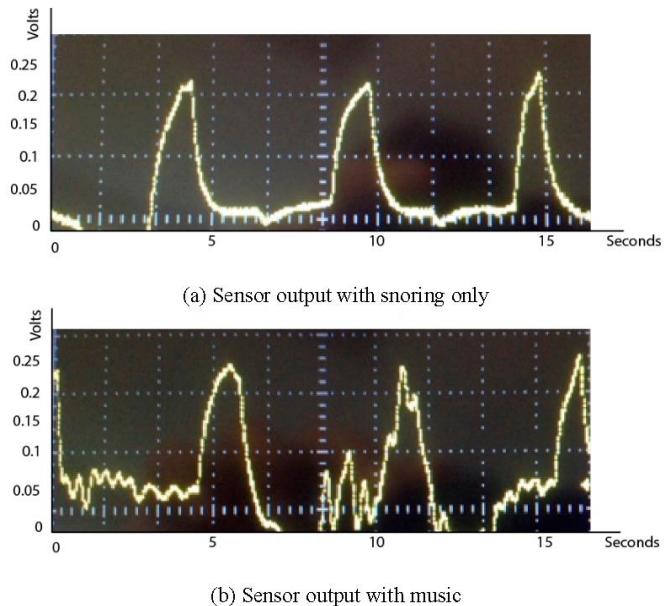
The operation of the pillow was tested in accordance with the specified functions. To illustrate, when the microphone module could detect snoring sounds, it would command the motor to vibrate for alerting the snoring sleepers to be conscious or to change sleeping postures and stop snoring [4]. The command was written in C programming language for controlling the vibration of the motor as show in Fig.8.



**Fig. 8.** Flowchart of data analysis

## 5 Experimental Results

The research purposed to performance is compared result by Snoring only from the subject and snoring with ambient music. The experimental results are as follows:



**Table 4.** Summary of snoring values from two sensors

		Sound Sensor	Ratio (%)	Remarks
a	1	0.195	92.27	Snoring
	2	0.196	92.87	Snoring
	3	0.213	95.54	Snoring
b	1	0.035	-12.32	Music
	2	0.219	92.54	Snoring
	3	0.081	-102.35	Music
	4	0.218	92.48	Snoring
	5	0.032	-12.10	Music

From the experiments that have tested for 2 level, we found the ratio of the snoring only from the subject and snoring with ambient music performances are described as follow;

As can be seen from the table 4, the positive values of ratio are from 92.27% up to 95.54%. Ratios with negative values mean that the decision of noise [11]. Therefore, the ratio of higher than 70% is sufficient as a threshold for consideration snoring sounds from ambient noises.

## 6 Conclusion

In this paper, after the researcher had evaluated the efficiency of the pillow's performance, it was discovered that the distance between the microphone module and snoring sounds that could make the motor vibrate at its best was 10 cm. And when the pillow was actually applied to the 3 samples, it was unveiled that the results of the test depended on the loudness of snoring sounds (dB). Moreover, the longer the distances between snoring sounds and the sound sensor module were, the more difficult to detect snoring sounds. The vibration of the motor would also delay or even not operate at all. Apart from the aforesaid factors, it hinged on ages, genders, or sleeping postures of the samples as well.

## References

1. Satoshi, I., Satoshi, U., Nakamura, Y., Motegi, M., Ogawa, K., Shimokura, K.: A basic study of a pillow-shaped haptic device using a pneumatic actuator. In: IEEE International Symposium on Mechatronics and its Applications, Amman, Jordan, 27-29 (2008)
2. Azarbarzin, A., Moussavi, A.: Snoring sounds variability as a signature of obstructive sleep apnea. *Med. Eng. Phys.*, vol. 35, no. 4, pp. 479–485, Apr. (2013)
3. Shumit, S., Mahsa, T., Zahra, M., Azadeh, Y.: Effects of Changing in the Neck Circumference during Sleep on Snoring Sound Characteristics. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.2235-2238 (2015)
4. Jin Zha., Qian Zhang., Yuanpeng, W., Chen, Q.: A Real-time Auto-Adjustable Smart Pillow System for Sleep Apnea Detection and Treatment. IPSN, pp.179-190 (2013)
5. Ran, W., Xing Li., Hee Sun Kim., Jae Joong, Im., Hyun Jeong, K.: A Development of Pillow for Detection and Restraining of Snoring. In: IEEE Int. Conf. on Biomedical Engineering and informatics, pp.1381-1385 (2010)
6. Rajendra, S., Shashikant, D.: Mobile operated anti-snoring pillow. In: IEEE Int. Conf. on Environmental and Computer Science, pp.441-444 (2009)
7. Hyung, G., Dong, W.: Intelligent Pillow Type Wireless Charger for Fully Implantable Middle Ear Hearing Device with a Function of Electromagnetic Emission Reduction. In: IEEE International Symposium on Intelligent Information Technology Application, pp.835-838 (2008)
8. Xin Zhu., Wenxi Chen.: Automatic Home Care System for Monitoring HR/RR during Sleep. In: International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24 , pp.522-525 (2008)
9. Kongsak, R., Nuntigar S.: Short-term outcomes of transoral radiofrequency somnoplasty treatment for snoring. [http://www.rcot.org/download/Shortterm outcomes of Transoral radiofrequency sommoplasty-treatment for snoring \(2009\)](http://www.rcot.org/download/Shortterm%20outcomes%20of%20Transoral%20radiofrequency%20sommoplasty-treatment%20for%20snoring.pdf)
10. Azarbarzin, A., Zahra, M.: A Comparison between Recording Sites of Snoring Sounds in Relation to Upper Airway Obstruction. International Conference of the IEEE EMBS, pp.4246-4249 (2012)
11. Duk Keun, J., Hee Sun, K., Ran, W., Jan Joong, Im.: A study for the development of PVDF vibration sensor and establishment of noise removal algorithm for snoring detection pillow. In: IEEE Int. Conf. on Environmental Engineering and informatics, pp.1031-1035 (2011)
12. Boomidevi, R., Pandiyan, R., Rajasekaran, N.: Monitoring Cardio Respiratory and Gesture Recognition System Using Sensor Pillow System. International Research Journal of Engineering and Technology (IRJET), Vol. 2, No. 8, pp. 83-87 (2015)

## A multi-objective adaptive Invasive Weed Optimization intelligence approach for solving DNA sequence design

Qiang Zhang\*, Gaijing Yang, Changjun Zhou, Bin Wang  
Key Laboratory of Advanced Design and Intelligent Computing (Dalian university),  
Ministry of Education, Dalian, 116622, China  
zhangq30@yahoo.com

**Abstract.** DNA sequence design is a key factor to ensure the success of DNA computing. In order to make the DNA computing more reliable, many studies have focused on the DNA sequence design. DNA sequence design relates to several and conflicting design criteria, in this paper, DNA sequence design is formulated as a multi-objective optimization problem, and solved by using a multi-objective adaptive Invasive Weed Optimization intelligence approach (MA\_IWO for short). Concretely speaking, our approach (MA\_IWO) generates reliable DNA sequences with the consideration of six different conflicting design criteria by introducing the fast non-dominated sorting. Moreover, adaptive mechanism is introduced into the reproduction phase of the Invasive Weed Optimization algorithm (IWO for short), so that the standard deviation of each generation can be changed adaptively according to the fitness value. In addition, our results are validated by comparison with other related works published in the literature. What can be concluded is that the novel approach presented in the paper obtains very satisfactory results which significantly surpass the other previously published results.

**Keywords:** DNA sequence design; DNA computing; IWO algorithm; fast non-dominated sorting.

### 1 Introduction

The computational model of DNA computing was first proposed by Adleman [1]. He has used modern molecular biology, and operated the DNA molecular in vitro to solve the Hamiltonian path problem with seven vertices successfully. This computational model took DNA molecular as a tool, and made full use of the powerful parallel computing ability of DNA hybridization. DNA computing relies on the biochemical reactions of DNA molecules, which may result in incorrect or unsatisfactory results. So the hybridization between DNA sequences must be strictly controlled. Further, this also reflects the importance of DNA sequence design.

DNA sequence design is the guarantee of the success of DNA computing. However, the encoding of the specific problem needs to meet several and conflicting design criteria simultaneously. That is to say, the DNA sequence design is essentially a multi-objective optimization problem.

For the last few years, many researchers have focused on the DNA sequence design for the aim of accurate DNA computing. Among all the strategies, exhaustive and random searches [2-3] were the simplest, but at the expense of using an enormous amount of computational resources. In order to select dissimilar sequences from a great deal of sequences, Template-map strategies [4-5] were brought into existence. An expression based on directed graphs in which the nodes with four strands represented base strands was used to generate a limited number of sequences. And the four strands could appear in a long sequence as successors of its child nodes. In 1999, Marathe et al. [6] designed DNA sequences based on Hamming distance and free energy by using dynamic programming. On the other hand, there also have studies on using biological-inspired methods to design DNA sequence. Simulated annealing was used to generate reliable sequences in preference [7], in this paper, different biochemical criteria were combined into a fitness function. A method based on in-vitro evolution was proposed by Deaton et al. [8-9] to find non-cross-hybridizing DNA libraries. Other biologically inspired methods took other properties presented in DNA chains into account, such as thermodynamic properties or free energy.

However, the most widely used methods are intelligent evolutionary algorithms. Cui et al. used hybrid algorithm MPSO/GA (short for Modified Particle Swarm Optimization/Genetic Algorithm) [10] to solve this problem. Ren et al. proposed a discrete particle swarm optimization (DPSO) [11] to produce DNA sequences with multiple constraints. In [12], with the combination of the particle swarm optimization and the cultural algorithm, the hybrid algorithm generated DNA sequence set satisfying the relevant hamming distance constraints. Xiao et al. [13] applied the gravitational search algorithm to the design of DNA sequence. In [14], Xiao et al. solved the DNA sequence design problem with a membrane evolutionary algorithm. Zhang et al. [15] first introduced the IWO into DNA sequence design with three thermodynamic evaluation functions (h-measure, Tm and free energy). After that, Yin et al. [16] proposed a cultural evolution based IWO method. In the paper, the obstacles of the traditional IWO algorithm that can't be applied to discrete problems directly were solved by defining the colonizing behavior of weeds. Then, Luo et al. [17] also adopted IWO algorithm to produce reliable DNA sequences, meanwhile, they took into account hamming distance, similarity, continuity, hairpin and melting temperature. These literatures are not real multi-objective researches. The authors eventually convert multiple objectives into a single objective to simplify the scheme by using a constrained weighted summation strategy. Shin et al. proposed a multi-objective method based on six different evaluation criteria in 2002 [18], and improved the original approach in 2005 [19]. Wang et al. [20] also dealt the problem with the non-dominated sorting genetic algorithm. Chaves-Gonzalez et al. [21] used an adapted multi-objective version of the differential evolution to generate reliable DNA sequences. Then, a multi-objective method based on firefly behavior is proposed [22], and six different and conflicting design criteria are used to obtain the reliable DNA sequence.

In the paper, DNA sequence design is dealt with multi-objective method. We propose a novel multi-objective IWO algorithm. The fast non-dominated sorting and adaptive mechanism are combined with IWO, which not only can enrich the diversity of population, but also benefit to explore the potential areas of the optimal solution.

Moreover, our results are validated by comparison with other related works published in the literature. The comparison results prove the superiority of our algorithm.

## 2 Related background

### 2.1 DNA sequence design

DNA sequence design relates to several conflicting design criteria and these criteria must be satisfied at the same time. In the paper, these criteria are treated as objectives. Eventually, DNA sequence design is formulated as a multi-objective optimization problem. Because of the overlap between these objectives, we refer to the relevant publications and carefully study the relationship between these objectives. Finally, six design criteria including similarity, H-measure, Hairpin structure, continuity, GC content and Tm are selected to generate reliable DNA sequences. Among them, the first four objectives should be minimized, while the remaining two objectives should remain as constant as possible. These design criteria are introduced in the below, and the detailed descriptions please refer to reference [21].

**Similarity:** In DNA sequence design, each sequence should be kept the uniqueness. The criterion is to calculate the similarity of two given sequences in the same direction including position shift. The evaluation function is defined as follows [21]:

$$f_{sim}(L) = \sum_{i=1}^m \sum_{j=1, i \neq j}^m \max_{g,l} [S\_dis(x_i, shift((x_j(-)^g x_j, l)) + S\_con(x_i, shift((x_j(-)^g x_j, l))) \quad (1)$$

with  $0 \leq g \leq \text{round}(n/4)$ ,  $|l| \leq n-1$ .  $l$  denotes the shift position,  $(-)^g$  is the g gabs.  $S\_dis$  is the non-continuous similarity,  $S\_con$  is the continuous similarity.

**H-measure:** H-measure is similar to similarity. This criterion measures the degree of cross hybridizations between sequences by counting complementary bases with the other sequences. The specific calculation formula is as follows [21]:

$$f_{H\text{-measure}}(L) = \sum_{i=1}^m \sum_{j=1}^m \max_{g,l} [h\_dis(x_i^R, shift((x_j(-)^g x_j, l)) + h\_con(x_i^R, shift((x_j(-)^g x_j, l))) \quad (2)$$

with  $0 \leq g \leq \text{round}(n/4)$ ,  $(-)^g$  is the g gabs.  $x_i^R$  is  $x_i$ 's reverse sequence.  $H\_dis$  is the non-continuous H-measure,  $H\_con$  is the continuous H-measure.

**Hairpin structure:** Hairpin structure is a cyclic structure formed by self-complementary of DNA sequences. This criterion can effectively avoid the formation of secondary structures by estimating a penalty for DNA hairpin. The mathematical formula is as follows [21]:

$$f_{Hairpin}(L) = \sum_{i=1}^m \sum_{p=p_{\min}}^{n-R_{\min}} \sum_{r=R_{\min}}^{n-2p} T \left( \sum_{j=0}^{\text{pinlen}(p,r,i)-1} bp(x_{p+i-j}, x_{p+i+r+j+1}), \text{pinlen}/2 \right) \quad (3)$$

$$bp(x_a, x_b) = \begin{cases} 1 & x_a = \overline{x_b} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$pinlen(p, r, i) = \min(p + i, n - p - i - r) \quad (5)$$

Where  $\overline{x_b}$  is the complement of  $x_b$ ,  $p$  is the length of hairpin stem,  $r$  is the length of hairpin loop.  $R_{\min}$  and  $P_{\min}$  are respectively the minimum length of forming a hairpin loop and a hairpin stem.

**Continuity:** This criterion is usually used to indicate the degree of continuous repetitions of the same base in a given sequence. The specific calculation formula is as follows [21]:

$$f_{con}(L) = \sum_{i=1}^m \sum_{j=1}^{n-i+1} T(C_\rho(x_i, j), t)^2 \quad (6)$$

$$T(a, t) = \begin{cases} a & \text{if } a > t \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$C_\rho(x_i, j) = \begin{cases} c & \text{if } \exists c \text{ s.t. } x_i^j \neq \rho, x_i^{j+k} = \rho, \text{ and } x_i^{j+k+1} \neq \rho \text{ for } 1 \leq k \leq c \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where  $L$  represents the DNA library,  $m$  is the number of DNA sequences,  $n$  is the length of each individual,  $x$  is a DNA sequence with a length of  $n$ , and  $t$  is the continuity threshold.

**GC content:** The GC content is defined as the percentage of base G and C in a DNA sequence. This criterion can also effectively prevent the occurrence of non-specific hybridization.

**Tm:** Tm is an important parameter to determine the reaction efficiency. In order to effectively keep the DNA duplex stability, the Tm values of DNA sequences which participate in the reaction should be the same as possible. Many methods are used to calculate the Tm, such as the nearest neighbor model, and the GC% technique.

The nearest neighbor method [21]:

$$f_{T_m}(L) = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln([C_T]/4)} - 273.15 \quad (9)$$

GC% method [21]:

$$f_{T_m}(L) = 81.5 + 16.6 \times \log_{10} \left( \frac{[salt]}{1.0 + 0.7 \times [salt]} \right) + 41 \times GC_{content}(x) - \frac{500}{|x|} \quad (10)$$

Among them,  $[salt]$  is the concentration of salt,  $R$  is the concentration of gas,  $[C_T]$  is mole concentration of DNA molecule. In this paper, the nearest neighbor method is used to calculate Tm.

**Multi-objective function:** As is well known, in a multi-objective environment, the result is a set of alternative high quality solutions known as Pareto-optimal solutions, rather than a unique solution that simultaneously satisfies every objective. The problem of DNA sequence design can be formulated as a multi-objective problem as follows.

Minimize  $f_i(x)$

Where  $i \in \{similarity, H-measure, continuity, hairpin\}$

Subject to GC content = 50%

## 2.2 Basic IWO algorithm

In the basic IWO algorithm [15-17], the feasible solution of the problem is expressed by the weed, and the population is the collection of all the weeds. During the process of evolution, the weeds produce seeds through the propagation; the seeds develop into weeds through spatial diffusion, so repeatedly. When the number of weeds in the population reaches the maximum population size, the weeds can survive by the competition. The weed with good fitness is reserved, while the weed with poor fitness is eliminated.

## 3 The proposed method

### 3.1 Fast non-dominated sorting

DNA sequence design is a multi-objective optimization problem in essence. Non-dominated sorting genetic algorithm (NSGA-II) proposed by Deb et al. [23] in 2002 is recognized as a highly efficient multi-objective algorithm. The fast non-dominated sorting strategy is the core of NSGA-II. It can simplify multiple objectives to a fitness function, so that it can solve any number of target problems, and can solve the extreme value problems.

After determining the rank of each individual in the population, a series of non dominated solutions ( $F_1, F_2, \dots, F_m$ ) are obtained, which are known as the first rank, the second rank .... Assignment to fitness can be easily based on the rank of the individual. In the paper, we adopt the idea of Dong et al. [24]. The specific formula is as follows:

$$f = \frac{1}{1 + \frac{r(i)-1}{R}} \quad (11)$$

Where  $r(i)$  is the rank of individual  $i$  in group ranking, while  $R$  is the total rank of contemporary group ranking. The above method can avoid the shock phenomenon caused by the large gap of total rank of different generations.

### 3.2 Adaptive mechanism

In the basic IWO algorithm,  $\sigma_t$  of each iteration only decreases with the increase of the number of iterations; while in a certain generation, it keeps the same, which is obviously not conducive to the convergence of the algorithm. Especially, in the late phase of evolution, along with the increase of the iterations,  $\sigma_t$  is getting smaller and smaller, such that the new seeds can only be distributed in the neighborhood of the parent, which is easy to fall into local optimum. Therefore, we introduce the adaptive mechanism. So that  $\sigma_t$  of each iteration adaptively adjusts according to the

fitness value of each individual and the maximum and minimum fitness values of the current iteration. The specific formula [25-26] is as follows:

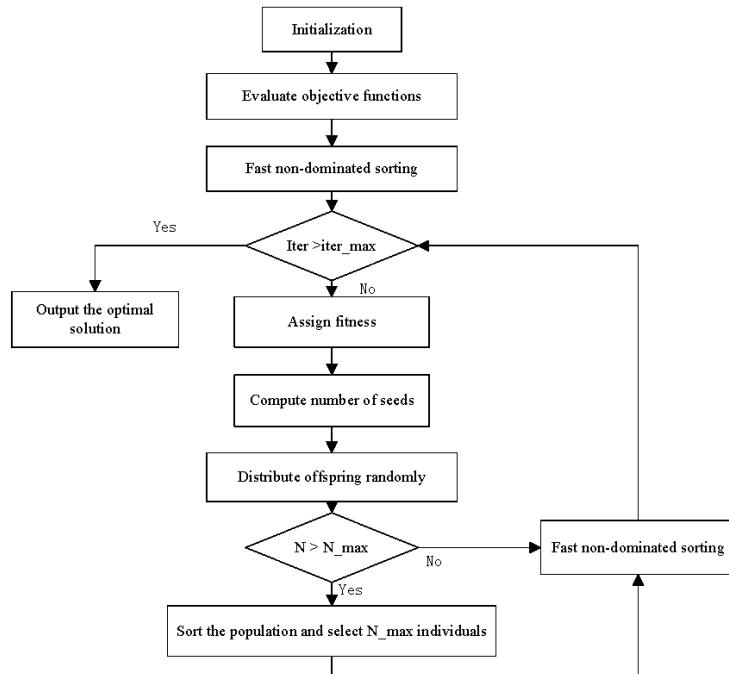
$$\sigma_{i,t} = \begin{cases} [1 + 0.5 * (f(x_{i,t}) - f_{avg,t}) / (f_{worst,t} - f_{avg,t})] * \sigma_t & f(x_{i,t}) \geq f_{avg,t} \\ [1 - 0.5 * (f_{avg,t} - f(x_{i,t})) / (f_{avg,t} - f_{best,t})] * \sigma_t & f(x_{i,t}) < f_{avg,t} \end{cases} \quad (12)$$

In which,  $\sigma_t$  is the same with formula ,  $f_{avg,t}$ ,  $f_{best,t}$  and  $f_{worst,t}$  are respectively the average, the best and the worst fitness value of the current iteration.

From the above formula, we can find the rule: in a certain generation, the standard deviations of individuals with good fitness are smaller, which is helpful for the seeds to distribute in the neighborhood of the better weeds; while the individuals with poor fitness are bigger, which is conducive to the distribution of the more excellent seeds in the far range. Moreover, the range of  $\sigma_{i,t}$  is  $[0.5, 1.5] * \sigma_t$ . These rules not only help the new seeds produced jump out of local optimal, speed up the convergence rate, but also effectively balance the global and local search capabilities.

### 3.3 Multi-objective IWO intelligence approach

In the paper, the DNA sequences are encoded in the following way:  $C - 0, T - 1, A - 2, G - 3$ . The flow chart of the algorithm in shown in Fig.1.



**Fig.1.** Flow chart of the algorithm

The specific steps for this algorithm are summarized as follows:

Step 1. Initialize a population of  $N_0$  solutions ( $P$ ).

Step 2. Evaluate objective functions for all the individuals in  $P$ .

Step 3. Assign the rank based on fast non-dominated sorting for each individual  $p \in P$ .

Step 4. When the condition of stopping ( $iter > iter_{max}$ ) is not satisfied, perform steps 5 to 9.

Step 5. Assign fitness based on the rank of each individual.

Step 6. Compute number of seeds of  $p$ .

Step 7. Randomly distribute generated seeds over the search space with normal distribution which has an adaptive standard deviation around the parent plant ( $p$ ).

Step 8. Add the generated seeds to the solution set ( $P$ ). When the population exceeds the preset value ( $N_{max}$ ), the weeds and seeds are sorted according to the fitness values, and the best  $N_{max}$  individuals are allowed to survive. Otherwise, go to step 9.

Step 9. Assign the rank based on fast non-dominated sorting.

#### 4 Experimental results and analysis

An adaptive IWO algorithm based on multi-objective optimization problem model was implemented in the MATLAB to search for best sequences. For fair comparison, we chose the same objectives, and the specific parameters of evaluation system were as follows. For similarity and H-measure, the lower limits for the continuous case and the discontinuous case respectively equaled to six bases and 17%. The threshold value for continuity was 2. For hairpin structure, at least a six base stem and a six base loop were required. As shown in Table 1, we use the same basic parameter settings just as the other experiments about IWO algorithm.

**Table 1.** Parameters used in our algorithm

Symbol	Quantity	Value
$N$	Number of the initial population	10
$N_{max}$	Maximum number of population	20
$iter\_max$	Maximum number of iterations	200
$seed\_max$	Maximum number of seeds	3
$seed\_min$	Minimum number of seeds	0
$n$	Nonlinear modulation index	3
$\sigma_{int}$	Initial value of standard deviation	5
$\sigma_{final}$	Final value of standard deviation	1

**Table 2.** Comparison results

DNA Sequences (5'—3')	Continuity	Hairpin	H-measure	Similarity	Tm	GC (%)
Our Sequences						
AGGAATGACGAGGGTAGA	0	0	64	45	63.89	50
ACCTACTCACACACCTACCA	0	0	59	55	64.34	50
TCTGCTGCCAGTCCTCTCT	0	0	62	49	64.68	50
GTTAACACTTGAGGCGTCCT	0	0	62	50	64.07	50
CTTGTACTTCTCGCTCGCT	0	0	61	52	64.70	50
ACACCAATACGCAAGAAC	0	0	62	48	65.26	50
GGATATGTTCGGTCGTGGA	0	0	62	46	64.20	50
Group results	0	0	432	345		
CE-IWO [16]						
CTCTTCATCCACCTCTTCTC	0	0	47	58	61.38	50
CTCTCATCTCTCGTCTTC	0	0	39	57	61.44	50
TATCCTGTGGTGTCCCTCT	0	0	51	54	64.46	50
CACAGGCAGATCTAGTCAG	0	0	69	51	62.02	50
TCTCTTACGTTGGTGGCTG	0	0	54	49	64.63	50
GTATTCCAAGCGTCCGTGTT	0	0	56	48	65.30	50
AAACCTCCACCAACACACCA	9	0	53	47	66.71	50
Group results	9	0	369	374		
IWO Sequences [17]						
GATGGATTACCTGCACCT	9	4	60	54	62.29	45
CCTTCTCTCGTCTCATACA	0	0	61	53	60.72	45
ACGATCGATTAATGGGAGTC	9	3	66	50	61.52	45
ATAAGTAGGGACTGCTCTAC	9	0	68	52	59.84	45
CCTAAGAACACAGGGCATAG	9	4	65	53	62.04	50
CTGGAAGCGTTGCTAATT	9	6	66	52	63.38	45
GCAGATTCCCGGATACTCAG	9	7	66	56	64.34	55
Group results	54	24	452	370		
NACST/Seq Sequences [19]						
CTCTTCATCCACCTCTTCTC	0	0	43	58	61.38	50
CTCTCATCTCTCGTCTTC	0	0	37	58	61.44	50
TATCCTGTGGTGTCCCTCT	0	0	45	57	64.46	50
ATTCTGTCGTTGCGTGT	0	0	52	56	65.83	50
TCTCTTACGTTGGTGGCTG	0	0	51	53	64.63	50
GTATTCCAAGCGTCCGTGTT	0	0	55	49	65.30	50
AAACCTCCACCAACACACCA	9	0	55	43	66.71	50
Group results	9	0	338	374		
DEPT [21]						
ACCACAAACAACACACACCC	9	0	29	55	65.97	50
CCATACCAGCCAACCGAAAA	16	0	39	56	65.33	50
GAGAGAAGAGAAGAGGCCAA	0	0	39	53	63.17	50
CCATTCTTAACCTCTCTCC	0	0	59	39	61.40	50
GGAGCAATGGAGAATAAGGG	9	0	48	47	62.42	50
ACACACACACACACACAC	0	0	27	49	65.86	50
GGAAGGAGGAGGAAGAAGAA	0	0	37	45	62.84	50
Group results	34	0	278	344		

As shown in Table 2, by examining design criteria values, the quality of the generated sequences respectively in our paper and references [16] [17] [19] [21] is evaluated. NACST/Seq approach and DEPT algorithm are related to the algorithm proposed in our paper, they all transform DNA sequence design as a multi-objective optimization problem. Besides, our algorithm is improved based on the IWO. So, we select the data in the reference [16], the reference [17], reference [19] and reference [21] as reference. In the paper, the lower values of the design criteria are the better individual.

Before the comparison, we recalculated the Tm (as shown in the sixth column) in the references according to our method, and corrected the errors in the reference [17] (the hairpin value of the sixth sequence was 6 instead of 3). In addition, the continuity in reference [21] was calculated at the threshold value of 4, the lower the threshold value, the better the performance of the sequence in terms of continuity, so we recalculated the values of sequences in the continuity at the threshold value of 2 (recalculated values are shown in Table 2).

Shin et al. [19] used a multi-objective evolutionary optimization algorithm based on the standard NSGA-II. Yin et al. [16] proposed a cultural evolution based IWO method. Luo et al. [17] adopted IWO algorithm to produce reliable DNA sequences. Chaves-Gonzalez et al. [21] used an adapted multi-objective version of the differential evolution. The generated sequences (7 sequences and 20 bases per sequence) of them and objective values are shown in Table 2. In terms of group results, our sequences have the minimal values for continuity; the same values of hairpin and GC content with reference [16], reference [19] and reference [21]; the values of hairpin in them are smaller than in reference [17], the values of GC content in them are strictly 50%, whereas the values of GC content in reference [17] range from 45% to 55%. This means that our sequences have reduced the secondary structures. The values of H-measure in our paper are smaller than in reference [17], but larger than in the other three references. The values of similarity in our paper are almost the same with in reference [21], and smaller than in the other three references. That is to say, our sequences have higher probability to hybridize with its correct complementary sequences.

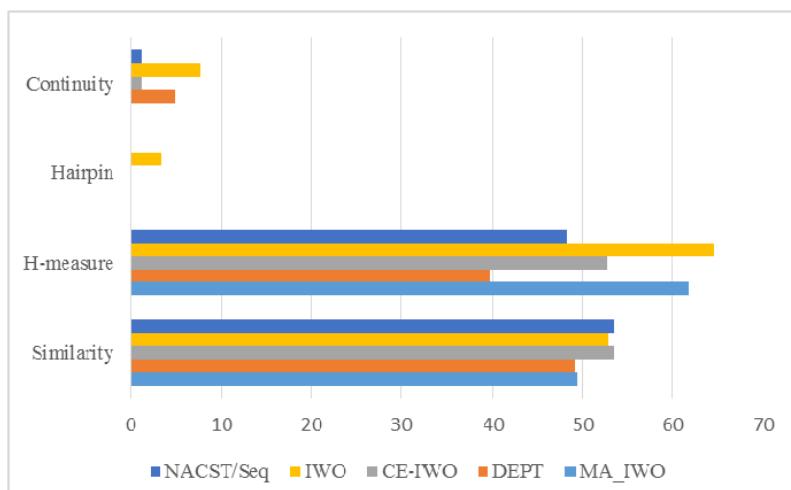
To make a more intuitive performance for the comparison results, we take the average of each objective and show it in the form of a graph (Fig.3). Every objective is to be minimized. From Fig.2, our sequences show minimal values for continuity and hairpin, meanwhile, lower values for H-measure and similarity. Therefore, our sequences have little secondary structure and higher probability to hybridize with their correct complementary sequences.

A stable temperature environment is crucial for biochemical reactions. The volatility of a set of data is often measured by variance. Therefore, we evaluate Tm by calculating the variance. The smaller the variance is, the smaller the volatility is and the more stable of the data is. The comparison results of Tm are shown in Table 3, and easily find that our sequences have the minimum variance.

**Table 3.** Comparison results of NACST/Seq, IWO, DEPT and MA\_IWO in Tm

	NACST/Seq [19]	IWO [17]	CE-IWO [16]	DEPT [21]	MA_IWO
variance	4.3287	2.3330	4.3958	3.3758	0.2173

In summary, MA\_IWO obtains very satisfactory results which are better than the other previously published results.



**Fig. 2.** Comparison results among average values of NACST/Seq, IWO, CE-IWO, DEPT and MA\_IWO in continuity, hairpin, H-measure and similarity

## 5 Conclusion

In the paper, we eventually transformed DNA sequence design as a multi-objective optimization problem by introducing the fast non-dominated sorting. Moreover, in the reproduction phase of the IWO algorithm, we introduced adaptive mechanism to make the standard deviation of each generation change adaptively according to the fitness value. Thus, the new algorithm (MA\_IWO) which is not prone to fall into local optimum was generated. From the comparison results, the seven sequences generated by MA\_IWO present minimal values for hairpin and continuity, lower values for H-measure and similarity. That is to say, our sequences have little chance to form secondary structures and more opportunity to hybrid with their correct complementary sequences. Altogether, MA\_IWO is feasible and effective. However, the time complexity of MA\_IWO is high. In the future, we will attempt combine with other approaches to reduce the time complexity.

## 6 Acknowledgment

This work is supported by the National Natural Science Foundation of China (Nos. 61425002, 61572093, 61402066, 61402067, 61370005), Program for Changjiang Scholars and Innovative Research Team in University (No.IRT\_15R07), the Program

for Liaoning Innovative Research Team in University(No.LT2015002), the Basic Research Program of the Key Lab in Liaoning Province Educational Department (Nos.LZ2014049, LZ2015004), and the Program for Liaoning Key Lab of Intelligent Information Processing and Network Technology in University.

## References

1. Adleman, L. M.: Molecular computation of solutions to combinatorial problems. *Science* 266, 1021-1024 (1994).
2. Hartemink, A. J., Gifford, D. K., Khodor, J.: Automated constraint-based nucleotide sequence selection for DNA computation. In: Proceedings of 4th DIMACS Workshop on DNA Based Computers, 227-235 (1998).
3. Penchovsky, R., Ackermann, J.: DNA library design for molecular computation. *Journal of Computational Biology*, 215-229 (2003).
4. Frutos, A. G., Thiel, A. J., Condon, A. E., Smith, L. M., Corn, R. M.: DNA computing at surfaces: four base mismatch word design. In: Proceedings of the 3rd DIMACS Workshop on DNA Based Computers, 238 (1997).
5. Feldkamp, U., Saghafi, S., Banzhaf, W., Rauhe, H.: DNA sequence generator – a program for the construction of DNA sequences. In: Proceedings of 7th International Workshop on DNA Based Computers, 179–188 (2001).
6. Marathe, A., Condon, A. E., Corn, R. M.: On combinatorial DNA word design. In: Proceedings of 5th DIMACS Workshop on DNA Based Computers, 75–89 (1999).
7. Tanaka, F., Nakatsugawa, M., Yamamoto, M., Shiba, T., Ohuchi, A.: Developing support system for sequence design in DNA computing. In: Proceedings of 7th International Workshop on DNA Based Computers, 340–349 (2001).
8. Deaton, R., Chen, J., Bi, H., Garzon, M., Rubin, H., Wood, D. H.: A PCR-based protocol for *in vitro* selection of non-cross-hybridizing oligonucleotides. In: Proceedings of the 8th International Workshop DNA Based Computers, 196–204 (2002).
9. Deaton, R., Chen, J., Bi, H., Rose, J. A.: A software tool for generating non-cross-hybridization libraries of DNA oligonucleotides. In: Proceedings of the 8th International Workshop DNA Based Computers, 252–261 (2002).
10. Cui, G. Z., Li, X. G.: The optimization of DNA encoding based on modified PSO/GA algorithm. In: Proceedings of 2012 International Conference on Computer Design and Applications, 609-614 (2010).
11. Ren, X.N., Zhang, D.F., Xiang, X.Y.: A combination model to optimize DNA encoding based on discrete particle swarm optimization. *Computer Engineering & Science*, 33(3), 179-184 (2011).
12. Yin, Z., Ye, C.M., Ma, H.M.: Cultural evolution based particle swarm optimization algorithm for DNA sequence design. *Computer Engineering and Application*, 47(1), 40-42 (2011).
13. Xiao, J.H., Cheng, Z.: DNA sequences optimization based on gravitational search algorithm for reliable DNA computing. In: Proceedings of 6th International Conference on Bio-Inspired Computing: Theories and Applications, Penang, 103-107 (2011).
14. Xiao, J.H., Zhang, X.Y., Xu, J.: A membrane evolutionary algorithm for DNA sequence design in DNA computing. *Computer Science & Technology*, 57(6), 698-706 (2012).
15. Zhang, X. C., Wang, Y. F., Cui, G. Z., Niu, Y., Xu, J.: Application of a novel IWO to the design of encoding sequences for DNA computing. *Computers & Mathematics with Applications* 57, 2001-2008 (2009).
16. Yin, Z., Ye, C. M., Yin, W. H., Wen, M.: A cultural evolution based on IWO approach for DNA sequence optimization. *Computational Information System*, 5715-5722 (2011).

17. Luo, D. F., Luo, D. J.: The research of DNA coding sequences based on Invasive Weed Optimization. *Science Technology & Engineering* 13, 3545-3551 (2013).
18. Shin, S. Y., Lee, I. H., Kim, D. M., Zhang, B. T.: Evolutionary sequence generation for reliable DNA computing. In: *Proceedings of Congress on Evolutionary Computation*. 79-84 (2002).
19. Shin, S. Y., Lee, I. H., Kim, D. M., Zhang, B. T.: Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing. *IEEE Trans. Evol. Comput.* 9, 143-158 (2005).
20. Wang, Y. F., Shen, Y. P., Zhang, X. C., Cui, G. Z.: DNA codewords design using the improved NSGA-II algorithms. In: *Proceedings of 4th International Conference on Bio-Inspired Computing*. 48-52 (2009).
21. Chaves-Gonzalez, J. M., Vega-Rodriguez, M. A.: DNA strand generation for DNA computing by using a multi-objective differential evolution algorithm. *Biosystems* 116, 49-64 (2014).
22. Chaves-González J M, Vega-Rodríguez M A. A multiobjective approach based on the behavior of fireflies to generate reliable DNA sequences for molecular computing. *Applied Mathematics & Computation*, 227(2), 291-308 (2014).
23. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* 6, 182-197 (2002).
24. Dong, W.: The optimal design for crane main box beam based on Pareto Genetic Algorithms. Masteral dissertation, Dalian University of Technology, (2005).
25. Basak, A., Pal, S., Das, S., Abraham, A.: Circular antenna array synthesis with a differential invasive weed optimization algorithm. *2010 10th International Conference on Hybrid Intelligent Systems*, 153-158 (2010).
26. Basak, A., Maity, D., Das, S.: A differential invasive weed optimization algorithm for improved global numerical optimization. *Applied Mathematics & Computation* 219, 6645-6668 (2013).

## Fatigue Classification of Military Mission by EEG signals via Artificial Neural Network (ANN)

Worawut Yimyam<sup>1</sup>, and Mahasak Ketcham<sup>2</sup>

<sup>1</sup> Department of Computer Business, Phetchaburi Rajabhat University, Thailand

worawut\_yimyam@hotmail.com

<sup>2</sup> Department of Information Technology Management, King Mongkut's University of Technology North Bangkok, Thailand

mahasak.k@it.kmutnb.ac.th

**Abstract.** This paper proposes the development of an algorithm used for monitoring fatigue of the soldiers while they perform their duty. Electroencephalography (EEG) signals are analyzed by an Artificial Neural Networks (ANN) technique and compared with other techniques. The experimental results show that the ANN provides more accurate results than Bayesnet, Support Vector Machines (SMO), and Naïve Bayes techniques. The result of the ANN technique provides the accuracy, recall, and precision values at 83.77, 0.838, and 0.838, respectively.

**Keywords:** Fatigue, Electroencephalography, Artificial Neural Networks

### 1 Introduction

Nowadays, there are many important duties of soldiers in the mission of the military to perform against the law breaking such as the maintenance of the independence, sovereignty, national security, border security, and forest patrol. There are also the offenses that affect the national security such as drug and weapon trade, and deforestation. Each mission need lots of times to achieve on it. From that, soldiers may feel fatigued because of their mission. [1] When fatigue occurs, it causes accidents during the mission and also reduce work performance of the soldiers. For example, there was an accident occurred to US military called M985 in truck drive. The investigation found that the accident happened because of the driver's fatigue which leads him to death because of the lack of sleeping. [3] This type of problem has affected to the wrong decision, communication errors, and risk assessment. All these consequences have systematic relationships. [6]

Military's mission may face the risk any time. One mission needs many hours for working and uses a lot of military personnel. However, a number of soldiers are not enough in some military base. [13-14] Hence, these actions cause fatigue because the soldiers might work several tasks. For example, the mission of the military is to train the pilots to fly helicopter. In facts, the pilots should sleep appropriately because they have to work all day based on FFA rule. They have to wake up at 5.00 am, moni-

tor the helicopter's engine before flying at 4 pm, take off at 5.30 pm, land at 10.30 pm, and store the helicopter at 12.30 am. Then, they move to another base to prepare helicopter at 2.30 am., re-check again at 6.00 am. until military mission is completed. As from their daily routine duties, it can be seen that the pilot's problem is the lack of sleep. Thus, the system is developed to monitor fatigue from the eyes [10-11]. There are many researches working in this area. However, the measurement of fatigue is not certainly accurate and effective.

Saeid Fazli et al.,[8] proposed eyes tracking in order to monitor fatigue by using image processing technique to find the eyes position and check whether the eyes open or close. The results showed some errors because the experiments of the closed and opened eyes depend on an individual eye. Ye sun et al.,[16] proposed the fatigue monitoring system to detect the physiological sign consisting of eyes and heart. Moreover, sensor technology has been used, but it has a problem in terms of the limitation on transmission distance. Edward et al., [4] proposed the evaluation technologies that can help to monitor fatigue. Researchers found that the eyes monitoring technology has high reliability in case of indicating the fatigue of the body. YenWei Chen and Kenji Kubo [17] proposed the development of face detection and eye movement via webcam by using Gabor filter technique. Filter technique works with color filter data. For face detection, the system monitors a geometric shape from face structure and uses Gabor filter to shift the image. The system runs continuously and displays the results of face detection via a monitor. Sung-Uk Jung and Jang-Hee Yoo [9] proposed the method to increase the quality of eye detection. Researchers cut an irrelevant distraction by SQI method. The three-dimensional image conversion helps detect the eye position. Then, AdaBoost was used in identifying the eye position more precisely.

This paper proposes the development of an algorithm used for monitoring fatigue of the soldiers while they perform their duty. Electroencephalography (EEG) signals are analyzed by an Artificial Neural Networks (ANN) technique and compared with other techniques. Preliminary

#### **a. Fatigue**

In case of fatigue, sleepiness, and loss of concentration in perspective of Intelligent Transportation System: ITS, it was found that it has the same meaning of science, in which there is no criteria for measure of fatigue precisely [10] [11]. Fatigue can be measured by nerve, muscle, body temperature, eye movement, respiratory rate, heart rate and brain function [2] [5] [11] [12]. In addition, the brain function provides better results for analyzing the fatigue and sleepiness. According to the Psychomotor Vigilance Task (PVT) research, it has mentioned that the visual stimulation and visual responsiveness are the main factors of brain monitoring [7].

#### **b. EEG Monitoring**

Human brain whether it has been sleeping or awaking, it has different frequencies which can observe through EEG monitoring. The frequency of signal occurs when there is a change of the electrical stimulation. The analysis of EEG signal has been developed by many researchers [18-20]. EEG monitoring system has been uti-

lized by Digital Signal Processing Units due to its frequency range which can show the behavior of patients and the signal measurement [21]. The types of EEG signal are shown in Table.1.

TABLE I. Types of EEG Signal [24]

Types of EEG Signal	Frequency Spectrum (Hz)	Amplitude ( $\mu V$ )	Significance
Delta	0.1 – 0.3	100 – 200	Deepest, dreamless sleep, unconscious state, cognitive tasks by frontal lobe
Theta	4.0 – 7.5	<30	REM sleep, dreaming, physiological at the age of 1-6, cognitive task by frontal lobe (Fourier analysis), intuition, creativity
Alpha	8.0 – 12.0	30 – 50	The “basic” wave of the brain occurred when stimulating high frequency (alpha block). Relaxed but not sleepy state
Beta	13.0 – 30.0	<20	Sensory and emotional influences, harmonic, wide awake, exciting, conscious states
Gamma	30.0 – 50.0	<10	High mental activity

## 2 System Design

This research focuses on the fatigue monitoring system of the military mission. The system is implemented by the EEG sensor in which it can send signal to smartphone via ZigBee and the frequency radio wave, and can collect EEG data via smartphone program. The system is able to analyze the soldier’s fatigue conditions and sends the alert to the admin. The Fig. 1 shows the overview of system.

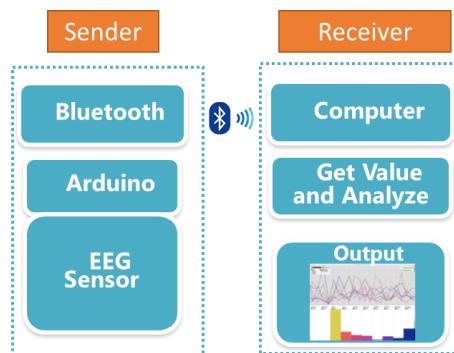


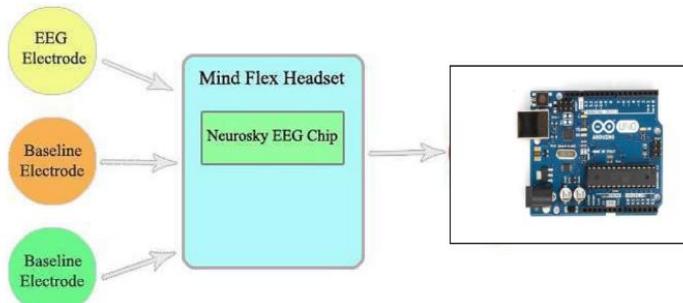
Fig. 1. Overview of system

### a. EEG Sensor

This part is the analysis of the EEG signal received from EEG sensor. The fatigue from researcher's EEG signal is tested as shown in Fig.2.



**Fig. 2.** MindFlex Headset



**Fig. 3** Neurosky EEG Chip

### b. Communication Link Arduino

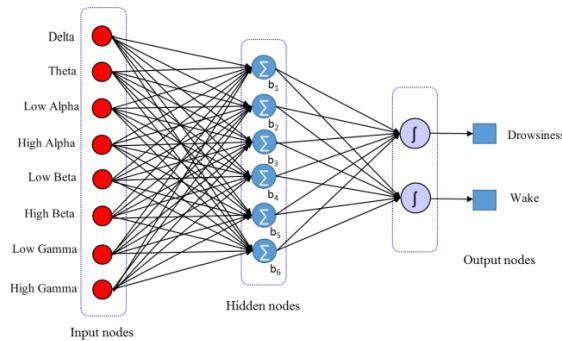
EEG monitoring data is received from MindFlex Headset device. Arduino board is also used to convert value received from sensor to different types of signals which can be divided into 8 ranges: Delta 1-3 Hz, Theta 4-7 Hz, Low Alpha 8-9 Hz, High Beta 18-20 Hz, Low Gamma 31-40 Hz, and High Gamma 41-50 Hz. These frequency waves are transformed in ASCII coding.

### c. Data reception

In data reception, the system connects to the Bluetooth module in order to send signal to computer for analyzing the fatigue in the next step.

#### d. Receive value and analyze with ANN

Artificial Neural Network technique (ANN) is used to analyze data receiving from the EEG signals. The processing applies with neural network of the human brain. The EEG signal is an input of the ANN technique. The EEG data composes of Delta, Theta, Low Alpha, High Alpha, Low Beta, High Beta, Low Gamma, and High Gamma signals. All inputs are multiplied with weight which is represented as w<sub>1</sub>, w<sub>2</sub>, w<sub>3</sub>, w<sub>4</sub>, w<sub>5</sub>, w<sub>6</sub>, w<sub>7</sub>, and w<sub>8</sub>. Each neuron is a bias adjustment with the weighting. It has been sent to the transfer function in order to calculate the result as shown in Figure 4.



**Fig. 4.** Example of Artificial Neural Network Technique

The Equation is shown as below:

$$a^m = f^{m+1}(w^{m+1}x^m + b^{m+1}) \quad (1)$$

Where

$a^m$  means Output Node

$f^m$  means Transfer Function

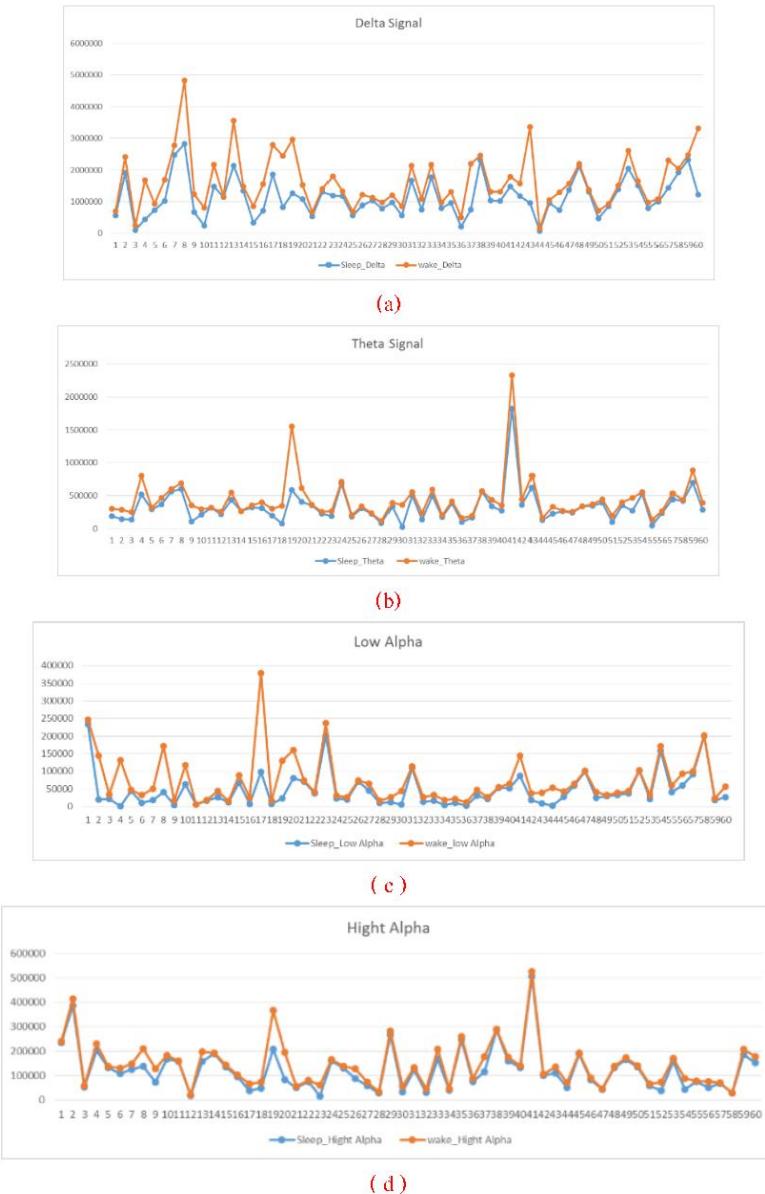
$w^m = 0.2$

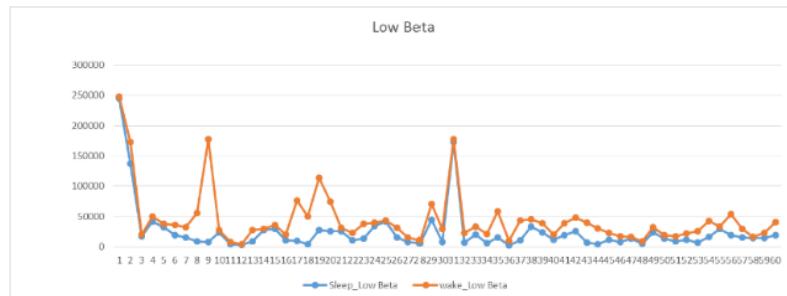
$b^m = 0.1$

$x^m$  means Delta, Theta , Low Alpha, High Alpha, Low Beta ,High Beta ,Low Gamma , High Gamma

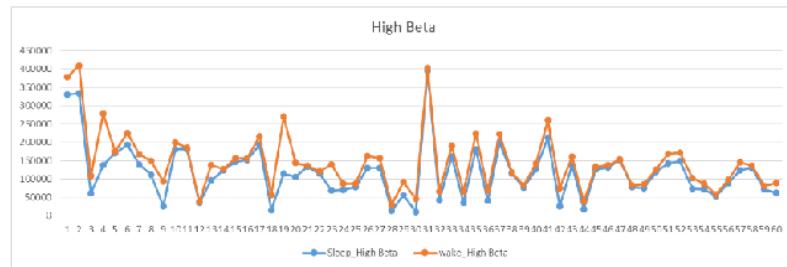
### 3 Experimental Results

The dataset collected from user's EEG signals are measured using EEG sensor. The system requires the data record from user's EEG signals for 20 times, 10 times for falling asleep and 10 times for waking up. Each record contains five minutes. The dataset is experimented for finding the fatigue symptom. As a result, the signals can be divided from each record into 8 signal waves as shown in Figure 5.

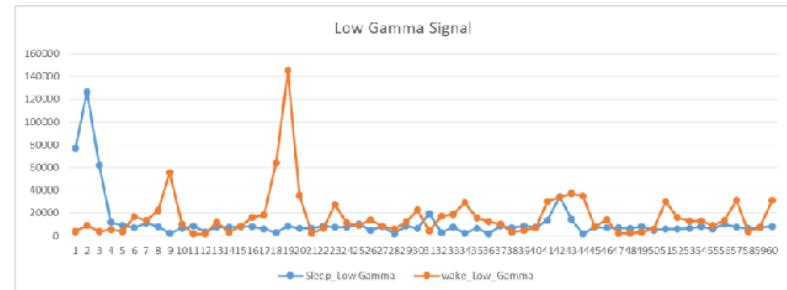




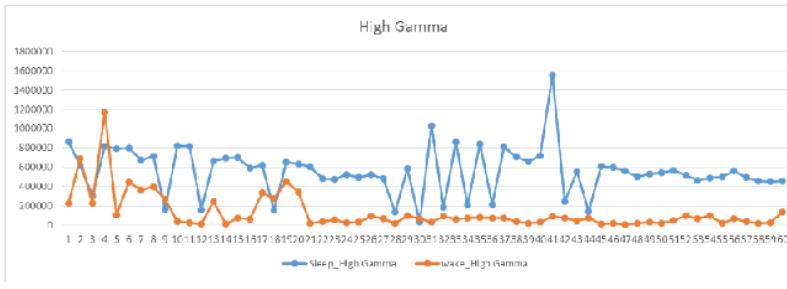
(e)



(f)



(g)



(h)

**Fig. 5.** (a) The graph of signals including (a) delta signal (b) theta signal (c) low alpha signal (d) high alpha signal (e) low beta signal (f) high beta signal (g) low gamma signal (h) high gamma signal

The performance of classification is conducted by the ANN technique. It considers the accuracy of precision and recall as shown equation (2), (3), and (4). Table 1 shows the experimental results of the performance of predicted class. Table 2 shows the comparison of the performance of classification.

$$\text{Precision}(p) = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}(r) = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Accuracy}(A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

**Table 2.** The result of data experimental classification

	PREDICTED CLASS	
	Drowsiness	Wake
ACTUAL CLASS		
Drowsiness	3546	595
Wake	667	2993

**Table 3.** The comparison of the performance of classification

Model	10-fold cross validation		
	Accuracy	Recall	Precision
ANN	83.77	0.838	0.838
Bayesnet	77.07	0.771	0.777
SMO	75.40	0.754	0.759
NaiveBaye	62.10	0.621	0.712

From the experiment, it was found that the ANN technique has higher performance than Bayesnet, Support Vector Machines (SMO), and Naïve Bayes techniques. the values of the accuracy, recall, and precision are 83.77, 0.838, and 0.838 percent, respectively.

#### 4 Conclusion

Researcher proposes the development of algorithm for monitoring fatigue in military mission based on brain signals. The ANN was applied to analyze the data. As a result, ANN performs higher performance than Bayesnet, Support Vector Machines (SMO), and Naïve Bayes. The experimental result of ANN technique showed the

percentage of its accuracy, recall, and precision values at 83.77, 0.838, and 0.838, respectively.

## References

1. Sicard, B.: Risk propensity assessment in military special operations. In: *Military medicine*. Vol. 166, No.10 , 871.(2001).
2. Lin, C.T., Ko, L.W., Chung, I.F., Huang, T.Y., Chen, Y.C., Jung, T.P., Liang, S.F.: Adaptive EEG-Based Alertness Estimation System by Using ICA-Based Fuzzy Neural Networks. In: IEEE Transactions on Circuits and Systems, vol. 53, no.11, (2006).
3. Department of the Army.: Leaders' Manual for Combat Stress Control, FM22-51, Washington DC, USA, Sept(1994).
4. Edwards,D.J., Sirois, B., Dawson,T., Aguirre,A.et al: Evaluation of fatigue management technologies using weighted feature matrix method. In: Processing of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,Stevenson , Washington(2007).
5. Cai, H., Lin, Y.: An Experiment to Non-Intrusively Collect Physiological Parameters Towards Driver State Detection. In: Proceedings of SAE 2007 World Congress, No.2007-01-0403.SAE Technical Paper, (2007)
6. How, J. M., Foo, S. C., Low, E., Wong, T. M., Vijayan, A., Siew, M.G., Kanapathy, R.: Effects of sleep deprivation on performance of Naval seamen: In Total sleep deprivation on performance. Annals of the academy of medicine. Vol. 23,No.5, pp. 669-675. Singapore(1994).
7. Rau, P. S.: Drowsy driver detection and warning system for commercial vehicle drivers: field operational test design, data analyses, and progress. In: National Highway Traffic Safety Administration, pp.05-0192. (2005).
8. Fazli, S., Esfehani, P.: Tracking Eye State for Fatigue Detection. In International Conference on Advances in Computer and Electrical Engineering In: ICACEE 2012, pp. 17-20. (2012).
9. Jung, S. U., Yoo, J. H.: Robust Eye Detection Using Self Quotient image. In: Intelligent Signal Processing and Communications. ISPACS'06, pp. 263-266.IEEE.Japan(2006).
10. Brandt, T., Stemmer, R., Rakotonirainy, A.: Affordable visual driver monitoring system for fatigue and monotony. In: Systems, Man and Cybernetics, IEEE International Conference on, Vol. 7, pp. 6451-6456. IEEE, (2004).
11. Von Jan, T., Karnahl, T., Seifert, K., Hilgenstock, J., Zobel, R.: Don't sleep and drive-VW's fatigue detection technology. In: Proceedings of 19th International Conference on Enhanced Safety of Vehicles, Washington, DC. (2005).
12. Nakagawa, T., Kawachi, T., Arimitsu, S., Kanno, M., Sasaki, K., Hosaka, H.: Drowsiness detection using spectrum analysis of eye movement and effective stimuli to keep driver awake. In: DENSO Technical Review, Vol. 12, No 1. (2006).
13. US Army Safety Center.: Sustaining Performance in Combat, Flight fax, (31)5.9-11(2003).
14. US Army Safety Center.:Fatigue, Countermeasure, (23)3.4-5(2002).

15. Raudonis, V., Simutis, R., Narvydas, G.: Discrete eye tracking for medical applications. In: Applied Sciences in Biomedical and Communication Technologies, ISABEL 2009. 2nd International Symposium on. pp. 1-6. IEEE. (2009).
16. Sun, Y., Yu, X., Berilla, J.: An Innovative Non-invasive ECG Sensor and Comparison Study with Clinic System. In: Bioengineering Conference (NEBEC), 39th Annual Northeast. pp. 163-164. IEEE. (2013).
17. Chen, Y. W., Kubo, K.: A robust eye detection and tracking technique using gabor filters. In: Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2007. Third International Conference on Vol. 1, pp. 109-112. IEEE. (2007).
18. Adeli, H., Ghosh-Dastidar, S., Dadmehr, N.: A spatio-temporal wavelet-chaos methodology for EEG-based Diagnosis of Alzheimer's disease. In: Neuroscience Letters, vol. 444, no. 2, pp. 190-194. (2008).
19. Dauwels, J., Vialatte, F., Musha, T.: A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. In: NeuroImage, vol. 49, no. 1, pp. 668-693. (2010).
20. Morison, G., Tieges, Z., Kilborn, K.: Multiscale Permutation Entropy Analysis of the EEG in Early Stage Alzheimer's Patients. In: Conf Proc. IEEE Enf Med. Biol. Soc. CA. (2012).
21. Kanton, J., Farkas, T., Dukan, P., Kovari, A.: Evaluation Of The NeuroskyMindFlex EEG Headset Brain Waves Data. In: IEEE 12th International Symposium on Applied Machine Intelligence and Informatics. (2014).
22. NeuroSky Inc.: The brain wave signal (EEG). NeuroSky Inc. (2009).
23. Onchira, O., Mawaporn, W.: Indoor localization of a wireless ad hoc sensor networks. In: The 9<sup>th</sup> National Conference on Computing and Information Technology. (2013).
24. Gy., Buzsaki.: Rhythms of the Brain. Oxford University Press. (2006)

## Arrival Time Prediction and Train Tracking Analysis

Somkiat Kosolsombat<sup>1</sup>, and Wasit Limprasert<sup>2</sup>

<sup>1, 2</sup> Department of Computer Science, Faculty of Science and Technology,  
Thammasat University, Pathumthani, Thailand

{somkiat.k, wasit\_1}@sci.tu.ac.th

**Abstract.** Rail transportation is a convenient and safe in many countries. However, Rail transportation in some countries has significant long delays. Arrival time prediction and rescheduling the time table are partial solutions to tackle the delay problem. In this paper, the relationship between measurable properties and the delay time are studied in order to develop an arrival time prediction. The result of this experiment has three parts. The relationship between properties and arrival late are then visualized and discussed. Some properties from the acquired database show that week, day and station, are important features and impact on the delay. Various regression methods are compared in our experiment and the result shows that best RMSE is  $\pm 3.863$  minutes by applying Random Forest Regression on train tracking dataset.

**Keywords:** arrival time prediction, train tracking analysis, arrival regression

### 1 INTRODUCTION

According to the data acquired from State Railway of Thailand (SRT) website and train tracking database system. The data has been collected for 1 year in 2015 consisting over 975,386 records. The mean difference between schedule time and actual arrive time is about  $18\pm 16$  minutes. This result occurred in huge economic impact and transportation delay.

One of the cause the arrival delay is that many trains running on the same track which has possibility to have more than one train in the same location. Mostly the railway agents have to solve the route conflict manually in real-time. From our preliminary analysis, we found that route conflict is the main reason that causes arrival delay about 50,000 incidents annually.

Many researches attempted to find solutions to reduce the arrival delay. In 2014 a study [1] used heuristic approach to find suitable solution for scheduling in order to prevent deadlock in a single track railway problem. The solution attempts to solve the route conflict and find the near-optimal travel strategies. Similar approach also considered in [2], which proposed a train scheduling system for double-track railway.

---

The system is able to predict a route conflict and the system is also able to reschedule to minimize the conflict using a stochastic graph.

One of the most important elements in train scheduling system is the ability to predict the arrival time of all trains arriving all stations. There are many studies attempting to find the best machine learning method to predict the arrival time. For example, In 2014 a study [3] developed train arrival time prediction model by comparing k-NN and moving average of time series data. In this experiment used arrival records collected from three different routes (#75, #201 and #407). The result shows there is no significant difference between k-NN and the moving average technique. A study in [4] compares performance between SVR and ANN implemented on MATLAB by using train arrival delay records and other related information from Serbian Railways Network consists 727 routes of the passenger trains. The result of SVR is better than ANN in this particular experiment. In [5], the data retrieved from Iranian Railways between 2005 and 2009 is used evaluating the delay prediction model using ANN, decision tree and logistic regression. The result suggested that ANN is outperform other two methods. From the related studies, SVR and ANN are likely to be the suitable regression method for this type of problem.

In this paper, we are going to compare three regression methods in order to find the possible good candidate for a train arrival time prediction system.

## 2 OUR DATA ANALYSIS METHOD

In this paper, the dataset in our experiment is acquired from the official web site, expert interview and train tracking database system of State Railway of Thailand (SRT). The data has been constantly collected for 1 year of 2015 consisting around 975,386 records and a summarized histogram of arrival late is shown in Figure 1. The original dataset consists many tables. Some table are dropped out because it is unrelated or consisting of too many missing data. For example, in *train\_tracking* table, *arrive\_note* field and *leave\_note* field are all empty.

There are three tables; *train\_running*, *train\_tracking* and *time\_table*. The *train\_tracking* table and *train\_running* table are merged by *train\_running* id. After merging, the number of record is 975,386 records. After that, we merge it with *time\_table* by *time\_table* id. The number of record after merging is 686,445 records. Then the merged table are validated and dropped all records containing Null or NAN or duplicate values. All fields containing date-time are converted to numeric data type. The *started\_on* id is converted to *week* and *day*, to represent index of week in year and index of day in week. Finally, the number of record after cleaning is 323,543 records. The result after this pre-processing is shown in

Table 1, which contains 11 fields. The arrive time is set to be output of the regression called  $y$  and the remaining columns are formed set of vector  $x$ .

We divide our study into three experiments. In the first experiment, the data is analyzed for basic visualization to represent important characteristics. In the second experiment the ExtraTrees classifier from Scikit-learn [6] is applied to extract key features that have significant impact on the arrival late. In the final experiment, three regression techniques are compared used the acquired for the evaluation.

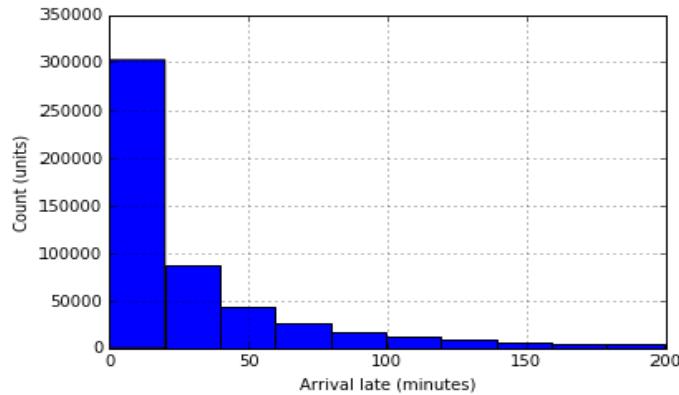


Figure 1: The histogram of arrival late (minutes)

### 3 THE EXPERIMENT AND RESULT

#### 3.1 Preliminary visualization

In this experiment, we defined two properties to express basics characteristic of the train arrival data.

*Delay magnitude* is the function of arrive late respected to value of particular feature. The delay magnitude is illustrated by a cumulative density function (CDF), where the 0.9 line expresses 90% of the trains arrive to the station by the time.

*Activity function* is the function to represent the relation between the number activities respected to value of particular feature. The activity function is computed by calculating a histogram of the number of arrival, where the bin is interval on the particular feature.

**Table 1: List of feature after pre-processing**

Symbol	Feature	Original data type	After pre-processing
x1	<i>default leavetime</i>	Date-time	Integer
x2	<i>default arrivetime</i>	Date-time	Integer
x3	<i>leave time</i>	Date-time	Integer
x4	<i>arrive time</i>	Date-time	Integer
x5	<i>leave cause</i>	Integer	Integer
x6	<i>arrive cause</i>	Integer	Integer
x7	<i>train</i>	Integer	Integer
x8	<i>station</i>	Integer	Integer
x9	<i>day</i>	Integer	Integer
x10	<i>week</i>	Integer	Integer
y	<i>arrive late</i>	Integer	Float

From Figure 2 to Figure 5, the figures show multiple-axis graph of the delay magnitude and activity function. The delay magnitude has a unit in minutes as represent in the left vertical axis. the activity function shows the number of arrival and the unit is in right vertical axis.

Figure 2 shows the delay magnitude and the activity function varying respect to week. The horizontal axis represent the index of week in a year. The dashed line indicates a number of arrival per week. The red solid line represents the delay magnitude with 0.9 confidence, 90% of trains arrive late less than the graph. Between 8th to 26th week, we found there is large magnitude of delay, while the activity is low, which in the same period of a long holiday. According to our interviews with some officers in SRT, during this period SRT needs to increase the number of carriage of the train. This increase loading time and reduce maximum speed of the trains.

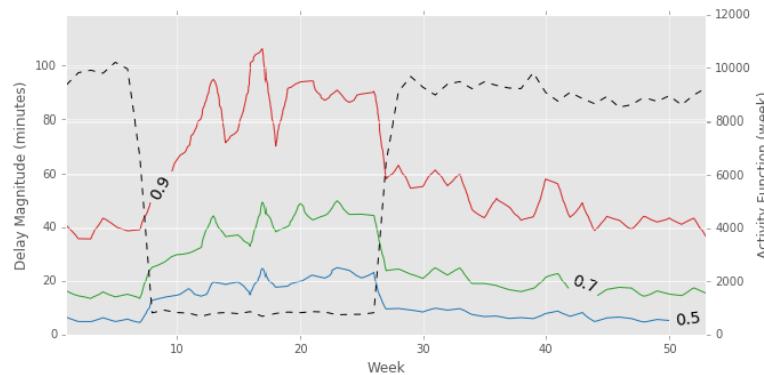
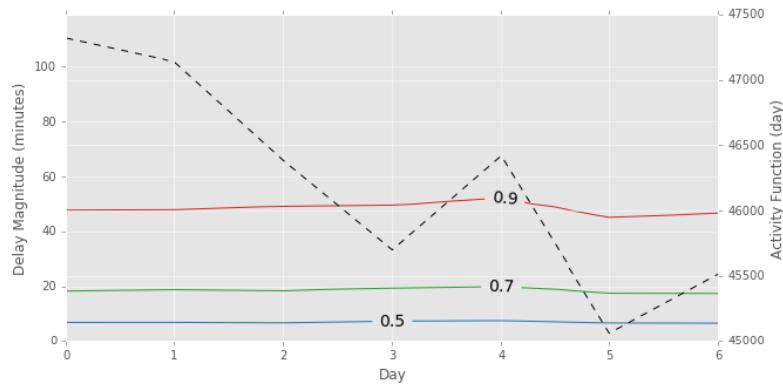
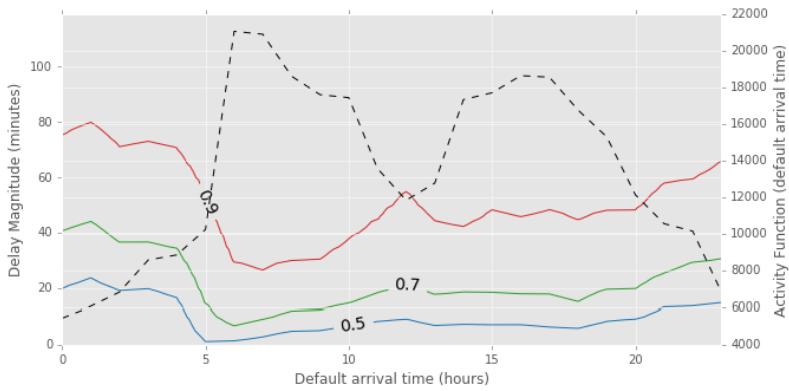
**Figure 2: Relationship between number of arrival per week and arrival late (minutes)**

Figure 3 shows the relationship of delay magnitude and activity function with the horizontal axis is index of day in week, where 0 is Monday and 6 is Sunday. The dashed line indicates a number of arrival per day. From the data, Monday Tuesday and Friday are high activity day. The magnitude of delay is almost the same during the week only about 10 minutes different between Friday and Saturday. In summary, more than 90% of arrival has delay time least than 50 minutes.

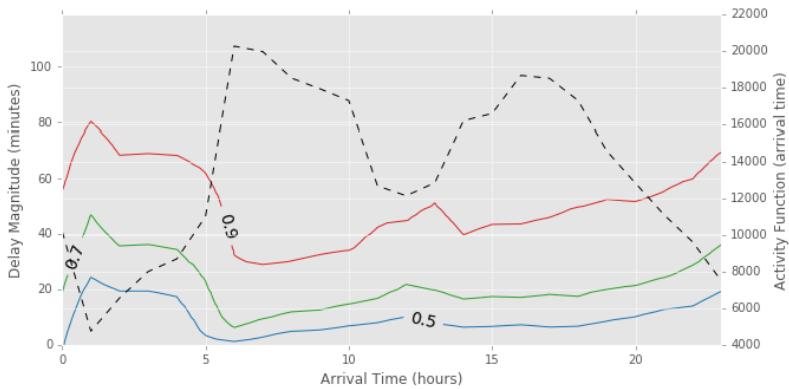


**Figure 3: Relationship b between number of arrival per day and arrival late (minutes)**

In Figure 4 and Figure 5 the horizontal axes are the default arrive time and the actual arrive time, respectively. The graphs in both figures have similar patterns. The activity function in both graphs are peak higher 18,000 arrival per hour at 6am and 4pm and the activity is dropped to 12,000 arrival per hour around midday. The delay magnitude is an increasing function with time between 6am and 4am of the next day. There are an abnormal characteristic around the midday, where the activity is off peak but the delay magnitude reaching the peak. This requires further investigation.



**Figure 4: Relationship between number of arrival per default arrival time and arrival late (minutes)**



**Figure 5: Relationship between number of arrival per arrival time and arrival late (minutes)**

### 3.2 Feature Importance

The feature importance score (FIS) is average impurity reduction for all partition in the decision trees [6], [7]. The feature receiving high FIS is the most frequently occurring and producing large impurity reduction.

After pre-processing data, All features ( $X$ ) is transformed to matrix. Similarly, the field *arrive\_late* is converted to a vector ( $y$ ). The ExtraTrees classifier [6] is applied to extract the feature importance score (FIS). The result ranking is sorted from most influent features, which is on the top to the least as shown in Table 2. *Week* is the most impact feature for predicting arrival time about receiving a FIS of 0.273. *Day*, *station* and *train* features have FIS at the same figure.

**Table 2:** The most influent features sorted by the number of impacts

Ranking	Feature	FIS
1	<i>week</i>	0.273
2	<i>day</i>	0.117
3	<i>station</i>	0.113
4	<i>train_no</i>	0.107
5	<i>arrive_time</i>	0.072
6	<i>leave_time</i>	0.054
7	<i>arrive_cause</i>	0.049
8	<i>default_arrivetime</i>	0.041
9	<i>leave_cause</i>	0.041
10	<i>default_leavetime</i>	0.041

### 3.3 Regression comparison

For most of regression studies, Root Mean Square Error (RMSE) (1) or Mean Absolute Error (MAE) (2) are commonly used to represent the error of model to fit the training data.

$$\text{RMSE} = \sqrt{\sum_{i=1}^N \frac{(y_{pi} - y_i)^2}{N}} \quad (1)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_{pi} - y_i| \quad (2)$$

After pre-processing data, all features ( $y_i$ ) is transformed to matrix. Similarly, the field *arrive\_late* is converted to a vector ( $y_{pi}$ ). The ExtraTrees classifier [6] from Scikit-learn is applied to fit the training data and to find out the RMSE and MAE value by using the estimator equal 100. The result is a prediction error of model as shown in Table 3. The Random Forest algorithm has the best result of RMSE equal 3.863. ANN and Linear have the value 124.907 and 25.380, respectively. The result of MAE, Random Forest has the prediction error less than other methods. Random Forest, ANN and Linear have the value are 2.001, 60.582 and 14.976.

**Table 3:** The result of RMSE and MAE by regression method

Regression	Random Forest	ANN	Linear
RMSE	3.863	124.907	25.380
MAE	2.001	60.582	14.976

## 4 CONCLUSION

This paper examines the arrival time prediction and train tracking analysis. All data retrieved from the official web site, expert interview and train tracking database of State Railway of Thailand (SRT) for 1 year of 2015 consisting 975,386 records. The relationship between many properties and the delay magnitude (minutes) are studied. The comparison of three regression methods; Random Forest Regression, ANN, Linear regression, are also studied. The mean RMSE are 3.863, 124.907 and 25.380, respectively. We also found the accuracy of the classifier are affected by the importance feature. The most important feature is week receiving importance score of 0.273. Whereas, *day*, *station* and *train\_no* receives no different importance score at 0.11. In future work, further analysis on arrival time and leave time will be examined in order to improve the accuracy of the current time table.

## 5 DISCUSSION

This experiment retrieves data from three tables; *train\_running*, *train\_tracking* and *time\_table*. The total after cleaning data is 323,543 records from all 975,386 records. It has many lost of data. Data is not complete or containing Null or NAN or duplicating value. It may have caused by data log recording, missing data, converting data, assigning the wrong data type, and so on. Then, data store or data pre-processing method are important to do experiment.

The result of regression model RMSE equal  $\pm 3.863$  minutes to indicate the value of prediction error for arrival time. The optimize this value may be consisting of the corrective data, rescheduling the time table, improving the recorded data, and adjusting the relationship between the actual time and default time table.

## 6 ACKNOWLEDGEMENT

We would like to very thank Mr. Chokdee Suwanrat and his department of information technology and State Railway of Thailand for providing the data and consulting on a workflow of train operations and many definitions of the technical term.

## 7 REFERENCES

- [1] F. Li, J.-B. Sheu, and Z.-Y. Gao, “Deadlock analysis, prevention and train optimal travel mechanism in single-track railway system,” *Transp. Res. Part B Methodol.*, vol. 68, pp. 385–414, Oct. 2014.
- [2] P. Kecman and R. M. P. Goverde, “Online Data-Driven Adaptive Prediction of Train Event Times,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 465–474, Feb. 2015.

- [3] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah, “Improving arrival time prediction of Thailand’s passenger trains using historical travel times,” in *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2014, pp. 307–312.
- [4] N. Marković, S. Milinković, K. S. Tikhonov, and P. Schonfeld, “Analyzing passenger train arrival delays with support vector regression,” *Transp. Res. Part C Emerg. Technol.*, vol. 56, pp. 251–262, Jul. 2015.
- [5] M. Yaghini, M. M. Khoshraftar, and M. Seyedabadi, “Railway passenger train delay prediction via neural network model,” *J. Adv. Transp.*, vol. 47, no. 3, pp. 355–368, Apr. 2013.
- [6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” *ArXiv12010490 Cs*, Jan. 2012.
- [7] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

## The Limb Leads ECG Signal Analysis in Myocardial Infarction Patients

Anchana Muankid<sup>1</sup> and Mahasak Ketcham<sup>2</sup>

<sup>1</sup>Department of Information Technology, Faculty of Information Technology,  
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand  
muan.anchana@gmail.com, s5607011910072@email.kmutnb.ac.th

<sup>2</sup>Department of Information Technology management, Faculty of Information Technology,  
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand  
mahasak.k@it.kmutnb.ac.th

**Abstract.** Cardiovascular disease is one of the most serious diseases in the world. An electrocardiogram is a tool in the diagnostic of Myocardial Infarction which detects abnormal wave patterns. This paper purposed the limb leads; I, II and III electrocardiogram analysis algorithm using Wavelet transform to classify Inferior Myocardial Infarction patients. And investigate the lead which relates to inferior infarcts. The processes in ECG signal analysis are noise elimination from the ECG signal, R peak Detection, QRS Complex Detection and inferior Myocardial Infarction Classification. The results show that 73.33% accuracy of inferior Myocardial Infarction. Lead III and Lead II are the most relevant to inferior Myocardial Infarction.

**Keywords:** ECG Analysis · Wavelets Transform · Limb leads

### 1 Introduction

Cardiovascular disease (CVD) is the most common cause of death in the world [1]. Myocardial Infarction (MI) is the necrosis of heart muscle when blood flow is obstructed. A Myocardial Infarction might cause acute heart failure, and cardiac arrest. The electrocardiogram is a tool in the diagnostic of Myocardial Infarction [2] which detects abnormal wave patterns. The electrocardiogram patterns were identifying Myocardial Infarction type of patient. There are involved to treatment decision making of physicians.

The electrocardiogram analysis by computerize system has been interested for 10 years ago [3] and has been increasing steadily. The researches mainly focused on analyzing electrocardiogram pattern, classifying diseases and applying a new technique to enhance the electrocardiogram signal quality. The electrocardiogram can help in identifying proximal occlusion of the coronary arteries, which results in the most extensive and most severe myocardial infarctions [2].

This paper purposed a limb leads; I, II and III electrocardiogram analysis algorithm using Wavelet transform to classify Inferior Myocardial Infarction patients. And investigate the lead which relate to inferior infarcts. The three steps of electrocardiogram signal analysis are noise elimination, feature extraction and inferior Myocardial Infarction classification. Accuracy and percentage of relate to inferior infarct are used to evaluate the algorithm. The remaining of the paper is organized as follows: the first section is Introduction, section 2 covers Electrophysiology, Wavelet Transform and related works, section 3 is Electrocardiogram Analysis, section 4 is Results and section 5 is the Conclusion.

## 2 Background and Notations

### 2.1 Electrocardiogram Signal

Electrocardiogram (ECG) is signal showing the cardiac electrical activity [4] that's detected by attaching electrodes to the skin on chest, arm and legs. An ECG normal waveform, the P wave occurs first, followed by the QRS complexes and the T wave. The ranges between the waves are called segments. The X-axis shows the record speed (millimeters/second), and the Y-axis shows the energy (amplitude).

The ECG lead placements are divided into three types, Limb Leads (Bipolar), Augmented Limb Leads (Unipolar) and Precordial Leads (Unipolar) [5]. The Limb Leads are the view of signal from electrodes that are attached to the limb, consisting of Lead I, II, and III. The various characteristic features of ECG are used to identify the cardiac abnormal area and support physicians treatment decisions making.

### 2.2 Wavelet Transform

A wavelet based signal technique is an effective tool for non-stationary ECG signal analysis and characterization of local wave (P, T and QRS complex morphologies) [6]. Even if a signal is not represented well by one member of the Daubechies family, it may still be efficiently represented by another [7]. This paper, selection of the wavelet decomposition of ECG signal at level 4, using Daubechies 4 is then undertaken since this waveform resembles the original ECG signal.

### 2.3 Related Works

ECG varies in time, researchers have developed a computerize system to monitor patients heart health accurately and easily [8]. The researches which concentrated on classifying Cardiovascular Disease patients were divided into two types, Arrhythmia and Myocardial Infarction (MI). The most common technique used for classification is data mining such as Neural Network, Least Square Support Vector Machine (LS-SVM) [8, 9] and Multi-layer back propagation neural network [10].

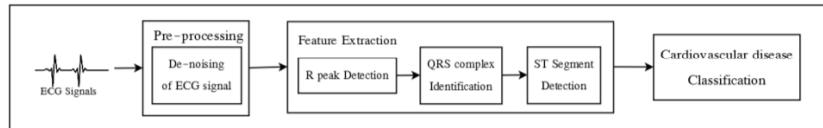
The survey of the ECG signal used in the analysis of Myocardial Infarction shows that the classification of Myocardial Infarction by ECG waveform analysis

based on Support vector machines (SVMs) and Gaussian mixture model (GMMs) provides 82.50% accuracy [11]. Banerjee analyzed the ECG signal from PTB-DB and classified IM patients using a Cross Wavelet Transform (XWT) technique [12] which provided accuracy up to 92.50%. The PTB-DB contains an ECG data associated with Myocardial Infarction and 15 leads of ECG signal, PTB-DB is appropriate for this research purpose [13]. The studies are appended by adjusting the XWT parameters for higher accuracy. The classification accuracy is up to 97.6% [14].

However there are many Myocardial Infarction Classification researches, the researches which concentrated on investigate the location of infarct area is not found.

### 3 Proposed Method

The processes in ECG signal analysis are Pre-processing, Feature Extraction and inferior Myocardial Infarction Classification. The ECG analysis process is shown in Figure 1.



**Fig. 1.** ECG Analysis Process

#### 3.1 Preprocessing Stage

The ECG signal data from the PTB Diagnostic ECG database is used to analyze ECG signal waveform [15]. The samples are 30 inferior Myocardial Infarction patients which inferior infarct, using 60 seconds of ECG signal from limb leads; lead I, II and III.

In this process, the noise is removed to improve ECG signal quality [16]. A wavelet based signal technique is an effective tool for non-stationary ECG signal analysis and characterization of local wave (P, T and QRS complex morphologies) [6]. Selection of the wavelet decomposition of ECG signal at level 4, using Daubechies4 is then undertaken since this waveform resembles the original ECG signal.

#### 3.2 R-peak Detection

R-peak detection is the most important task in ECG signal analysis as there is an obvious peak detected first [17]. Assign the threshold of R-peak. The peak which has a value in criteria range is R-peak. The R-peak detection algorithm is as follows.

---

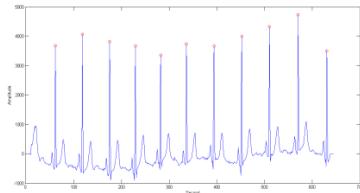
*Algorithm 1: R-peak Detection*

---

*If ( $R_x > R_{x-1}$ ) (and) ( $R_x > R_{x+1}$ ) (and) ( $R_a \geq Threshold$ ) then  
R peak = ( $R_x, R_a$ ) end*

---

Where  $R_a$  is R peak amplitude and  $R_x$  is R peak position. The results of R-peak detection is shown in figure 2



**Fig. 2.** R-peak detection result

### 3.3 QRS Complex Detection

QRS complex identifies by calculating the lowest point in the previous position and the next position. The QRS complex detection algorithm is as follows.

---

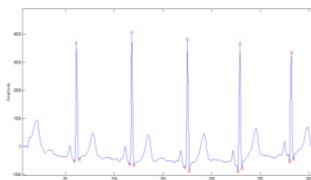
*Algorithm 2: QRS Complex Detection*

---

```
If (Qx < Rx) (and) (Qx < Qx-1) (and) (Qx < Qx+1) then
    Qpoint = (Qx, Qa) end
If (Sx > Rx) (and) (Sx < Sx-1) (and) (Sx < Sx+1) then
    Spoint = (Sx, Sa) end
```

---

Where Qa is Q wave amplitude, Qx is Q wave position, Sa is S wave amplitude and Sx is S wave position. The detection of QRS complex is shown in Figure 3.



**Fig. 3.** QRS complex Identification result

Choosing the appropriate features are very important because this will affect the classification of the ECG signal. If the feature is not appropriate, classification results in an error [18]. This paper select abnormal of Q-Wave, ST-Segment and T-wave feature to classifying Inferior Myocardial Infarction patients.

### 3.4 ST Segment Detection

In normal ECG pattern, T wave could not be invert. The normal ST segment detection algorithm is as follows.

---

*Algorithm 3: ST segment detection*

---

```
If (Tx > Sx) (and) (Tx > Tx-1) (and) (Tx > Tx+1) then
    Tpoint = (Tx, Ta) end
```

---

Where Tx is T wave position, Ta is T wave amplitude and Sx is S wave position.

### 3.5 ECG Signal Classifications

Analyses of the ECG signal to classify inferior Myocardial Infarction patients by detected developing Q-wave, elevated ST-segment or inverted T-wave following the QRS complex. The inferior Myocardial Infarction classification algorithm is as follows.

---

*Algorithm 4: Inferior Myocardial Infarction Classification*

---

*If ( $Qa < threshold$ ) or ( $Sa > 0$ ) or ( $Ta < Sa$ ) then  
Inferior Myocardial Infarction end*

Where  $Qa$  is Q wave amplitude,  $Sa$  is S wave position and  $Ta$  is T wave amplitude.

## 4 Results and Discussion

This algorithm classifies 22 correctly out of 30 Myocardial Infarction patients or 73.33% accuracy. About the leads that relate to inferior infarct area, Lead II is the most relevant to inferior infarct area, 66.67% relate to inferior infarct. Second is Lead III, 53.55% relate to inferior infarct. Lead I is the less relevant to inferior infarct area, 16.67% relate to inferior infarct. The overall result provides 73.33% accuracy of Myocardial Infarction patient classification. The classification result is presented in Table 1.

**Table 1.** ECG Analysis Results

	Lead I	Lead II	Lead III
% Relate to inferior infarct	16.67	66.67	53.53
% Accuracy		73.33	

## 5 Conclusions

Electrocardiogram (ECG) is a signal showing the cardiac electrical activity. The electrocardiogram is a tool in the diagnostic of Myocardial Infarction which detects abnormal wave patterns. The electrocardiogram patterns were identifying Myocardial Infarction type of patient. This paper purposes a limb leads; I, II and III electrocardiogram analysis algorithm using Wavelet transform to classify Inferior Myocardial Infarction patients. And investigate the lead which relate to location of infarcts. The steps in ECG signal analysis are noise elimination of ECG signal, R peak Detection, QRS Complex Detection and inferior Myocardial Infarction Classification.

In the feature extraction stage, a threshold base is applied in the algorithm for ECG signal classification. The analysis of ECG signals to classify inferior Myocardial Infarction patient using Wavelet transform, the results show that 73.33% accuracy of inferior Myocardial Infarction. Lead III and Lead II are the most relevant to inferior Myocardial Infarction. In future work, variable values in the feature extraction algorithm will be adjusted to further improve the classification performance.

## References

1. WorldHealthOrganization. [cited 2014 Oct, 13]; Available from: <http://www.who.int/mediacentre/factsheets/fs317/en/>.
2. Zimetbaum, P.J. and M.E. Josephson, *Use of the electrocardiogram in acute myocardial infarction*. New England Journal of Medicine, 2003. **348**(10): p. 933-940.
3. Maglaveras, N., et al., *ECG pattern recognition and classification using non-linear transformations and neural networks: A review*. International Journal of Medical Informatics, 1998. **52**(1-3): p. 191-208.
4. Klabunde, R., *Cardiovascular physiology concepts*. 2011: Lippincott Williams & Wilkins.
5. Acharya, R., et al., *Advances in cardiac signal processing*. 2007: Springer.
6. Manikandan, M.S. and S. Dandapat, *Wavelet-based electrocardiogram signal compression methods and their performances: A prospective review*. Biomedical Signal Processing and Control, 2014. **14**: p. 73-107.
7. Saritha, C., V. Sukanya, and Y.N. Murthy, *ECG signal analysis using wavelet transforms*. Bulg. J. Phys, 2008. **35**(1): p. 68-77.
8. Martis, R.J., et al., *Cardiac decision making using higher order spectra*. Biomedical Signal Processing and Control, 2013. **8**(2): p. 193-203.
9. Martis, R.J., et al., *Application of principal component analysis to ECG signals for automated diagnosis of cardiac health*. Expert Systems with Applications, 2012. **39**(14): p. 11792-11800.
10. Thomas, M., M.K. Das, and S. Ari, *Automatic ECG arrhythmia classification using dual tree complex wavelet based features*. AEU-International Journal of Electronics and Communications, 2015.
11. Chang, P.-C., et al., *Myocardial infarction classification with multi-lead ECG using hidden Markov models and Gaussian mixture models*. Applied Soft Computing, 2012. **12**(10): p. 3165-3175.
12. Banerjee, S., R. Gupta, and M. Mitra, *Delineation of ECG characteristic features using multiresolution wavelet analysis method*. Measurement, 2012. **45**(3): p. 474-487.
13. Sun, L., et al., *ECG analysis using multiple instance learning for myocardial infarction detection*. Biomedical Engineering, IEEE Transactions on, 2012. **59**(12): p. 3348-3356.
14. Banerjee, S. and M. Mitra, *Application of Cross Wavelet Transform for ECG Pattern Analysis and Classification*. 2014.
15. Goldberger, A.L., et al., *Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals*. Circulation, 2000. **101**(23): p. e215-e220.
16. Zadeh, A.E., A. Khazaee, and V. Ranaee, *Classification of the electrocardiogram signals using supervised classifiers and efficient features*. Computer methods and programs in biomedicine, 2010. **99**(2): p. 179-194.
17. Manikandan, M.S. and K.P. Soman, *A novel method for detecting R-peaks in electrocardiogram (ECG) signal*. Biomedical Signal Processing and Control, 2012. **7**(2): p. 118-128.
18. Rai, H.M., A. Trivedi, and S. Shukla, *ECG signal processing for abnormalities detection using multi-resolution wavelet transform and Artificial Neural Network classifier*. Measurement, 2013. **46**(9): p. 3238-3246.

## Estimating PSD Characteristics of ECG in Comparison between Normal and Supraventricular Subjects

Thaweesak Yingthawornsuk<sup>1</sup>, Siriphan Phetmuam<sup>1</sup>, Saowaros Singkhal<sup>1</sup>, Waraporn Pattarason<sup>1</sup>  
Media Technology, King Mongkut's University of Technology Thonburi, Thailand  
thaweesak.yin@kmutt.ac.th, siriphan.jaa@gmail.com, jeabsaowaros@gmail.com, krajib.pat@gmail.com

### Abstract.

The aims of project are to develop an arithmetic program that can detect irregularity in electrocardiogram (ECG) and classify between two groups of normal and supraventricular ECG waveforms by using Auto regressive (AR) estimators with various model orders starting from 3rd to 9th. All AR estimators are associated with the PSD of ECG waveforms collected from a group of 30 subjects at 200Hz sampling frequency. The best classification scores found on the 5<sup>th</sup>-order AR model are 95.99% and 72.17% obtained from training and testing the C4\_5 classifier with the fifth-order coefficients. By classifying the 7<sup>th</sup>-order AR coefficients with Linear Least Squared (LS) classifier the accurate scores of 86.43% and 80.85% were obtained from training and testing cases respectively. These performance accuracies show that the proposed method is highly effective in parameterizing and classifying PSD feature as quantitative measure that can characterize the ECG signals of normal and supraventricular cardiac conditions.

**Keywords:** ECG, PSD, AR, Supraventricular

### 1 Introduction

Cardiac Arrhythmia is a common type of heart disorder found in all aging persons and some with cardiac disorders. The cause of arrhythmia is various in each person such as the one who has a heart wall with very thickness or aging persons with cardiac disorder since they were born. The ECG signal is commonly used to represent on how our heart functions and reflects the healthiness condition of heart itself shown as a form of bioelectric signal. The shape of ECG waveform can feature a heart in terms of functionality, structural components or illness affection. This kind of signal can indicate changes in cardiac condition that mediate in its waveform.

Some of the most distressing types of heart malfunction occur not as a result of abnormal heart muscle but instead abnormal rhythm of the heart. Abnormality of any portion of the heart, including arterial and ventricle can sometimes causes a rapid rhythmic discharge of the impulse that spreads in all directions throughout a heart. This rapid heart rate, as determined from the time intervals between QRS complexes,

is approximately 150 per minute instead of 72 per minute [4]. Supraventricular ECG is categorized in rapid tachycardia with heart rate above 100 per minute which is caused by electrical impulses originated above the heart's ventricles [6].

In this work, two different types of ECG signal comprised of normal ECG and Supraventricular ECG were comparatively studied to determine any significant difference in their power spectral distribution and other distinctions.

The following sections organized in paper are methodology, experimental results and discussion, and conclusion.

## 2 Methodology

The study procedure mainly consists of data acquisition, preprocessing signal, feature extraction including a model fitting and then classification. After all algorithms designed in each step correctly, the GUI for Matlab was designed and implemented to make it friendly for users. The main task of this study is to detect QRS complexes and the estimation of Power Spectral Density via AR modeling. The QRS complex detection is a major challenge. The Hamilton-Tompkins algorithm to detect complexes is divided into following steps [1, 4].

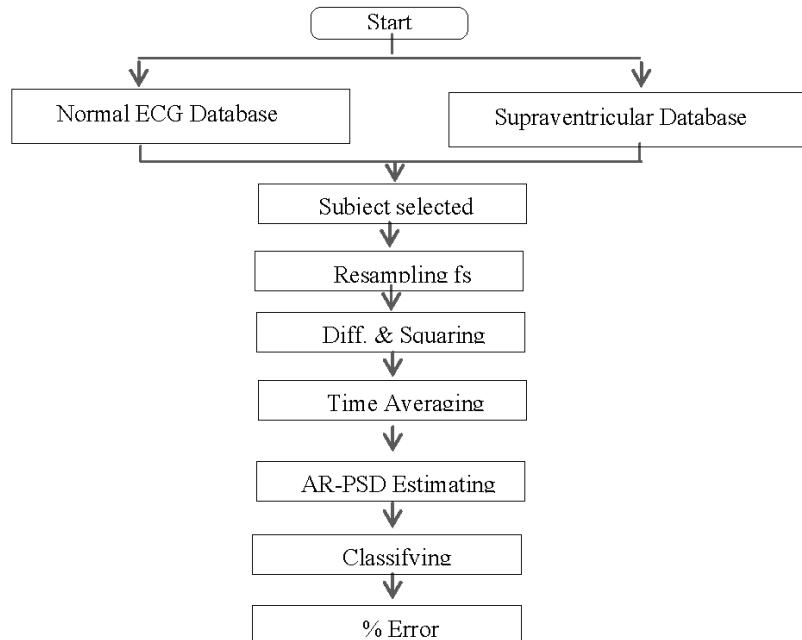


Figure 1 : Flowchart of the algorithm

The pre-processor section performs linear and nonlinear filtering of the ECG signal and produces a set of periodic vectors. Next, the decision rule operates on the output from first section by classifying it as either a QRS complex or noise, and then save it for further steps. Filtering is used to attenuate noise. The Low and high filters are combined together to form a bandpass filter. The procedure of signal processing was followed by a differentiation, squaring and time averaging of the signal. Time averaging was done by adding the 32 most recent values from squared values and divided by 32. Figure 1 shows a workflow of PSD estimation and its comparison between two different databases. All ECG signals were down-sampled at 200 Hz which is adequate in acquisition of all information and frequency response components contained in signals. All procedure steps were repeated for individual recorded ECG signals in database until all thirty ECG file completed.

The signal databases used in this study consist of fifteen normal ECG signals and another fifteen supraventricular ECG signals. In main step of PSD estimation, the AR model based on Yule Walker's technique was applied to signals to determine the best fitted model coefficients that represent the signal in a form of spectral power distributed along a low frequency range [2, 3]. The coefficients belonging to the best fitted-order of AR model were used as a set of feature input to classifier in pairwise manner. The trials of possible fitting models to signal were performed on the model orders of AR.

### 3 Experimental Results and discussion

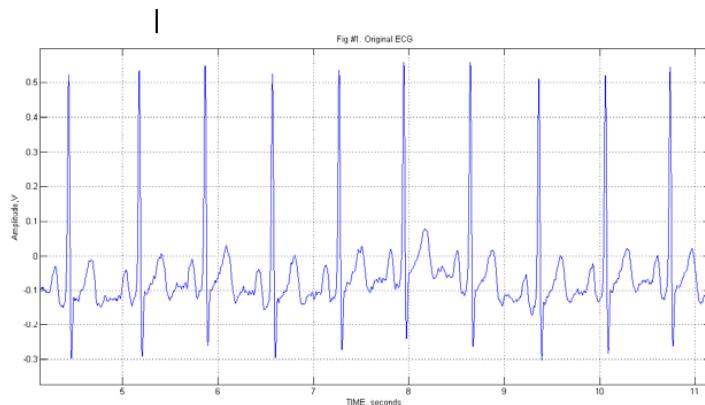


Figure 2 : Normal ECG waveform

In processing state each normal ECG signal as shown in figure 2 and abnormal ECG signal shown in figure 3 were observed visually first, then analyzed via our designed program. As one can notify, in case of supraventricular ECG the T wave in a

current complex and P wave of a following complex are messy mixed to each other and look noisy alike with the dropped amplitude level as compared to the normal ECG. The S wave obviously diminishes and there is no resetting back interval to the signal baseline. In Figures 4 and 5 show the original ECG signals from Normal and abnormal cases respectively being down resampled to have a new sampling frequency at 200 Hz, filtered by derivative filter, squared in its amplitude to have an absolute value in term of power and then time averaged with every 32 data-points window frame to have very evenly smoothed peaks. Results from processing two different groups of the categorized ECG signals via our program revealed the significant difference in term of quantitation of peaks found in state of time averaging. As one can see in case of abnormal ECG, it has a very less number of time averaged peaks as compared to normal one. In Figure 6, the middle points of positive and negative slopes in individual time averaged peaks were detected automatically as indicated in red and blue markers shown in the same lower subplot.

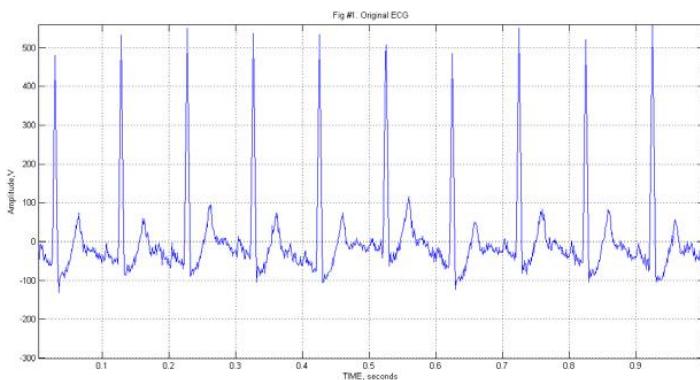


Figure 3 : Supraventricular ECG waveform

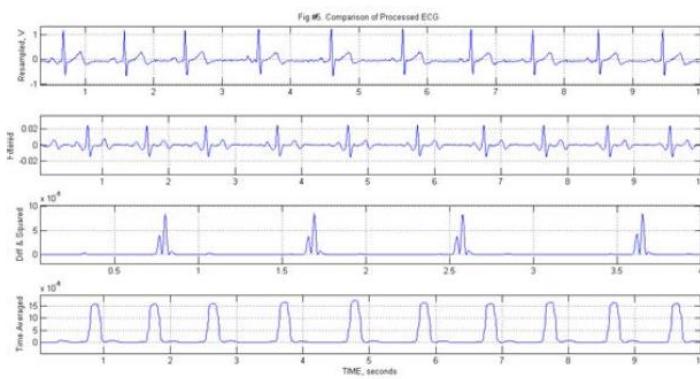


Figure 4 : Plots of the processed normal ECG signal (a) resampled, (b) filtered, (c) squared, and (d) time averaged

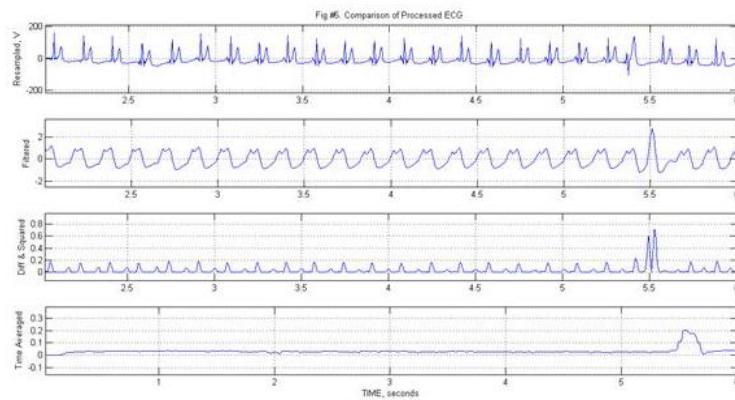


Figure 5 : Plots of the processed abnormal ECG signal (a) resampled, (b) filtered, (c) squared and (d) time averaged

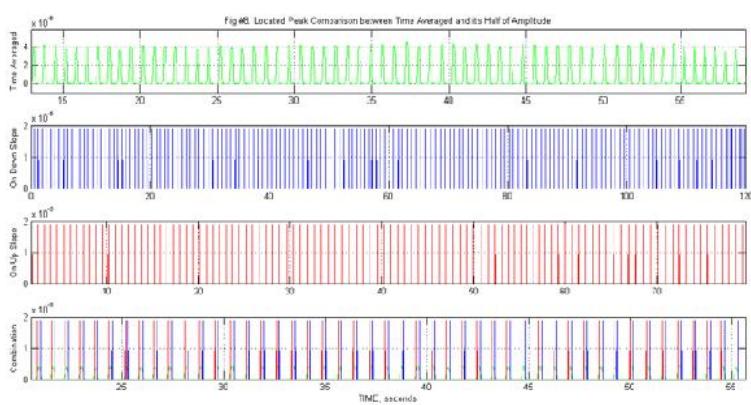


Figure 6 : Plots of (a) time averaged signal, (b) detected down-slope, (c) detected up-slope and (d) both markers

```

arAll =
Columns 1 through 12
1.0000 -2.5679 2.3331 -0.7637 0 0 0 0 0 0 0 0
1.0000 -3.2473 4.4087 -3.0482 0.8896 0 0 0 0 0 0 0
1.0000 -3.8967 6.6338 -6.2663 3.2600 -0.7300 0 0 0 0 0 0
1.0000 -4.4697 9.1930 -11.1857 8.4678 -3.7890 0.7850 0 0 0 0 0
1.0000 -5.0543 12.0147 -17.4916 16.7977 -10.6350 4.1136 -0.7447 0 0 0 0
1.0000 -5.5473 14.7375 -24.5310 27.9163 -22.2129 12.0662 -4.0002 0.6619 0 0 0
1.0000 -5.9393 17.1602 -31.6780 41.0733 -38.7482 26.5964 -12.8195 3.9476 -0.5923 0 0
1.0000 -6.0915 18.1745 -34.9720 47.9073 -48.7045 37.1502 -20.9592 8.3570 -2.1184 0.2570 0
1.0000 -6.0696 17.9939 -34.2594 46.1200 -45.5365 32.9969 -16.8739 5.3747 -0.5686 -0.2625 0.0853
1.0000 -6.0902 18.0573 -34.1220 44.8212 -41.4589 25.0231 -5.8699 -5.7703 7.7103 -4.6108 1.5520
0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0
Column 13
0
0
0
0
0
0
0
0
0
-0.2417
0
0

```

Figure 7 : The estimated AR coefficients based on Yule Walker's technique

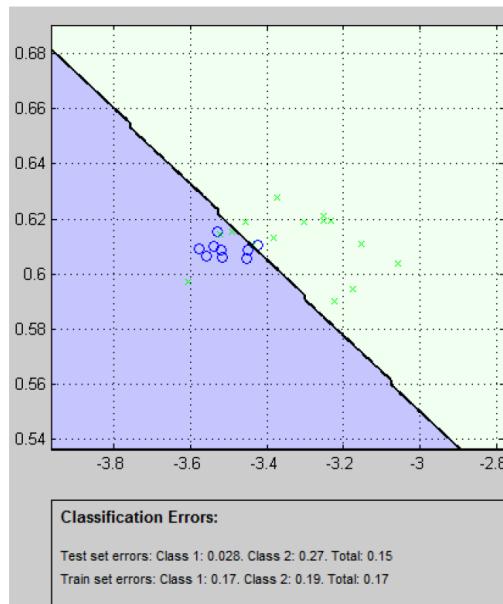


Figure 8 : Region of class discrimination and errors in classification

Figure 7 illustrates the values of the AR- model coefficients that were estimated from fitting the models to the time averaged signals. The highest model order of AR used in signal modeling is the 12-th order and the Power Spectral Density of signal was therefore estimated based on the coefficients of AR model that were found in each model order. Next step is classification made in pairwise manner and the 20% of total number of AR coefficients was randomly selected and used to train the selected classifiers and the rest of coefficients used to evaluate the classification. The best classification scores found on the 5<sup>th</sup>-order AR model are 95.99% and 72.17% obtained from training and testing the C4\_5 classifier with fifth-order coefficients. By classifying the 7<sup>th</sup>-order AR coefficients with Linear Least Squared (LS) classifier the accurate scores of 86.43% and 80.85% were obtained from training and testing cases respectively.

In order to simplify all complicated steps in executing all Matlab scripts to perform all steps in a workflow depicted in Figure 1, the Graphic User Interface (GUI) was purposely designed to be much friendly and facilitated for user. Figure 9 presents GUI screens with plots of processed signals in result of executing each functional button that was designed to callback the Matlab scripts to perform tasks in selected function and then plot out the result of that state as shown in Figure 9. User can make an observation on how original ECG signal is analyzed step by step clearly.

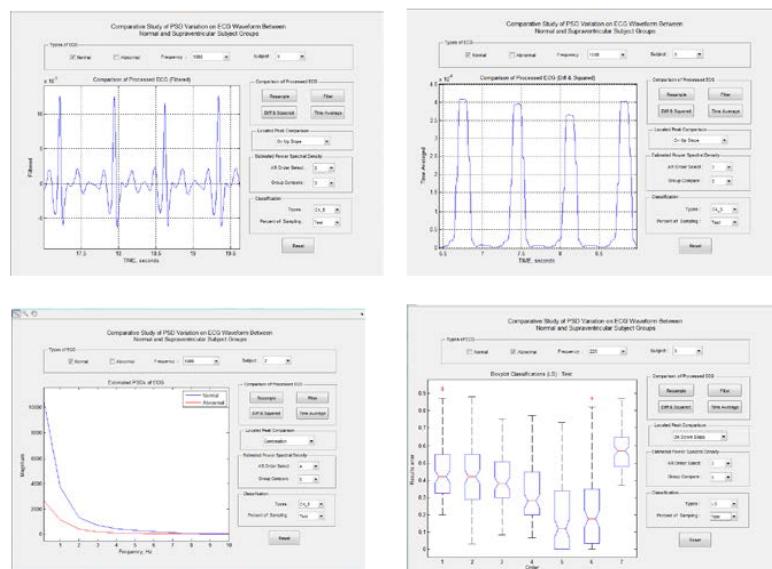


Figure 9 : Analyzed Results on GUI design

#### 4 Conclusion

Different significance can be identified in term of quantitative PSD estimated by modeling ECG with AR coefficients which best fitted to the processed waveform of ECG in each categorized signal database. The PSD characteristics obtained from estimation made on between different groups of ECG signals have revealed the certainly different frequency response at a very-low frequency range. Moreover, the Graphic User Interface (GUI) is purposely provided to be friendly facilitated in execution the processing program by user who may not be experienced with Matlab environment.

As shown in study on performance and computation cost, other alternative techniques are required in combination with the presently proposed work to gain more satisfactory expectation on accuracy and speed of program execution.

#### References

1. Patrick S. Hamilton, Willis J. Tompkins, "Quantitative Investigation of QRS Detection Rules Using the MIT/BH Arrhythmia Database", IEEE Transaction on biomedical engineering. Vol. BME-33, No.12 December 1986
2. Chusak Thanawattano, Thaweesak Yingthawornsuk, "Cardiac Arrhythmia Classification using Beat-by-Beat Autoregressive Modeling", 2013 3<sup>rd</sup> International Conference on Computer and Electrical Engineering(ICCEE 2010)
3. Junyou Huang, "Study of Autoregressive (AR) Spectrum Estimation Algorithm for Vibration Signals of Industrial Steam Turbines", International Journal of Control and Automation Vol.7, No.8 (2014),
4. C. Guyton, "Textbook of Medical Physiology" 8th Edition, Harcourt College Pub; October 1990
5. W. J. Tompkins, "Electrocardiography," in Biomedical Digital Signal Processing, W. J. Tompkins, Ed. New Jersey: Prentice Hall, 2000
6. O. A. Obel, A. J. Camm, "Supraventricular tachycardia ECG diagnosis and anatomy", European Heart Journal, No.18 (1997)

## Evolving Public Opinion Mining Methods on Decision Support System in Thai E-Government

Jeerana Noymanee<sup>1</sup>, Wimol San-Um<sup>2</sup> and Thanaruk Theeramunkong<sup>3</sup>

<sup>1</sup>*Electronic Government Agency (Public Organization) of Thailand , Bangkok, Thailand*

<sup>2</sup>*Intelligent Electronic System Research Laboratory, Thai-Nichi Institute of Technology, Bangkok, Thailand*

<sup>3</sup>*School of Information, Computer, and Communication Technology (ICT) Sirindhorn International Institute of Technology, Thammasat University P.O.Box 22, Pathum Thani 12121, Thailand.*

jeerana @ega.or.th, wimol@tni.ac.th, thanaruk@siit.tu.ac.th

**Abstract.** Public opinion mining is a combination of Natural Language Processing (NLP) and Sentiment Analysis. To make appreciate decisions in policy, it is necessary to utilize sentiment classification efficiently. While reviews usually contain sentiment which is expressed in a different way in different domains, it is costly to annotate data for each new domain. E-Government refers to the use of information and communications technologies (ICT) to improve quality of services and information offered to citizens, and government in order to obtain more accountable and transparency towards governance in public sector. Recently, have been widely discussed governmental decisions within digital societies. This paper provide and exploration of opinion mining and text mining techniques towards apprehending the public's opinion communicated online and concerning governmental decisions. Regarding the objective of study is focuses on the understanding of the citizen opinions about e-Government issues and on the exploitation of these opinions in subsequent governmental actions. This paper also examine several features in the user generated content discussing governmental decisions in an attempt to automatically extract the citizen opinions from online posts on public sector regulations and thereafter it can be able to organize the extracted opinions not only into polarized clusters but also collect the potential word that able to declare the citizen demand. The objective is to identify the public's stance against governmental decisions automatically and It can be deduced that how the citizen's attitudes may effect to government actions. To demonstrate the usability and added value of the proposed the architecture of e-Government will be presented and discussed in paper.

**Keywords:** Opinion mining, opinion classification, E-Government, Decision support system, Online Social Network

### 1. Introduction

Although the resources that provide to e-Government. It is mainly perceived as a service system to support the activities of governments and prevent the resist issues of the sociopolitical impact of those activities. But since this day, there are no concrete plan to manage public social issue.

In this paper, shown the possible solution by providing an innovative e-Government protocol that captures the societal impact of public sector regulations in a challenge to decrypt the public's attitude towards governmental decisions. Especially, propose the exploitation of data mining techniques towards firstly capturing the public's opinions about governmental decisions and secondly analyzing are the polarity of the mining opinions accordingly to they are considered in following governmental decisions. Specifically, introduce a method for decomposing citizen's opinions and comments that are posted in online fora and blogs, in order to evaluate the government decision righteousness as a result of feedback from citizen.

The findings of experimental study clearly demonstrate that e-Government services invoke the citizen's active participation in the decision making process and indicate that by putting together inter-disciplinary methods and tools can transform e-Government from a technological infrastructure to a powerful interactive manifestation of e-inclusion and e-participation. In

Section below, the introduction of the utilization of text and data mining techniques for identifying and deciphering the citizen opinions about governmental issues that are communicated online. Online content via the use of natural language processing and text mining tools in order to firstly mine user opinions from their posts and then annotate the mined opinions with a suitable polarity la-bel depending on the orientation of the latent user opinions. In Fig 1, present preliminary experiments carried out in which relied on real user comments about governmental decisions and tried to organize the mined user opinions into polarized clusters of citizen comments. And present the architecture framework of opinion mining based decision support system for e-Government.

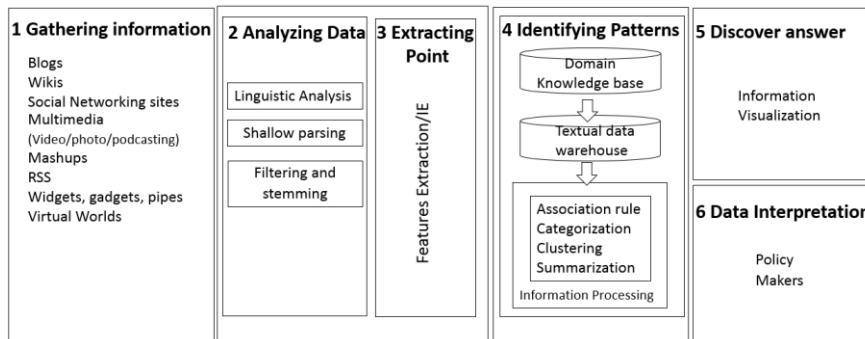


Fig. 1 Architecture framework of opinion mining based decision support system for e-government

## 2. Relate Work

Resent Problem the aim of research is to design and implement a method able to automatically detect and analyses the public's stance towards governmental decisions. In this respect, considering that people verbalize their opinion in natural language need to rely on the textual data of their comments on governmental issues. Such content can be easily harvested either manually by domain experts who indicate the data sources that need to be collected, or automatically via a trained focused crawler application to e-Government social media sites.

The main issue of existing e-Government services and applications is their failure to capitalize on societal factors. The first part of approach, relying on user online comments, concerns a technique that detects and extracts phrases containing user opinions from their posts. Then, at the second part of analysis annotate the sentiment orientation of the identified opinion phrases in order to assign them with a positive or negative polarity label depending on their publishers underlying stance against the issues they discuss. Based on the output of the above analysis, may not only capture the citizen's viewpoints on governmental issues but with the proper tools and techniques may also be able to build predictive models about how citizens value public sector regulations. In the following paragraphs, describe in detail how process the user postings to identify and evaluate opinion phrases as well as how to utilize the mined opinions along with their polarity labels in order to train an opinion classifier.

### 2.1 e-Government

Definition by the EU is "Electronic Government (EGovernment) is the use of information and communication technologies in public administrations combined with organizational change and new skills to improve public services and democratic processes and to strengthen support to public policies. Regarding Thailand e-Government is the use of information technology to

support government operations, engage citizens, and provide improved quality of government services. In addition to able to transparency

## 2.2 Opinion Mining

Table 1 Types of Opinion Mining

Types of Opinion	Purpose
Anomaly Detection	Detects an unusual emotion that can change the crowd trend.
Concept Extraction	Hot topic that can become talk of the town.
Emotional Signature Detection	Emotions around important topics can be tracking for estimate future behavior.
Fine-Grained Citizen Satisfaction	Polarity are positive, negative and neutral, but the hidden point is how to measure polarity.
Opinion Classification	For complex topics, opinion may rapidly diverge significantly between positive and negative.
Polarity Classification	It's easy to separate into positive, negative and neutral. The point is how to define the hidden agenda in that opinion.
Subtle/Hidden Expression	Many of positive opinion are untruth. Extraction for the real meaning is the key.
Tribal Ethos Identification	For social animals, better look overview better then point to spot.

### Opinion analyzer architecture

One can consider document-level polarity classification to be just a case of text categorization with sentiment- rather than topic based categories. Hence, standard machine learning classification techniques, such as support vector machines (SVMs), can be applied to the entire documents themselves, as was done by Pang, Lee, and Vaithyanathan (2002). Refer to such classification techniques as default polarity classifiers. However, as noted above, we may be able to improve polarity classification by removing objective sentences. Therefore propose, as depicted in Figure 1, to first employ a subjectivity detector that determines whether each sentence is subjective or not: discarding the objective ones creates an extract that should better represent a review's subjective content to a default polarity classifier.

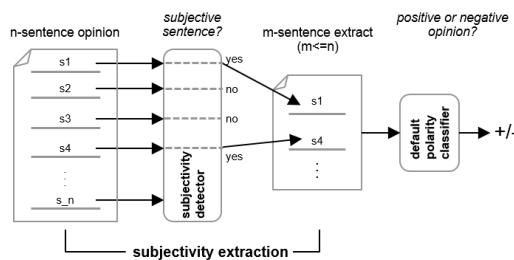


Fig.2 A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.

## 2.3 Decision Support System

Decision Support Systems (DSS) are a specific class of computerized information system that supports business and organizational decision-making activities. A properly designed Decision Support System is an interactive software-based system intended to help decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions.

### 3. Evolving process

#### 3.1 Opinion mining approach.

From many user generate input intruding social media, bulletin board or public block there are millions of text that can defend into many type of linguistic. For this paper only mention at opinion sentences as show in Fig.3. The opinion mining can classification into Fact Sentiment and Question. The easiest way to measure value of the opinion is focus on the sentiment as shown in Ding T. and Pan S. (2016). The sentiment can be divide into three type as Simple Sentence complex sentient and paragraph. When deep attention into the simple sentient there are polarity and some condition.

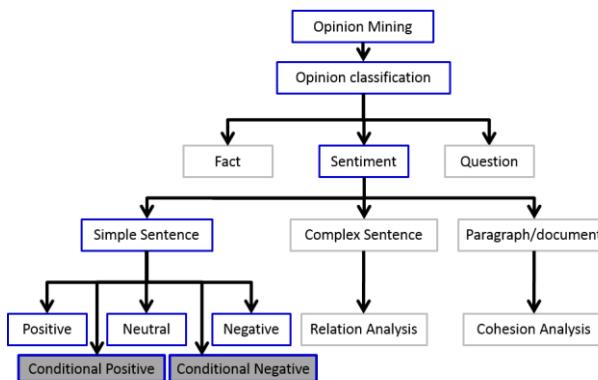


Fig.3 *Opinion mining classification model Conceptual*

#### 3.2 Technical Approaches

##### 3.2.1 Linguistic NLP

Linguistic NLP uses grammatical and lexical processing rules to identify and extract entities and their features, associate them with specific topics, isolate subjective statements of sentiment, and assign a positive-negative polarity rating. Sentiment analysis is typically an additional processing module of an NLP-based text analytics system. Linguistic NLP requires an industry or enterprise domain expert, and some-times a linguistics expert, to build the language models and rules required to interpret the language snippets that are being analyzed for sentiment. Sentence detection understands the full use of punctuation and capitals. Most text analytics products have functions to clean the text. Tokenization breaks a sentence down into its smallest units, which are usually words, but will also separate out contractions. Tokenization removes punctuation as preparation for further processing. And last Part of speech (POS) tagging: POS tagging assigns a POS tag to each token in a sentence. For a list of commonly used POS tags from the Penn Treebank project. As Peleja F. and Magalhães J. (2015).

##### 3.2.2 MACHINE LEARNING STATISTICAL ANALYSIS

Machine learning statistical analysis (MLSA) reduces reliance on linguistic-based analysis of documents. In-stead, MLSA relies on statistical and probabilistic analyses of words in documents to determine topic, sentiment and polarity. Most of the MLSA approaches are supervised, meaning that the systems are trained on document sets that already have sentiment and polarity assigned to them.

*Latent semantic indexing (LSI)*

LSI is a document-term matrix processing approach that is based on the assumption that words that are used in the same contexts tend to have similar meanings. Technically, LSI uses a mathematical technique called singular value decomposition (SVD) to identify patterns that exist between the terms and concepts contained in a text corpus. LSI is an application of principal component analysis (PCA) or independent component analysis (ICA) applied to text:

**PCA :** The goal of PCA is to identify the most meaningful basis to re-express a confusing dataset. It arranges data in a matrix form and, similar to SVD, compresses the data while maintaining the relationships between the documents and concepts.

**ICA :** ICA is a mathematical method for separating data into underlying informational components. SVD can also be used to process the matrices yielded by ICA. ICA is primarily used on voice and video data applications.

There are other machine learning data-mining techniques that use a linear algebra approach for determining feature selection, including Weka and Clementine.

#### *Sentiment Classification*

MLSA classification of sentiment analysis is a statistical and probabilistic modeling activity. The purpose of the statistical model is to classify sentiment based on a previously learned set of data. The feature creation and feature extraction steps create the data to be run through the classification model. The next challenge of MLSA is to select a modeling approach that is best suited to handle the features that have been created.

**Naïve Bayes (NB) :** An NB classifier assumes that each feature is independent from every other feature. Based on the training data, the NB classifier assigns a probability to each feature that is describing one sentiment or another and then sums the probabilities for a final prediction.

**Maximum entropy (ME) :** The concept behind the ME modeling approach is also known as the principle of Occam's Razor, which states that the right solution is usually the one that is least complex. In modeling terms, this means that MLSA should model all that is known and assume nothing about the unknown. The model should be consistent with all the facts but otherwise as uniform as possible.

**Support vector machines (SVM) :** The basic idea behind SVM is to find a hyperplane that not only separates the document vectors in one class from those in the other, but for which the separation is as large as possible.

MLSA classification engines generally work as black-box systems and provide little feedback as to why a document or sentence is classified the way it is, so the systems are not easy to tune without additional training. Angioni M. and Tuveri F. (2012).

#### 3.2.3 HYBRID SENTIMENT ANALYSIS

The decision regarding linguistic NLP versus MLSA does not need to be an either scenario. There are four main components. Rules, models and lexicons is in practice made up of multiple modules. It contains government-specific lexicons, dictionaries and word lists that are used to add context and meaning to the text being brought in for analysis. There are rules and text processing models to guide the linguistic NLP analysis and the statistical analysis processing. Linguistic NLP analysis is discussed in detail in the Linguistic NLP section of this assessment. MLSA is discussed in detail in the Machine Learning Statistical Analysis section of this assessment.

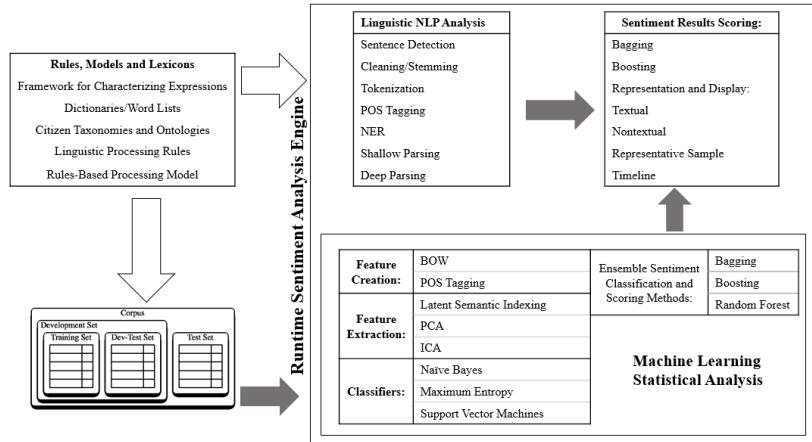


Fig.3 Architecture framework of decision support system based hybrid opinion analyzer engine architecture for e-Government

#### 4. Discussion

Evaluating classification accuracy of polarity-labeled user opinions So far have

described method for automatically identifying citizen opinions about governmental decisions in an attempt to assist both the public and governments successfully interact with each other as well as a metric for quantifying the impact of the mined opinions in formulating the public's stance against E-Government issues. The core of method is an opinion mining framework, which manages to automatically identify and validate citizen opinions. Have also proposed the architecture of an interactive e-Government framework that collects the mined opinions, processes them and feeds them back to governmental bodies so that they account for them in their decision making duties. In this section will focus on the novel method for capturing and assessing citizen opinions on governmental issues trying to experimentally evaluate the performance of technique in automatically organizing mined opinions in terms of their polarity, i.e., to group positive and negative user opinions on governmental issues separately so that governments can exploit and account for them in their subsequent regulations.

##### *Experimental setup to assess study objective,*

Collected a set of real citizen opinions, export 10,000 twits from twitter. Then trained three different classifiers incorporated into the RapidMiner 6. The three classifiers use are Support Vector Machine, Maximum entropy (ME) and Naïve Bayes. Before the training phase, software prepares the data by randomizing the full dataset and then stratify it because the classification class is nominal. Then, in order to reduce variability, performs a 10 fold cross validation and generates training and test sets using different partitions and the validation results are averaged over the rounds. Classification training is vital in order to learn the classifiers discriminate between positive and genitive citizen opinions.

##### *Experimental results after the training,*

Each classifier returns a summary of the results. The following chart shows the average of cases that were correctly and incorrectly predicted for the three classifiers for each dataset. In particular, the figure depicts the fraction of opinion phrases that were correctly and incorrectly identified by the classifiers as positive or negative. As the figure shows, the algorithm with the best classification performance is the Support Vector Machine where the average accuracy is about 83% while the worst performing classification algorithm is Naïve Bayes with 78.92% accuracy. Note that Support Vector Machine is more

suitable for text attributes in contrast to Naïve Bayes which has better performance for numerical attributes. Table 2 summarizes the performance details of the three classification modules employed in study and as results suggest the proposed method is quite effective into automatically organizing opinion phrases in terms of their polarity.

Table 2: Evaluation of opinion classification accuracy

	Naïve Bayes	Support Vector Machine	Maximum entropy
True Positive Rate	12.21%	67.74%	68.58%
True Negative Rate	64.88%	83%	85.6%
False Positive Rate	0.54%	3.80%	4.56%
False Negative Rate	78.92%	32.25%	32.41%

Based on the experimental findings, may deduce the following. First, that proposed opinion mining and evaluation technique is quite effective in automatically identifying the public's stance towards governmental decisions. Moreover, results demonstrate that method can be easily integrated into existing classification modules in order for the latter to automatically organize mined user opinions according to the positive or negative orientation. Above all, experimental study shows that with today's technological advancements it is feasible to deploy existing mechanisms into novel applications such ones related to e-Government.

## 5. Conclusion

As already pointed out by other researchers, one of the most important issues for making e-Government effective is to enable citizens participate in the decision-making process. Via the proposed approach ensure that citizen opinions and comments are properly received by public bodies and that they are accounted for in subsequent governmental actions as well as provide both citizens and governments with the means to effectively interact with each other and actively participate into common actions from which both would benefit. Although the work presented in this chapter is still in early stages and only gives a general notion with respect to how opinion mining techniques can be successfully explored in the course of e-Government and e-inclusion approaches believe that it will pave the ground for more initiatives in this respect. Another aspect of future work would be to rely on the mined and polarity annotated user opinions in order to build and train effective prediction models that would be able to approximate the potential impact of planned governmental decisions on citizen's stance. Finally, it would be interesting to apply opinion mining technique towards a wide variety of user opinions on governmental decisions and identify the regulations that interest citizens the most and thus offer them the infrastructure to interact with governmental figures.

## Acknowledgements

Authors are grateful to Electronic Government Agency (Public Organization) of Thailand for financial supports. Grateful acknowledgement is also made for Thai-Nichi Institute of Technology for conducting research, seminar, and discussions.

**REFERENCES**

1. Angioni M. and Tuveri F. (2011). A SEMANTIC APPROACH TO THE EXTRACTION OF FEATURE TERMS.In Proceedings of the 6th International Conference on Software and Database TechnologiesISBN 978-989-8425-77-5, pages 402-407.
2. Angioni M. and Tuveri F. (2012). AN AUTOMATIC APPROACH TO FEATURE EXTRACTION.In Proceedings of the 4th International Conference on Agents and Artificial IntelligenceISBN 978-989-8425-95-9, pages 473-476
3. Cercel D. and Trausan-Matu S. (2015). Modeling Post-level Sentiment Evolution in Online Forum Threads.In Proceedings of the International Conference on Agents and Artificial IntelligenceISBN 978-989-758-074-1, pages 588-593.
4. Ding T. and Pan S. (2016). An Empirical Study of the Effectiveness of using Sentiment Analysis Tools for Opinion Mining.In Proceedings of the 12th International Conference on Web Information Systems and TechnologiesISBN 978-989-758-186-1, pages 53-62
5. Fernandes Caíña M., Díaz Redondo R. and Fernández Vilas A. (2014). Marble Initiative - Monitoring the Impact of Events on Customers Opinion.In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (IC3K 2014)ISBN 978-989-758-048-2, pages 403-410.
6. Firmino A., Baptista C., Alves A., Andrade D., Figueirêdo H., Filho G. and de Paiva A. (2016). Towards Metadata Analysis on Opinionated Content in Tweets.In Proceedings of the 18th International Conference on Enterprise Information SystemsISBN 978-989-758-187-8, pages 314-320. DOI: 10.5220/0005890803140320
7. Pang, Lee, and Vaithyanathan (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of EMNLP, pp. 79--86,
8. Peleja F. and Magalhães J. (2015). Learning Text Patterns to Detect Opinion Targets.In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge ManagementISBN 978-989-758-158-8, pages 337-343.
9. Sangiorgi P., Augello A. and Pilato G. (2014). An Approach to Detect Polarity Variation Rules for Sentiment Analysis.In Proceedings of the 10th International Conference on Web Information Systems and TechnologiesISBN 978-989-758-024-6, pages 344-349.
10. Tuveri F. and Angioni M. (2012). Definition of a Linguistic Resource for Opinion Mining.In Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2012)ISBN 978-989-8565-16-7, pages 64-73.

## Author Index

### A

Asawasutsakorn, Pitinat, 33

Amin, Ashraful M., 46

### B

Buatoom, Uraiwan, 52

Buranarach, Marut, 157

Boonarchatong, Chutiwan, 298

### C

Chai, Wasan NA, 157

Chaowalit, Orawan, 120

Cu, Jocelynn, 253

### D

Dobashi, Konomu, 133

### F

Fushimi, Takayasu, 182

Fukui, Ken-ichi, 218

Fukui, Ken-ichi, 243

### H

Hnoohom, Narit, 64, 78

Hossan, Zakir Md., 46

Hagad, Juan Lorenzo, 243

### I

Ikeda, Tetsuo, 182

### J

Jinawath, Natini, 33

Janphat, Jittima, 120

### K

Kamolsantiroj, Suwatchai, 21

Kongprawechnon, Waree, 52

Kojiri, Tomoko, 145

Ketui, Nongnuch, 101

Khongchai, Pornthep, 139

Kazama, Kazuhiro, 182

Klangsakulpoontawee, Nipaporn, 304

Ketcham, Mahasak, 278, 286, 298, 327, 346

Kosolsombat, Somkiat, 337

### L

Lertritchai, Penpichaya, 304

Limprasert, Wasit, 337

Luekhong, Prasert , 101

Lutfi Lebai Syaheerah, 261

Luz, Beatrice Ma., 253

### M

Meeboon, Benjamard, 33

Mokhtar Sadiq Najlaa, 261

Muankid, Anchana, 346

### N

Nateeraitaiwa, Suppanut, 64

Noymantee, Jeerana, 360

Ngnotchouye, Jean Medard, 166

Nanba, Hidetsugu, 199

Nozawa, Yuya, 206

Numao, Masayuki, 218

Nocum, McAnjelo, 253

Numao, Masayuki, 243

**O**

Ong, Ethel, 108

**P**

Pipanmaekaporn, Luepol, 21

Proma, Tanjina Piash, 46

Purganan, Timothy Jasper, 253

Pramkeaw, Patiyuth, 304

Phetnuam, Siriphan, 352

Phattarachairawee, Siriya, 286

**R**

Ruangrajitpakorn, Taneth, 157

Rodrigo, Mercedes Ma., 231

Rattanasiriwongwut, Montean, 286

**S**

Suttichaya, Vasin, 14

Sivaraksa, Mingmanas, 33

San-Um, Wimol, 360

Supnithi, Thepchai, 157

Songmuang, Pokpong, 139

Saito, Kazumi, 182

Smanchat, Sucha, 298

**T**

Taewijit, Siriwon, 1

Theeramunkong, Thanaruk, 1, 52, 360

Tanthuwapathom, Ratikanlaya, 78

Tilahun, Surafel, 166

Thammasan, Nattapong, 218

Thammachantuek, Ittikon, 278

**V**

Vea, Larry, 231

**W**

Wong, Wing San, 253

Wang, Bin, 315

Wisitpongphan, Nawaporn, 298

**Y**

Yang, Gaijing, 315

Yimyam, Worawut, 327

Yingthawornsuk, Thaweesak, 352

**Z**

Zhang, Qiang, 315

Zhou, Changjun, 315

Zbrzezny, Agnieszka, 272