The 1st Joint Seminar on Computational Intelligence (JSCI 2016)

IEEE Computational Intelligence Society Thailand Chapter

28 April 2016

Sponsors

JSCI2016 is supported by the following sponsors:

- Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang

- IEEE Computational Intelligence Society Thailand Chapter

- International Neural Network Society Thailand Chapter

Contents

x

Program

Time	Topics
08:30–09:00	Registration
09:00–09:05	Welcoming Speech Chanboon Sathitwiriyawong (IEEE CIS Chair)
09:05–09:10	Welcoming Speech Nopporn Chotikakamthorn (IT-KMITL Dean)
09:10–10:10	Keynote Speaker – Three NLP Key Challenges in Big Data Virach Sornlertlamvanich Sirindhorn International Institute of Technology, Thammasat University, Thailand
10:10–10:30	Keynote Speaker – Supporting Web Searches by Intelligent Local Agents Herwig Unger Fern Universität in Hagen, Germany
10:30–10:50	Coffee Break
10:50–11:10	Interval Type 2 Fuzzy Logic System: Construction and Applications Phayung Meesad, KMUTNB
11:10–11:30	Learning to Predict Where People Look with Tensor-based Multiview Learning Kitsuchart Pasupa, KMITL
11:30–11:50	TNorm: An Unsupervised Batch Effects Correction Method for Gene Expression Data Classification Praisan Padungweang, KMUTT
11:50–12:10	Image Representations using Local Image Descriptors Sanparith Marukatat, NECTEC
12-10–12:15	Photo Session
12:15–13:15	Lunch
13:15–13:35	Map Reduced Technique for Big Data Mining Van-Hieu Duong, KMUTNB
13:35–13:55	Hinge Loss Projection for Classification Syukron Alfarozi, KMITL
13:55–14:15	Semi-Automatic Construction of Thyroid Cancer Intervention Corpus from Biomedical Abstracts Wutthipong Kongburan, KMUTT
14:15–14:35	ARMFEG: Association Rule Mining by Frequency Edge Graph Pramool Suksakawpong, KMUTNB
14:35–14:45	Coffee Break
14:45–15:05	Text-Background Decomposition Using Adaptive Boundary Clustering Ungsumalee Suttapakti, KMITL
15:05–15:25	Utilization of Parallel Computing and Affinity Propagation Clustering in Identifying Sub-network Biomarker Genes of Cancer on Web Application Thanyathorn Thanapattheerakul, KMUTT
15:25–15:45	Applying Expert for Recommending Items Thanaphon Phukseng, KMUTNB
15:45–16:45	IEEE Computational Intelligence Society Thailand Chapter Meeting

Invited Talks

Three NLP Key Challenges in Big Data

Virach Sornlertlamvanich

School of Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology, Thammasat University
virach@siit.tu.ac.th

The talk formalizes the three key issues in natural language processing (NLP), namely word segmentation, named entity recognition or keyword extraction, and semantic relation extraction, which are crucial in handling the explosively growth of text in the Big Data era. In the flood of information today, we spend most of the time to grasp the essence of the information rather than to enjoy the reading. Many approaches have been proposed to handle these fundamental issues, however, there is still much room for improvement. The talk elaborates the problematic issues in text processing and aim to make the problems well recognized. In a non-segmenting language such as Thai, Lao, Cambodian, Chinese, or Japanese, word segmentation to determine the word boundary in a sentence is an introductory process of an input text. Alternatively, in character based approaches, mutual information and entropy are effective measures to uncover the possible word boundary for the non-segmenting languages such as the Thai language. Keyword labeling is also a task that we can effectively apply a machine learning approach such as MIRA (Margin Infused Relaxed Algorithm) to capture the word surrounding context. This can be done on the result from the word segmentation task. Undoubtedly, the accuracy of the annotated tag is ranked descendently from person (PER), date (DAT), location (LOC), and organization (ORG). This is because tag for person has the least ambiguity. The pattern for extracting the semantic relation between the type-annotated keywords is accordingly assigned to the word form of the disambiguated verb phrase. The experimental result shows that most of the distance between the keyword and the target verb phrase is not more than one word. Therefore, we can find the target verb phrase in the adjacent position or one word skipped position with the highest probability.

The talk demonstrates the three constructive applications on the huge generated data, i.e. linked data formation for knowledge map reasoning; keyword tracking on social media to understand the online social movement; and online local news generation.

The task of natural language processing today is not just only for the language study, but it can bring along the possibilities on the advantages of the Internet, big data, and machine learning technique.

Keyword: NLP; word segmentation; named entity recognition; semantic relation extraction; knowledge map; social media text understanding.

Biopic: Dr. Virach Sornlertlamvanich received his bachelor and master degrees in engineering from Kyoto University in 1984 and 1986 respectively, and doctoral degree in Computer Science from Tokyo Institute of Technology in 1998. Since 2015, he is a lecturer at Sirindhorn International Institute of Technology, Thammasat University. He was esteemed "The Researcher of the Year 2001" by the Nation Newspaper in 2001, and was conclusively awarded by the National Research Council of Thailand as The Most Outstanding Researcher of the Year in Information Technology and Communication in 2003.

Supporting Web Searches by Intelligent Local Agents

Herwig Unger

Faculty of Mathematics and Computer Science
Fern University in Hagen, Germany
herwig.unger@gmail.com

Manually formulating queries to search in the World Wide Web is a tedious, time-consuming and error-prone task, which cannot sufficiently be supported by any existing search engine. Various user characteristics, preferences and needs, the consideration of search histories as well as disambiguation problems are just a few reasons.

A solution of this problem is to introduce intelligent local search agents, which are able to evaluate local, user-specific files, emails and (former) user activities. It will be shown, that as result not only the initial search request may satisfactorily be completed employing additional, well-fitting keywords, but that also the results may be represented in a more ergonomic manner than before.

In this talk, an overview on existing methods for such searches is given, and the tool "docanalyser" is introduced as an example of a software realisation.

Keyword: Self-organization; Adaptive learning systems; Internet algorithms; Simulation systems.

Biopic: Prof. Dr.-Ing. habil. Herwig Unger got his PhD with a work on Petri Net transformation in 1994 from the Technical University of Ilmenau and his habilitation with a work on large distributed systems from the University of Rostock in 2000. Since 2006, he is a full professor at the FernUniversität in Hagen and the head of the Chair of Communication Networks. His research interests are in self-organization, adaptive and learning systems, Internet algorithms and simulation systems.

Faculty Talks

Interval Type 2 Fuzzy Logic System: Construction and Applications

Phayung Meesad

Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand
phayung.m@it.kmutnb.ac.th

Constructing interval type 2 fuzzy systems is one of a challenge in machine learning field. In this talk, an approach for generation of interval type-2 fuzzy inference systems will be presented. The method comprises of two phases: 1) Structure initialization and 2) parameters fine tuning. In the first phase, a one-pass clustering method is carried out to find both a suitable number of rules and a suitable number of fuzzy sets of each variable in which inputs and targets are used as training data. In the second phase, an evolutionary algorithm can be applied to fine-tune the membership function parameters to increase the performance of a developing fuzzy system. The evaluation of the proposed method will be presented based on pattern classification and time series prediction.

Keyword: Fuzzy logic; Interval type 2 fuzzy logic system; Clustering; Learning; Pattern classification; Time series prediction.

Learning to Predict Where People Look with Tensor-Based Multi-view Learning

Kitsuchart Pasupa[1,✉], and Sandor Szedmak[2]

[1]Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Thailand
[2] Institute of Computer Science
University of Innsbruck, Austria
✉kitsuchart@it.kmitl.ac.th

Eye movements data collection is very expensive and laborious. Moreover, there are usually missing values. Assuming that we are collecting eye movements data on a set of images from different users (views). There is a possibility that we are not able to collect eye movements of all users on all images. One or more views are not represented in the image. We assume that the relationships among the views can be learnt from the complete items. The task is then to reproduce the missing part of the incomplete items from the relationships derived from the complete items and the known part of these items. Using the properties of tensor algebra we show that this problem can be formulated consistently as a regression type learning task. Furthermore, there is a maximum margin based optimisation framework where this problem can be solved in a tractable way. This problem is similar to learning to predict where human look. The proposed algorithm is proved to be more effective than well-known saliency detection techniques.

Keyword: Multi-view learning; Missing data; Tensor algebra; One-rank tensor approximation; Maximum margin learning; Eye movements.

TNorm: An Unsupervised Batch Effects Correction Method for Gene Expression Data Classification

Praisan Padungweang$^{\boxtimes}$, Worrawat Engchuan, and Jonathan H. Chan

School of Information Technology

King Mongkut's University of Technology Thonburi, Thailand

$^{\boxtimes}$praisan.pad@sit.kmutt.ac.th

In the field of biomedical research, gene expression analysis helps to identify the disease-related genes as genetic markers for diagnosis. As there is a huge number of publicly available gene expression datasets, the ongoing challenge is to utilize those available data effectively. Merging microarray datasets from different batches to improve the statistical power of a study is one of the active research topics. However, various works have addressed the issue of batch effects variation, which describes variation in gene expression levels induced by different experimental environments. Ignoring this variation may result in erroneous findings in a study. This work proposes a method for batch effect correction by mapping underlying topology of different batches. The mapping process for cross-batch normalization is examined using basic linear transformation. The comparative study of three cancers is conducted to compare the proposed method with a proven batch effects correction method. The results show that our method outperforms the existing method in most cases.

Keyword: Gene expression; Batch effects; Clustering; Linear regression; Classification; Cancer; Diagnosis.

Image Representations using Local Image Descriptors

Sanparith Marukatat

National Electronics and Computer Technology Center

sanparith.marukatat@nectec.or.th

Nowadays, image classification techniques are dominated by image descriptors provided by trained convolutional neural network (CNN). CNN combines feature extractor and classifier that are trained together for a specific task. Hence, it can achieve very high classification result in detriment of high computational cost. In this talk, we consider a lighter and more flexible approach based on local image descriptors. Indeed, recent works have shown that we can represent an image using these local descriptors and achieve similar performance, or better in some cases, as CNN. We will review some classical image descriptors (e.g. SIFT and LBP) as well as the representations such as bag of features (BoF), the spatial pyramid matching (SPM) and the locality-constrained linear coding (LLC). We compare these representations on a real world dataset of food and restaurant images.

Student Talks

A Combination of Cell-Based and Graph-based Techniques for Clustering Big Datasets on Computers with Limited Memory

Van-Hieu Duong$^{\boxtimes}$, and Phayung Meesad

Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand
$^{\boxtimes}$dvhieu@gmail.com

Big datasets not only bring potential benefits to enterprises, business and scientific applications but also result in challenges for data analyzing and mining experts. Data clustering techniques are essential tools for analyzing and mining big datasets and have been applied to many areas of life, business, engineering, science, and education. Beside data clustering, outlier analysis also plays an important role in data science because its applicability to various applications such as credit card fraud detection, insurance, health care, security intrusion detection, interesting sensor events detection, fault detection in safety systems, military surveillance, medical diagnosis, law enforcement, and earth science. Existing partitioning-based clustering methods have some limitations in terms of overwhelming memory and intolerable executing time when working with big datasets on computers with limited memory. Moreover, the well-known outlier detection algorithms cannot be applied on datasets having large numbers of objects due to their very long executing time. This study proposes comprehensive solutions to detect outliers in big datasets and to cluster big datasets on a personal computer with limited memory. The proposed solutions not only can solve the outlier detection problem with very large datasets but also the data clustering problems of big dataset clustering on computers with limited memory. Firstly, a fast outlier detection algorithm is proposed. This is a combination of cell-based algorithms and the rank-based outlier detection method with various depths to identify and eliminate outliers in provided datasets (abbreviated to Cell-DROS). This proposed algorithm is designed in such a way that it not only utilizes the strengths of the RDOS algorithm in term of accuracy level, but also minimizes the limitation of the RDOS algorithm in term of executing time. Secondly, Cell-MST based clustering algorithms are proposed. They are combinations of cell-based algorithms and minimum spanning trees. These proposed clustering algorithms are better than the similarity-based and theoretic-through set methods in term of memory use, and better than the QEM method in terms of accuracy level and executing time.

Keyword: Cell-based clustering; Big dataset clustering; MST-based clustering; Outlier detection.

Hinge Loss Projection for Classification

28 Apr
13:35

Syukron Abu Ishaq Alfarozi✉, Kuntpong Woraratpanya, and Kitsuchart Pasupa

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Thailand
✉syukron@outlook.com

Mean squared error (MSE) loss function is well-suited for regression task. Moreover, it can also be applied to classification problem. On the other hand, hinge loss function is one-sided loss function proposed only for classification problem, it allowed data point has value greater than +1 for the positive class and less than -1 for negative class, otherwise it will calculate as error or loss. However, it is not differentiable, to minimize the hinge function uses sub-gradient method or uses quadratic programming (QP) in case of support vector machine (SVM). In this paper, we used simple projection method to minimize hinge loss function for classification case. We modified the least squared method so that it can adopt using hinge loss instead of MSE and we also applied it to ridge regression and least Square-SVM (LS-SVM) for classification problem. The results show the effectiveness of our method in terms of accuracy to the ordinary least squared method and also more robust to the outliers and the distribution of data.

Keyword: Hinge loss; Least squares; Projection; Ridge regression.

Semi-Automatic Construction of Thyroid Cancer Intervention Corpus from Biomedical Abstracts

28 Apr
13:55

Wutthipong Kongburan✉, Praisan Padungweang, Worarat Krathu, and Jonathan H. Chan

School of Information Technology
King Mongkut's University of Technology Thonburi, Thailand
✉58130800102@st.sit.kmutt.ac.th

Thyroid cancer is a common endocrine tumor that is experiencing a steady increase in incidence worldwide. The latest discoveries on disease and its treatment are mostly propagated in the form of biomedical publications such as those in PubMed. Unfortunately, this information is distributed in unstructured text with over two thousand articles being added annually. Text mining technology plays an important role in information extraction, since it can be used to uncover hidden value from the vast amount of text in reasonable time. In general, a preliminary task of text mining is Named Entity Recognition (NER). In this case, a gold standard corpus is needed, since the capability of NER depends on a trustworthy corpus. However the construction of gold standard corpus is a laborious and time-consuming process. In order to obtain a reasonably practical corpus in a limited time, this paper consequently proposes a semiautomatic approach to construct a thyroid cancer interventions corpus. The experimental results demonstrate that the proposed method can be used to construct a thyroid cancer intervention corpus reasonably in terms of both performance and overfitting avoidance.

ARMFEG: Association Rule Mining by Frequency-Edge-Graph

Pramool Suksakaophong✉, and Phayung Meesad

Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand
✉pramool.s@email.kmutnb.ac.th

Association rules mining from transaction data can be used to recommend the items that are often purchased together frequently. However, it is difficult to set minimum support threshold. If the minimum support threshold is set too high, then there may be only a small or even no result. If the threshold is set too low, it may generate many uninteresting associations. In addition, each supporting a different set of data, enabling users to find the optimal difficult. This paper presents a new approach to the collection frequency by using top weight of complete symmetric digraphs. Using the top weight, the association rule with the maximum support can be calculated and it works with any dataset.

Text-Background Decomposition Using Adaptive Boundary Clustering

Ungsumalee Suttapakti✉, Kuntpong Woraratpanya, and Kitsuchart Pasupa

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Thailand
✉un.ung@hotmail.com

Thai text localization and recognition in natural scenes is still a grand challenge in current applications. However, the efficiency of recognition rates depends on text localization, i.e., the higher purity of text-background decomposition leads to the higher accuracy rate of character recognition. In order to achieve this purpose, the text-background decomposition methods, namely adaptive boundary clustering (ABC) and n-point boundary clustering (n-PBC), are proposed to improve a precision of text localization. These methods are evaluated by self-entropy for purity measure. Based on 300 test images, the experimental results demonstrate that the ABC method achieves the very low self-entropy, i.e., the low self-entropy implies the good decomposition of text and background. Furthermore, based on 8,077 characters in natural scene test images, the ABC method helps increase the precision of text localization and improves the accuracy rate of character recognition, when compared to the conventional methods.

Utilization of Parallel Computing and Affinity Propagation Clustering in Identifying Sub-network Biomarker Genes of Cancer on Web Application

Jaturong Kongmanee, Thanyathorn Thanapattheerakul✉, and Jonathan H. Chan

School of Information Technology

King Mongkut's University of Technology Thonburi, Thailand

✉th.thanyathorn@gmail.com

This project focuses on improvement of the identification of genes related to complex disease development to unravel the underlying mechanisms of how irregular phenotypes are introduced. The complex disease such as cancers is driven by the accumulation of small contributions of multiple genes. Thus, gene co-expression network is applied in this project to identify a group of genes within the same module or sub-network, which tend to have similar expression pattern and correlate with each other. Therefore, identifying biomarkers using the co-expression network can lead to extract the functional group of genes relevant to the diseases of interest. This work applies Affinity Propagation Clustering (APC) algorithm in order to accelerate the gene modules identification from co-expression network and applies the benefits of parallel computing to improve the system performance.

Keyword: Affinity Propagation Clustering algorithms (APC), Biomarker, Gene co-expression network, Sub-network, Visualization, Parallel computing.

Applying Expert for Recommending Items

Thanaphon Phukseng✉, and Sunantha Sodsee

Faculty of Information Technology

King Mongkut's University of Technology North Bangkok, Thailand

✉s5607011910048@email.kmutnb.ac.th

This research is presented the recommendation model applying expert concepts to recommend items. The merits of this model are to allocate the best fit items to serve user needs. Three aspects were considered in the analysis; network connection focusing on indegree of network, frequency of rating by counted rating from each user and user likes by calculating average user rating gaining from other users in system. For evaluation, the Epinion dataset was employed as a benchmark data. Results of those experiments were put into evaluation. The results have proven that frequency of rating methods illustrates the greater percentage of recommendation accuracy and the lower error than other methods. The maximum accuracy was around 85.404% and the minimum mean absolute error was around 0.146.

Keyword: Recommender System; Expert System.

Author Index