IMMM 2014

The Fourth International Conference on Advances in Information Mining and Management

ISBN: 978-1-61208-364-3

July 20 - 24, 2014

Paris, France

IMMM 2014 Editors

Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
Abdulrahman Yarali, Murray State University, USA

IMMM 2014

Foreword

The Fourth International Conference on Advances in Information Mining and Management (IMMM 2014), held between July 20-24, 2014, in Paris, France, continued a series of academic and industrial events focusing on advances in all aspects related to information mining, management, and use.

The amount of information and its complexity makes it difficult for our society to take advantage of the distributed knowledge value. Knowledge, text, speech, picture, data, opinion, and other forms of information representation, as well as the large spectrum of different potential sources (sensors, bio, geographic, health, etc.) led to the development of special mining techniques, mechanisms support, applications and enabling tools. However, the variety of information semantics, the dynamic of information update and the rapid change in user needs are challenging aspects when gathering and analyzing information.

We take here the opportunity to warmly thank all the members of the IMMM 2014 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to IMMM 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the IMMM 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that IMMM 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of information mining and management.

We are convinced that the participants found the event useful and communications very open. We hope that Paris, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

IMMM 2014 Chairs:

IMMM Advisory Committee

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
David Newell, Bournemouth University - Bournemouth, UK
Kuan-Ching Li, Providence University, Taiwan
Abdulrahman Yarali, Murray State University, USA
Alain Casali, Aix Marseille Université, France
Ingrid Fischer, Universität Konstanz, Germany
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Paolo Garza, Politecnico di Torino, Italy
Bartlomiej Jefmanski, Wroclaw University of Economics, Poland
Nathalie Pernelle, Université Paris-Sud, France
Jürgen Pfeffer, Carnegie Mellon University, USA

Jörg Scheidt, University of Applied Sciences Hof, Germany
Ariella Richardson, Jerusalem College of Technology, Israel
Lorna Uden, Staffordshire University, UK
Eli Upfal, Brown University - Providence USA
Domenico Ursino, Università Mediterranea di Reggio Calabria, Italy
Jan Zizka, Mendel University - Brno, Czech Republic
Nima Hatami, University of California - San Diego, USA

IMMM Industry/Research Liaison Committee
Johannes Meinecke, SAP AG / SAP Research Center Dresden, Germany
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Feng Yan, Facebook Inc., USA
Katja Pfeifer, SAP AG, Germany
Arno H.P. Reuser, Reuser's Information Services, The Netherlands
Yulan He, Knowledge Media Institute / The Open University, UK
Artura Mazeika, Max Planck Institute for Informatics - Saarbrücken, Germany
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Wei Jin, Amazon.com, Seattle, USA
Olivier Caelen, Atos Worldline, Belgium
Yili Chen, Monsanto Company, USA
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Daniel Kimming, Karlsruhe Institute of Technology, Germany
Josiane Mothe, IRIT, France
Dirk Labudde, Hochschule Mittweida, Germany
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Robert Wrembel, Poznan University of Technology, Poland

IMMM Publicity Chairs
Alessia Saggese, University of Salerno, Italy
Ludovico Boratto, Università di Cagliari, Italy
Toshio Kodama, University of Tokyo, Japan

IMMM 2014

COMMITTEE

IMMM Advisory Committee

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
David Newell, Bournemouth University - Bournemouth, UK
Kuan-Ching Li, Providence University, Taiwan
Abdulrahman Yarali, Murray State University, USA
Alain Casali, Aix Marseille Université, France
Ingrid Fischer, Universität Konstanz, Germany
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Paolo Garza, Politecnico di Torino, Italy
Bartlomiej Jefmanski, Wroclaw University of Economics, Poland
Nathalie Pernelle, Université Paris-Sud, France
Jürgen Pfeffer, Carnegie Mellon University, USA
Jörg Scheidt, University of Applied Sciences Hof, Germany
Ariella Richardson, Jerusalem College of Technology, Israel
Lorna Uden, Staffordshire University, UK
Eli Upfal, Brown University - Providence USA
Domenico Ursino, Università Mediterranea di Reggio Calabria, Italy
Jan Zizka, Mendel University - Brno, Czech Republic
Nima Hatami, University of California - San Diego, USA

IMMM Industry/Research Liaison Committee

Johannes Meinecke, SAP AG / SAP Research Center Dresden, Germany
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Feng Yan, Facebook Inc., USA
Katja Pfeifer, SAP AG, Germany
Arno H.P. Reuser, Reuser's Information Services, The Netherlands
Yulan He, Knowledge Media Institute / The Open University, UK
Artura Mazeika, Max Planck Institute for Informatics - Saarbrücken, Germany
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Wei Jin, Amazon.com, Seattle, USA
Olivier Caelen, Atos Worldline, Belgium
Yili Chen, Monsanto Company, USA
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Daniel Kimming, Karlsruhe Institute of Technology, Germany
Josiane Mothe, IRIT, France
Dirk Labudde, Hochschule Mittweida, Germany
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Robert Wrembel, Poznan University of Technology, Poland

IMMM Publicity Chairs

Alessia Saggese, University of Salerno, Italy
Ludovico Boratto, Università di Cagliari, Italy
Toshio Kodama, University of Tokyo, Japan

IMMM 2014 Technical Program Committee

Aseel Addawood, Cornell University, USA
Zaher Al Aghbari, University of Sharjah, UAE
Riccardo Albertoni, Consiglio Nazionale delle Ricerche - Genova, Italy
César Andrés Sanchez, Universidad Complutense de Madrid, Spain
Annalisa Appice, Università degli Studi di Bari Aldo Moro, Italy
Avi Arampatzis, Democritus University of Thrace, Greece
Liliana Ibeth Barbosa Santillán, University of Guadalajara, Mexico
Shariq Bashir, National University of Computer and Emerging Sciences, Pakistan
Bernhard Bauer, University of Augsburg, Germany
Grigorios N. Beligiannis, University of Western Greece - Agrinio, Greece
Jorge Bernardino, ISEC - Polytechnic Institute of Coimbra, Portugal
Konstantinos Blekas, University of Ioannina, Greece
Jacek Blazewicz, Poznan University of Technology, Poland
Ludovico Boratto, Università di Cagliari, Italy
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Omar Boussaid, Université Lyon 2, France
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Alain Casali, Aix Marseille Université, France
Mirko Cesarini, University of Milano Bicocca, Italy
Nadezda Chalupova, Mendel University - Brno, Czech Republic
Chi-Hua Chen, National Chiao Tung University, Taiwan R.O.C.
Yili Chen, Monsanto Company, USA
Been-Chian Chien, University of Tainan, Taiwan
Sung-Bae Cho, Yonsei University, Korea
Kendra Cooper, University of Texas at Dallas, USA
Ronan Cummins, University of Greenwich, UK
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Lois Delcambre, Portland State University, USA
Frantisek Darena, Mendel University - Brno, Czech Republic
Andre Ponce de Leon F. de Carvalho, University of Sao Paulo at Sao Carlos, Brazil
Sébastien Déjean, Université de Toulouse & CNRS, France
Mustafa Mat Deris, University of Tun Hussein Onn, Malaysia
Emanuele Di Buccio, University of Padua, Italy
Qin Ding, East Carolina University - Greenville, USA
Aijuan Dong, Hood College - Frederick, USA
Nikolaos Doulamis, National Technical University of Athens, Greece

Flaminia Luccio, Università Ca' Foscari Venezia, Italy
Qiang Ma, Kyoto University, Japan
Laura Maag, Alcatel-Lucent Bell Labs, France
Stephane Maag, Telecom SudParis / CNRS UMR Samovar, France
Ricardo J. Machado, Universidade do Minho, Portugal
Thomas Mandl, Universität Hildesheim, Germany
Ioannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Francesco Marcelloni, University of Pisa, Italy
Elena Marchiori, Radboud University - AJ Nijmegen, The Netherlands
Ali Masoudi-Nejad, University of Tehran, Iran
Artura Mazeika, Max Planck Institute for Informatics - Saarbrücken, Germany
Johannes Meinecke, SAP AG, Germany
Fabio Mercorio, University of Milano - Bicocca, Italy
Dia Miron, Recognos, Romania
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Charalampos Moschopoulos, Katholieke Universiteit Leuven, Belgium
Katarzyna Musial-Gabrys, King's College London, UK
Erich Neuhold, University of Vienna, Austria
Ulrich Norbisrath, BIOMETRY.com / University of Tartu, Estonia
Samia Oussena, University of West London, UK
José R. Paramá, University of A Coruña, Spain
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Nathalie Pernelle, Université Paris-Sud, France
Jürgen Pfeffer, Carnegie Mellon University, USA
Katja Pfeifer, SAP AG, Germany
Ioannis Pratikakis, Democritus University of Thrace - Xanthi, Greece
Nishkam Ravi, NEC Labs - Princeton, USA
Arno H.P. Reuser, Reuser's Information Services, Netherlands
Ariella Richardson, Jerusalem College of Technology, Israel
Paolo Rosso, Universidad Politécnica Valencia, Spain
Igor Ruiz-Agundez, University of Deusto - Basque Country, Spain
Alessia Saggese, University of Salerno, Italy
Jörg Scheidt, University of Applied Sciences Hof, Germany
Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
Gyuzel Shakhmametova, Ufa State Aviation Technical University, Russia
Mingsheng Shang, University of Electronic Science and Technology of China, China
Armin Shams, University of Tehran, Iran
Josep Silva, Universitat Politècnica de València, Spain
Simeon Simoff, University of Western Sydney, Australia
Cristina Solimando, University Roma Tre, Italy
Theodora Souliou, National Technical University of Athens, Greece
Michael Spranger, University of Applied Sciences Mittweida, Germany
Giovanni Squillero, Politecnico di Torino, Italy
Jaideep Srivastava, University of Minnesota, USA
Vadim Strijov, Computing Centre of the Russian Academy of Sciences, Russia
Tatiana Tambouratzis, University of Piraeus, Greece
Tõnu Tamme, University of Tartu, Estonia
Mehmet Tan, TOBB University of Economics and Technology, Turkey

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Semantic Tools for Forensics: Towards Finding Evidence in Short Messages

Michael Spranger, Eric Zuchantke and Dirk Labudde
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*name.surname*}@hs-mittweida.de

***Abstract*—Mobile devices are a popular means for planning, appointing and conducting criminal offences. In particular, short messages (SMS) and chats often contain evidential information. Due to the terms of their use, these types of messages are fundamentally different from other forms of written communication in terms of their grammatical and syntactic structure. Due to the low price of media storage, messages are rarely deleted. On one hand, this fact is quite positive as possible evidential information is not lost. On the other hand, considering only SMSs, 15,000 and more on only one cell phone is not uncommon. In the most cases of organized or gang crime, there is not one but many devices in use. Analysing this huge amount of messages manually is time consuming and therefore not economically justifiable in the cases of small and medium crimes. In this work, we propose a process chain that enables to decrease the analysis and evaluation time dramatically by reducing the messages which need to be examined manually.**

Keywords–forensic; short message; German; text processing

I. INTRODUCTION

Investigations in criminal cases involve more and more investigating computers, smart phones, tablets and other devices of modern communication. This trend applies not only to computer crime in the strict sense, but also to many cases of classical crime. This is true because, on one hand, victims are easier to find and to spy on in a networked world, and on the other hand, the communications via the Internet or mobile devices have become a standard for our society and hence for sub-societies and criminals. Some of the traces of a crime, in particular those of a textual nature, are accumulated more than ever in the storage of mobile devices. Criminals use modern means of communication to plan their activities or arrange cooperatively committed crimes, as well as to find and contact potential victims. Due to the low price of media storage, messages are rarely deleted. This fact is quite positive since possible evidential information is probably not lost. However, if we consider only SMSs, experience has shown that 15,000 and more on only one smart phone is not uncommon. In addition, mobile devices may contain messages from messenger like ”WhatsApp” whose volume often exceeds ten times the volume of SMSs. If we consider gang or organized crime in general we need to realise that there is not one but many devices in use. Nowadays, the analysis of this huge amount of messages is mainly done by hand using much intuition and experience to separate the significant texts. This manual task is very time consuming and therefore not economically justifiable in the cases of small and medium crimes. Analysing and evaluating forensic texts in an automatic way is generally challenging, as shown by the authors in previous work [1] [2]. Forensic texts, as considered in this work, are texts that are subject to legal considerations with the goal of taking evidence. The analysis of such texts is regularly a branch of general linguistics [3]. In order to perform such analyses on a large amount of texts, methods from the field of computer linguistics are required. These are originated in the crossover of linguistics and computer sciences [4].

Currently, our cooperating local criminal investigation department uses Cellebrite's *Physical Analyzer* [5] for reading data from mobile devices. As a result, a multi-paged Excel or XML-based report with all the raw data reconstructed and gathered from the examined device is generated. Even if the extracted data are presented in a structured way, this particularly does not apply to the contained textual data. These remain in their original form and need to be analysed manually. If this process should be supported by automation, the special characteristics of SMSs as considered in the next section must be taken into account. There are few works dealing with the processing of SMSs. For example, Cooper et. al [6] extracted information from SMSs using manual created structural patterns in order to enrich a library database with current information. Mangan [7] has introduced an approach using structural patterns as well, but generating them by analysing the interdependency distance between slots and keywords. Amailef et al. [8] presented a mobile-based emergency response system for intelligent m-government services based on ontologies. They used a maximum entropy model for extraction of event entities from SMSs. However, we show in Section II-B, that none of these approaches is actually suitable to extract evidential information from forensic texts.

Subsequently, we propose a process chain based on these insights that enables the criminalists to reduce their search space for evidential messages significantly and hence the amount of messages that need to be analysed manually. The proposed process chain is based on an automatic clustering of coherent messages and uses a dictionary and a bag-of-words model for calculating the significance of each cluster with respect to the area of crime under consideration. In this way, the most time-consuming part in analysing SMSs can be accelerated dramatically.

In Section II, the SMS corpus we used is presented. Further, we will introduce the characteristics of this text type we found through the manual analysis of this corpus. Subsequently, in Section III the methodology used for clustering and ranking the messages are described. In Section IV some preliminary results are presented in order to give a first impression of the performance of the presented methods. After a short

 ISBN: 978-1-61208-364-3

summary in Section V, we envision some approaches for further development in Section VI.

II. Subject of Study

A. Forensic Short Messages Corpus

Due to a cooperation agreement between the author's university, the local criminal investigation department and the local public prosecution department, a first dataset of two closed cases of drug crime is provided. In each case, one single smart phone of the suspect has been seized and a physical backup has been created using Cellebrite's *Physical Analyzer* [5]. The backup is provided as an Excel report containing all textual data and meta-data including references to binary files from the cell phone under examination. Table I shows the amount of data currently available for evaluation. Unfortunately, only the SMSs from the first device are manually marked as evidentiary or not. In order to evaluate the results, in this work, we consider only the SMSs as well as some contact information concerning the sender and receiver of such messages contained in the report of the *HTC Desire A9191*.

TABLE I. Corpus containing SMSs

Device	SMSs	Chat messages
HTC Desire A9191	**14,307**	132,345
P743T Skate	810	13,749

B. Characteristics of Forensic Short Messages

Forensic text in general refers to every textual data that may contain evidential information. Their structure and quality regarding grammar, syntax and wording strongly depends on the area of the crime committed by the offenders, their level of education and their social environment. A more detailed description of the general characteristics of forensic texts can be found in [2]. Personalized SMS form the extreme case of these characteristics. They are particularly marked by frequent lack of correct grammatical structures. Therefore it is difficult to use (lexico)-syntactic pattern as in [6] [9] for extracting information of criminalistic relevance. Further, the usage of non-standardised emoticons, abbreviations, emotionally intended character extensions and especially written effects of language erosion caused by language-economic processes make this task more difficult and lead to a failing of known techniques. The following list shows some example texts to illustrate the problem:

- *"aber was ich mein[e] is[t] wir müss[e]n wenn wir weihnacht[e]n gefeiert hab[e]n* ***übelst money hab[e]n****"*
- *"Beruhig[e] dich* ***ich zieh[e] denn*** *das nächste ma[l] rich[tig]* ***fette ab****! :))))))"*
- *"Ich schreib jetzt wegen dir hab ich mein 12g nicht bekommen Weil Du* ***ne*** *aus[de]m* ***knick*** *gekommen bist XD"*

Missing characters are included in square brackets, whereas additional characters are marked by strike-through. Slang-afflicted words and phrases are printed in bold. The most challenging problem in the considered context of SMS with criminalistic relevance is the usage of slang-afflicted language combined with terms of hidden semantics. Hidden semantics refers to one kind of a steganographic code. Such a term is used in its common innocent meaning but its actual semantic background is prearranged by a narrow circle of insiders. For example, the question

"Bringst du ein Wernesgrüner mit? (Can you bring a Wernesgrüner?)"

appears innocent because the term *Wernesgrüner* is used as a beer brand. But, within the actual context the meaning of this term is marijuana. Note that in this example we intentionally do not use slang to avoid misunderstandings. But commonly terms of slang are mixed in regularly. These characteristics make it difficult even for criminalists and linguists with years of experience to read and understand the semantics of forensic SMSs.

If it becomes clear that any information not found by the system may be crucial in proving the guilt or innocence of a criminal suspect, then it follows that decisions concerning the evidential value of forensic SMSs cannot be made by a machine.

III. Methodology

In the last section, we explained why a fully automated solution should be rejected currently. For this reason, the way we propose is to decrease the effort of a manual search by reducing the search space automatically. This way the criminalist is able to find evidentiary information in a significantly shorter time.

Therefore, in this work, we outline a process chain towards reducing the search space by filtering the contacts which have exchanged significant messages and providing the corresponding conversation. The process is divided in two steps:

1) clustering of coherent messages to conversations
2) calculating a significance value for ranking conversations

A. Clustering Coherent Messages

As we stated earlier, we cannot be sure to identify all of the significant exchanged messages contained in a corpus. But, we can increase this probability simply by trying to detect significant conversations instead of concrete messages. A conversation is by definition a set of semantically and temporally coherent messages.

More formally, let $c = m_0, ..., m_n$ whereby c is a conversation and $m \in M$ a concrete message from the set of all messages in a temporal relationship contained in the corpus under consideration. In order to create clusters of coherent messages we analysed the length of the pauses between two messages. Direct comparison of these values with the areas, manually marked as significant by criminalists, reveals that long pauses are rarely situated within these areas. Figure 1 shows a clipped part of such a diagram regarding the conversation between two persons. The x-axis represents all messages of the considered subset of the corpus. The y-axis in positive direction represents the pause length between one

 ISBN: 978-1-61208-364-3

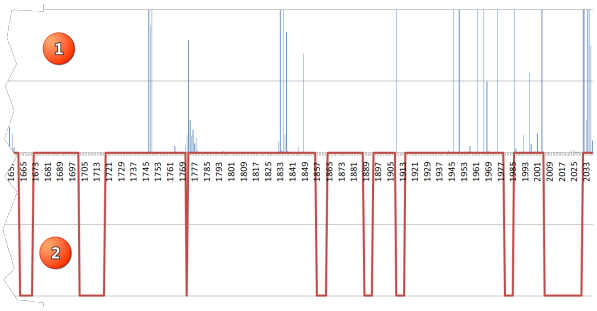


Figure 1. Clipped part of a diagram, that directly compares the pause lengths (1) of all SMS in the corpus and the manually marked significant conversation areas (2).

message to the corresponding answer. In negative direction, the y-axis shows the manually marked significant areas.

This observation leads to the approach to use the $Q_{0.75}$-Quantile of the length of pauses as threshold for the decision whether a message belongs to one conversation or is already part of the following. Applying this approach to a subset of the corpus with 3152 messages exchanged by two persons within one year, 352 conversations could be detected. The threshold was determined empirically. Experience shows that the pause length strongly depends on the individual communication behaviour. Therefore, the universality must be tested on other corpora, which, however, are currently not available.

B. Ranking Conversations

When the set of conversations $C = \{c_0, ..., c_n\}$ have been created the next step is to find out which of these are significant regarding the object of investigation. Respecting the insights from Section II-B, we decide to use a bag-of-words model combined with a domain specific dictionary d to assign a significance value to each conversation and hence to each person being part of it. This significance value S can be calculated depending on the frequency of domain-specific terms (see (1)).

$$S_i = bag(c_i, d), \forall c \in C \quad (1)$$

These values form the basis of a heat scale we use to colour the contacts in the contact network established using the report data. Figure 2 shows the overall process. The starting point is a contact network based on the data gathered by the *Physical Analyzer* [5]. The exchanged coherent messages are clustered into conversations as mentioned earlier. Subsequently, the significance value is calculated for each of these conversations. Based on these values suspicious contacts and communications will be marked using the corresponding colours from the heat scale.

The determining factor for a good result is a good dictionary. A dictionary that comprises local language conditions, as well as terms from different categories of offences, is currently not available (at least in Germany). Therefore, we need to create an appropriate dictionary for each offence category and each local cultural circle before we can start to calculate the significance of a conversation.

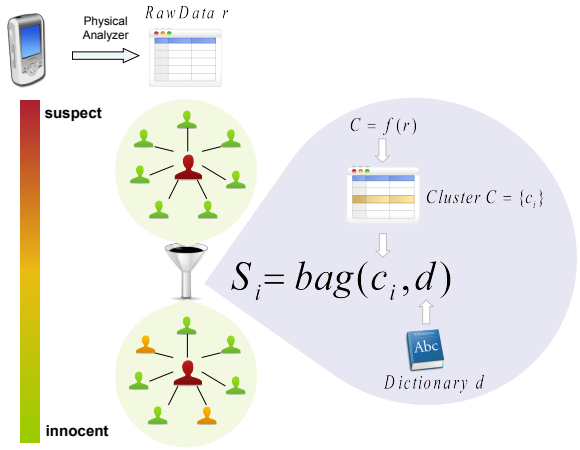


Figure 2. The process of detecting suspicious communication.

C. Creating the Dictionary

We started dividing the corpus into significant and non-suspicious parts and performing a discriminant analysis involving stop-word elimination and stemming. Considering only the frequency classes 1 and 2 (words exclusively in suspicious texts and words relatively more frequent in such texts) we found 882 ”suspicious” terms. Using these terms in turn for processing the whole dataset for evaluation we achieve 98.5% sensitivity with 100% precision. Looking at the distribution of hits, so we found that the most of them are unique. The reason for this is due to the high number of unique spellings, caused by syntactic and typographical errors as well as deliberate word extensions. However, these lists of terms can form a basis for the dictionary, especially if more than one corpus is taken into account and words are sorted out with respect to their frequency within all corpora. In addition, it is useful

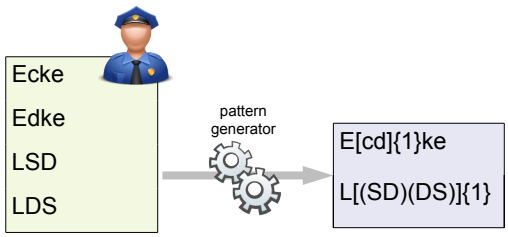


Figure 3. Generating a pattern dictionary by transforming criminalist's knowledge.

to integrate the knowledge of the local criminalists who deal with similar cases in a similar environment every day. This experiential knowledge is the best source of information for both, slang and hidden semantics. The manually added terms need to be extended automatically, for example, by twisting letters and transforming in patterns, e.g., regular expressions using an appropriate pattern generator (see Figure 3).

IV. Performance of a Prototype

Due to the lack of other annotated corpora the first dictionary described in Section III-C has been filtered and extended manually with the help of specialists in the field of drug crime.

 ISBN: 978-1-61208-364-3

In an effort to quantify the performance of the preliminary approach described in this paper, the 352 clusters of coherent messages (see Section III-A) have been filtered using the available dictionary and employing the algorithm described in Section III-B. This implementation of the proposed process chain achieved 67% sensitivity with 100% precision. The cause of the low sensitivity is due to the coverage of the created dictionary, which is, in comparison to the coverage of a comprehensive and ready-to-use dictionary, relatively low. Therefore, the improvement of the dictionary is in the focus of further development. However, the work load necessary for manual search decreased to only 15 %.

V. Conclusion

The manual analysis of forensic SMSs gathered from mobile devices during the criminal proceedings is very time-consuming and not economically justifiable for small and medium crimes. We have shown that extracting information from forensic SMSs in an automatic way is challenging. Considered existing methods are limited to specified domains and require a predominantly correct language usage and fail if applied on forensic SMSs. The reason for this is, among others, mainly missing standardized structures and the strong use of local dialect as well as an emotionally-influenced style of writing. Therefore, we proposed a process chain for decreasing the manual effort in analysing such messages by reducing the search space. In order to do this we cluster coherent messages to single conversations. Subsequently, we calculate a significance value using a bag-of-words model and a dictionary. Applied to a real world dataset - a closed case of drug crime provided by the local public prosecution department - we could evaluate the process chain with acceptable preliminary results.

VI. Further Work

Currently, we are trying to improve the calculation of the significance value by applying a similar bootstrapping algorithm as presented in [2] for the field of categorising forensic texts in general.

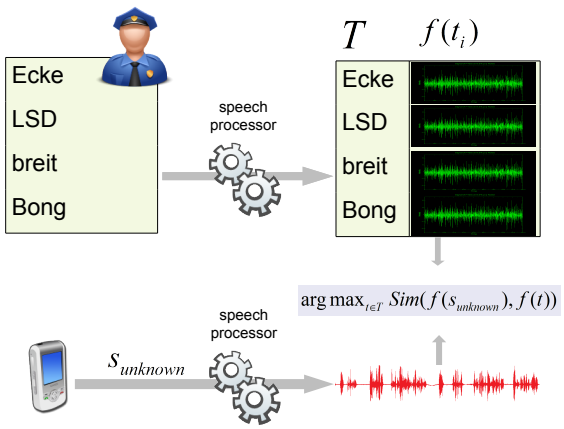


Figure 4. Dictionary containing pronunciation profiles as a basis for matching terms with high failure tolerance.

For testing the universality of the proposed process chain and especially the dictionary we need further corpora. Fortunately, the local prosecutor's office has announced the release of additional data. Another approach we currently consider is to create a dictionary, as well as a corresponding algorithm for calculating the significance value with a high failure tolerance as shown in Figure 4. Here, pronunciation profiles are used as a basis for understanding special terms.

Acknowledgment

The authors would like to thank the members of the criminal investigation department and the public prosecution department Chemnitz (Germany). We acknowledge funding by "Europäischer Sozialfonds" (ESF), the Free State of Saxony and the University of Applied Sciences Mittweida.

References

[1] M. Spranger, S. Schildbach, F. Heinke, S. Grunert, and D. Labudde, "Semantic tools for forensics: A highly adaptable framework," in Proc. 2nd. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2012, p. 27 to 31.

[2] M. Spranger and D. Labudde, "Semantic tools for forensics: Aproaches in forensic text analysis," in Proc. 3rd. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2013, p. 97 to 100.

[3] H. Kniffka, Working in Language and Law. A German perspective. Palgrave, 2007.

[4] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, Computerlinguistik und Sprachtechnologie - Eine Einführung, 3rd ed. Spektrum Akademischer Verlag, 2010.

[5] C. M. S. LTD". Ufed physical analyzer - mobile daten ermitteln, dekodieren und bereitstellen. [Online]. Available: http://www.cellebrite.com/de/mobile-forensics/products/applications/ufed-physical-analyzer (2014.05.21)

[6] R. Cooper and S. Ali, "Extracting data from short messages," in Natural Language Processing and Information Systems, LNCS 3513. LNCS, Springer, 2005, pp. 388–391.

[7] D. H. W. Dannis Muhammad Mangan, "Information extraction from short text message in bahasa indonesia for electronics," Jurnal Sarjana ITB bidang Teknik Elektro dan Informatika, vol. 1, 2012, pp. 29–32.

[8] K. Amailef and J. Lu, "Mobile-based emergency response system using ontology-supported information extraction," in Handbook on Decision Making, ser. Intelligent Systems Reference Library, J. Lu, L. Jain, and G. Zhang, Eds. Springer Berlin Heidelberg, 2012, vol. 33, pp. 429–449.

[9] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in Proceedings of the Eleventh National Conference on Artificial Intelligence, ser. AAAI'93. AAAI Press, 1993, pp. 811–816.

 ISBN: 978-1-61208-364-3

Class Strength Prediction Method for Associative Classification

Suzan Ayyat
Department of Informatics
Huddersfield University
Huddersfield, UK
U1277021@hud.ac.uk

Joan Lu
Department of Informatics
Huddersfield University
Huddersfield, UK
j.lu@hud.ac.uk

Fadi Thabtah
Ebusiness Department
Canadian University of Dubai
Dubai, UAE
fadi@cud.ac.ae

***Abstract*—Test data prediction is about assigning the most suitable class for each test case during classification. In Associative Classification (AC) data mining, this step is considered crucial since the overall performance of the classifier is heavily dependent on the class assigned to each test case. This paper investigates the classification (prediction) step in AC in an attempt to come up with a novel generic prediction method that assures the best class assignment for each test case. The outcome is a new prediction method that takes into account all applicable rules ranking position in the classifier beside the class number of rules. Experimental results using different data sets from the University of California Irvine (UCI) repository and two common AC prediction methods reveal that the proposed method is more accurate for the majority of the data sets. Further, the proposed method can be plugged and used successfully by any AC algorithm.**

Keywords-associative classification; data mining; prediction phase.

I. INTRODUCTION

Associative Classification (AC) is an emerging classification research topic which employs association rules to solve classification problems in data mining [1]. The goal of the AC method is to learn a classification model from input classified data (historical data) that in turn is used to assign the right target class in new data (test data). For example, in text categorization the target class is the document's category. This type of application can be seen as supervised learning because learning is focused on a special attribute in the training data set called the target class.

Recent experimental research [2][3] indicated that AC methods usually devise good classification models in terms of predictive accuracy when contrasted to other classification methods such as statistical, covering, and decision trees. For instance, in a recent research study [4], and using 20 University of California Irvine (UCI) data sets [5] the accuracy of an AC algorithm called MAC [4] is 1.86%, 3.12 % and 3.11% higher than PART[6], RIPPER [7], and C4.5 [8] algorithms, respectively. This evidence, if limited, reveals the predictive power of AC when building classification models which increase the usage of this type of classification models in applications. The main reason for learning high accurate classification models by AC approach is the new rules (knowledge) induced during the learning step where the majority of the items and the target class combinations in the training data are evaluated for possible positive correlations [9]. Nevertheless, the number of rules could be large [10].

Recently, a number of AC algorithms have been developed in the research literature like, CBA [1], CPAR [11], LC [9], MAC [4] and others. These methods utilise various methodologies to induce rules, store rules, prune rules and predict the class of test data. This paper concentrates on the class prediction step in AC. Predicting the right class of a test data by the classifier is considered the most important step in the AC algorithm's lifecycle. In this step, the AC algorithm uses the rules learnt to predict the class labels of the test data and coming up with the right class for each test data is crucial because the overall predictive performance of the classifier depends on this decision. In addition, choosing the right rules in the classifier to assign the predicted class is a challenging tasks [12][13]. This is since there could be multiple rules applicable to the test data yet associated with different class labels.

To deal with the class prediction step in AC, we develop a novel class prediction method that considers all possible rules applicable to the test data. This is unlike most existing methods that:

a) Either consider the first rule in the classification model that is similar to the test data items [14].
b) Or computes rules' weights based on complex mathematical formula [11][15].

The main problem that this paper addresses is the inability of existing AC prediction methods of making use of all class labels in the classification model in cases when there are more rules similar to the test data items. For example, suppose we have a test data (a,b,c) that requires classification, and we have in the classification model 3 rules R1: (a^b, class2), R2:(b^c, class1), and R3:(b, class1). Assume that R1>R2>R3 in the classification model. Now, most existing AC methods like CBA, MCAR and MAC allocate class2 to the test data dismissing rules (R2 and R3) which indeed jeopardizes the prediction decision. On the other hand, few AC algorithms, like CMAR, groups rules applicable to the test data, with respect to their class labels, and then computes each group's rules support and confidence. This is problematic especially when we have a large number of rules similar to the test data or the number of test data to be classified is huge. So we intend to use all classes of the rules that are similar to the test data items for prediction. Then, when computing the accuracy of the classification model, the fired rule's assigned class to the

 ISBN: 978-1-61208-364-3

test data is counted which makes the decision more reliable and will possibly enhance the accuracy of the model.

Our new class prediction method considers the class labels of rules similar to the test data and gives the test data the class belonging to the highest score. Later in Section 3, we show how the class score is computed. This prediction is more realistic than one existing rule prediction methods simply because none of the applicable rules that are similar to the test data is ignored. The research question that the article is trying to solve is:

- Can we come up with a prediction method that takes into account both the rules position in the classifier and the class representation in the context of number of rules in order to have a fair and accurate prediction decision?

This paper is structured as follows: The literature review is given in Section 2. Our prediction method is discussed in Section 3 along with a detailed example. Section 4 is devoted to present the comparison results between the proposed methods and other classification methods in AC. Finally, the conclusions are given in Section 5.

II. LITERATURE REVIEW

Generally, there are two main methods in predicting the class of test data in AC. The first method is a group-based method that assigns the class that belongs to a group of rules to the test data during the classification step. The second method takes on only a single rule class often the class of the first rule (highest position one) similar to the test data items. This is the class that these methods assign to the test data to determine whether the test data are a hit or a miss when computing the model's accuracy. Typical algorithms that employ this kind of prediction are MAC, MCAR, CBA and many others. This prediction method assumes:

1) The rules in the classification model are sorted based on certain criteria
2) Only one rule is used for prediction

The second condition above has been criticized by several researchers [4][15], due to the following facts:

1) There could be more than one rule similar to the test data
2) These matching rules may have close ranking position in the classification model

Therefore, using a single rule is seen to be biased and an unfair decision. Nevertheless, this approach is simple, especially in circumstances when there is only one rule similar to the test data.

In circumstances when multiple rules are similar to the test data during the prediction step, the decision to only fire one rule becomes debatable. Consequently, a more fair decision is to use the information provided by all matching rules for the class prediction decision. In 2011, two prediction methods were proposed by Thabtah F. et al. [12] on using the classifier's rules confidence values matching the test data to make the prediction decision. The first prediction method groups all rules matching the test data into collections based on their classes and then the average confidence for each collection is computed. This method assigns the class of the group with the largest average confidence. The other method described by Thabtah F. et al. [12] is similar to the method described above but it does not require that all items of the rules be identical to the test data items by allowing partial similarity than full similarity aiming to have larger number of rules within the collections.

Veloso A. et al. [16] developed a prediction method in AC that utilises all rules applicable to the test data after dividing them into groups with respect to their class labels. Then, a group score consisting of the confidence and support value of the rule(s) is computed and the group class with the largest score is given to the test data. A few years ago, a greedy AC called CPAR used a multiple rules prediction method based on the Laplace expected accuracy. This method works as follows: For a test data (t) that is about to be predicted, the method groups all rules in the classifier contained in t in groups based on the rules class labels. Then the group expected accuracy average is calculated and t is allocated the largest group's expected accuracy class. The group's Laplace expected accuracy is computed according to the equation below:

$$\text{Laplace (Cluster)} = \frac{(D_c(r\_group) + 1)}{(D_{tot}(r\_group) + D)} \quad (1)$$

where

D_c (r_group) is the number of training instances covered by the group's rule (head and tail).

$D_{tot}(r\_group)$ is the number of training instances similar to the group's rule body.

D is the number of class labels in the training data.

III. THE PROPOSED PREDICTION METHOD

In this section, we discuss the main contribution which is the development of a novel AC prediction method that will enhance the predictive power of any AC algorithm in forecasting test data. Our method falls under the category of a group-based method to come up with the most accurate class to assign to the test data during the classification step. The method assumes the following before it gets invoked:

1) All rules are generated and the classifier is built.
2) All rules within the classifier are sorted according to any sorting procedure.

So, when a test case is demanding a class during the prediction step, our method (Figure 1) works as follows:

It scans the classifier and marks any rule that is similar to the test data items. Here, we have several situations:

a) When there is only a single rule matching the test data items the situation is simple and we assign the class of that rule to the test data.
b) When more than one rule is similar to the test data and all of them are connected with a similar class, our

method assigns that class to the test data in a straightforward manner.

c) When multiple rules are similar to test data and these rules are connected with different class labels the situation becomes challenging and this is where the novelty of our method applies. Firstly, our method clusters the applicable rules into groups with respect to their classes. Then, based on both the rules rank in each cluster and the cluster size the decision of which class to assign to the test data is decided. We have combined both "the rules rank per group" and the "size of the group" into a ranking formula that we name the Class_Strength as shown in the equation below.

Class_Strength Ci=Score Ci + Ci Number of Rules (1)

$$\text{Score Ci} = \sum_{i=1}^{j} n - (Xci - 1) \quad (2)$$

Where

Xi is the number of rules matching the test data for

class ci

n is the total number of rules matching the test data

So, for each group's class, its strength will be calculated and the class belonging to the group that has the largest strength gets assigned to the test data. In the case that more than one group has the same number of rules; the choice will be based on the class representation (number of rules per group).

This method takes advantage of two previous group-of rule prediction methods in AC; the one that considers the rules rank are the primary criteria (confidence) and the other that considers the class representation per rule (number of rules). We have combined both procedures into a novel measure called the Class_Strength for a more legitimate prediction decision. The class assignment of test data has improved when contrasted with classification procedures such as that of CBA and its successors that take the class of the first ranked rule in the classifier matching the test data to make the prediction decision. Furthermore, it also overcome multiple rule prediction methods in AC like CPAR and CMAR that employ mathematical based attribute assessment formulas, e.g., confidence, support, weighted Chi-Square. Now, instead of favouring rules with high confidence (ranking position) or rules belonging to the most representative class to make the classifications decision, the new measure takes advantage of both approaches which give the decision of assigning class to test data legitimacy and accuracy. Finally, when no rules in the classifier are applicable to the test case, the default class rule will be assigned to that case.

Input: test data set (*T*), Classifier (C)
Output: Error rate *E*

1 Iterate over *T*s
2 Iterate over C
3 locate rules that are similar to the current test data
4 cluster the rules per class label
5 compute the class strength per cluster according to Equation (1)
6 assign the class with the largest strength to the current test data
7 end
8 end if
9 else assign the default rule to the current test data
10 end if
11 end
12 end
13 compute the number of errors of *T*

Figure 1 The new prediction method

Example

Consider the test data shown in Table I to be predicted, Table II shows all relevant rules from the classifier that are similar to the test data. The similarity has been based on the rule's body and the test data items. Now a typical AC like CBA or MCAR will take on rule rank # 1 and will allocate its class, i.e. (c3), to the test data simply because this is the best ranked rule matching the test data. On the other hand, a class representation based prediction method, like MAC, assigns class (c1) to the test data since this class has the most number of rules matching the test data. For our prediction method, we first divide the rules in Table II into groups based on their class labels as shown in Table III. We then compute the new rule score based on the Equation (1), i.e $n - (Xi - 1)$ and the rules score are shown below in Table IV. Finally, we sum up each class score with the number of rules belonging to it to derive the class strength, as shown in Table V. In this example, we have a tie score between class c_3 and c_1. Nevertheless, we assign class c_1 to the test data since it has a larger number of rules.

TABLE I. TEST DATA

Attribute1	Attribute2	Attribute3	Attribute4	Class
a_1	b_1	l_1	g_5	?

TABLE II. RULES MATCHING THE TEST DATA OF TABLE I

Rank	Rules
1	g_5 ^ $\rightarrow c_3$
2	l_1^b_1 $\rightarrow c_2$
3	a_1 ^ $\rightarrow c_1$
4	a_1 ^b_1 $\rightarrow c_1$

TABLE III. NUMBER OF RULES PER CLASS

Class	# of Rules
c_3	1
c_2	1
c_1	2

TABLE IV. RULES SCORE CALCULATIONS

Rank	Rule	Score Calculation	Rule weight
1	g_5 ^ $\rightarrow c_3$	4 - (1-1)	4
2	l_1^b_1 $\rightarrow c_2$	4 - (2-1)	3
3	a_1 ^ $\rightarrow c_1$	4 - (3-1)	2
4	a_1 ^b_1 $\rightarrow c_1$	4 - (4-1)	1

TABLE V. CLASS SCORES

Class	Class Score Eq. 1	Class Strength Score+#of Rules
c_3	4	4+1=5
c_2	3	3+1=4
c_1	3	3+2=5

IV. EXPERIMENTAL RESULTS

A. Settings

Different data collections from the UCI repository [5] have been utilised to measure the impact of the new prediction method on the classification accuracy of the classifiers resulting from the experiments. We have used 12 data sets that have different size and attribute types. Tenfold cross validation testing method has been used to run the experiments. This method is used in data mining research to derive accurate and fair results. In particular, it divides the input data set into 10 folds randomly and the classifier is trained on 9 folds and then tested on the hold out fold to derive its error rate. The same process is repeated 10 times and the accumulated results are then averaged.

We have selected two main prediction methods in AC to compare our results with mainly because they use different prediction methods for assigning the class labels to test data. The main measure used for comparing these methods and ours is the error rate since we would like to answer the question "whether combing rules rank with class number of rules enhance the predictive power of AC algorithms?". The first methodologies used are based on CBA and MCAR and consider the highest rank rule prediction [1][14]. The second method was recently developed and uses a group-based method that considers the class associated with the maximum number of rules [4]. All these methods and ours have been implemented in Java in MCAR algorithm.

In the experiments, the AC main parameters, which are minimum support (*minsupp*) and minimum confidence (*minconf*), have been set to 2% and 50%, respectively. The reason for setting the *minsupp* to 2% is because it has been used previously by many research studies and proved to be fair in compromising between the number of rules extracted and the accuracy rate. The *minconf* has a limited effect on the performance of the AC algorithm so we have set it to 50%. Lastly, the experiments have been performed on I3 PC with 4.0 GB RAM and 2.7 GH processor.

B. Results Analysis

We have generated the error rate of the considered prediction methods on 12 UCI data sets, as shown in Figure 2. The figure clearly demonstrated that the proposed prediction method has enhanced the predictive rate of the classifiers devised on the data sets. In particular, our method achieved a decrease in the error rate on average by 1.18% and 1.12% on the 12 data sets we consider when contrasted with MAC and CBA algorithms respectively. A possible reason for the decrease in the error rate is mainly due to that new prediction methodology that allocates the test data the most appropriate class based on the class strength that we compute during the prediction step and for each test case. The fact that we consider both the rules rank and the class of rules for each class cluster gives a legitimate and accurate decision of which class to assign. This is since we have accounted for multiple rules and considered these rules rank in a new formula that assures a score for each class. In other words, we allocate the class of the cluster having the largest score (strength) to the test data based on both number of rules applicable to the test data and these rules

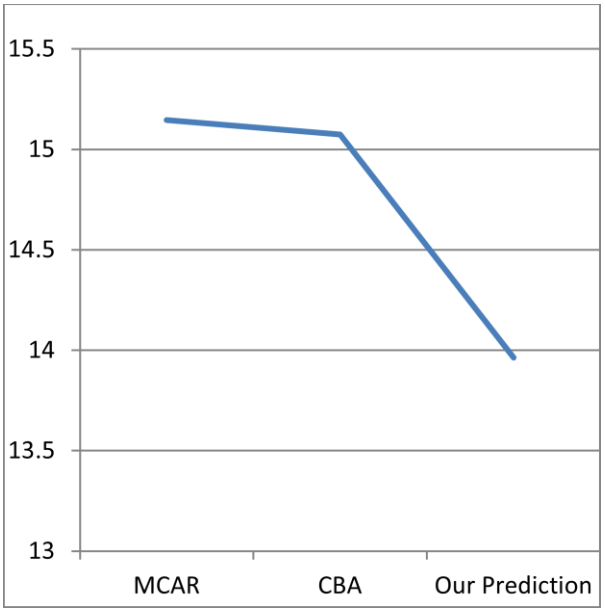


Figure 2 Average error rate produced from the UCI data sets by the prediction methods

rank in the classifier. This, surely, should minimize the error rate of the classifier, as shown in figure 2.

We have looked into more detailed results and for each UCI data sets, as depicted in Table VI. The figures in the table reveal consistency in the error rate results between the prediction methods we consider in this article. This means there are no large significant differences in the error rate results for most of the data sets for both CBA and MAC, except the fact that we have improved the predictive power of the classifiers for most of the data sets. Precisely, our prediction method outperforms both MAC and CBA prediction methods on most of the data sets and the won-tie-lost records are 10-2-2 and 9-3-0, respectively. The fact that our method investigates both the rank of the rules and the class representation per rules has a definite advantage and the most suitable class gets allocated to the test data.

Table VII displays the runtime for the prediction phase in seconds computed from two implementations (MAC single rule and our multi-rule prediction methods) for a sample of the data sets. It is obvious from the figures in the table that our prediction method normally takes longer to forecast test data than single rule based methods such as MAC. Nevertheless, the proposed prediction method has enhanced the predictive performance of the final classifiers if compared to those of MAC. In addition, the time spent in assigning test cases the right class labels is not excessive according to Table VII. There should be a tradeoff between precision and test data prediction time where longer time can be tolerated in exchange for higher level of predictive accuracy.

TABLE VI. THE ERROR RATE OF THE CONSIDERED PREDICTION METHODS ON 12 UCI DATA SETS

Data set	Size	MAC	CBA Prediction	Our Prediction
Breast	699	5.36	6.76	5.42
Cleve	303	18.54	16.9	18.36
Glass	214	24.76	23.47	22.58
Heart	294	18.8	18.13	18.1
Hybothroid	3772	6.3	7.68	6.3
Iris	150	7.06	6.69	5.74
Labor	57	16.49	13.67	14.04
Led	3200	28.1	30.53	25.2
Lymph	148	26.08	25.57	23.1
Pima	768	24.44	25.42	24.44
Tic-tac	958	1.02	1.04	0.18
Wine	178	4.8	5.04	4.1

TABLE VII. THE RUNTIME FOR PREDICTION IN SECONDS

Data set	MAC	Our Prediction
Cleve	0.06	0.17
Breast	0.25	0.38
Glass	0.03	0.16
Iris	0.02	0.09
Pima	0.08	0.19
Tic-Tac	0.14	0.22
Led	0.19	0.38
Heart	0.015	0.12

V. CONCLUSIONS AND FUTURE WORKS

Predicting test data in AC is an interesting research problem that requires careful consideration due to the fact that more than one rule could be similar to the test data and that makes the prediction decision a hard task. This paper presented a prediction method based on two main criteria:

- The class representation in the context of the numbers of rules
- The rules rank

The outcome is a novel method which considers all rules that are similar to the test data during the classification step and computes the class strength per class assigning the class that has the largest strength. The class strength is based on the rules ranking position as well as the number of rules per class. Experimentations using 12 data sets from the UCI data repository and two common AC prediction methods have been conducted to measure the success and failure of our method. The results with respect to one-error rate reveal that the new prediction method has enhanced the predictive power of the resulting classifiers and on most data sets we used. In the near future, we would like to use our prediction method on unstructured textual data in the domain of text mining.

REFERENCES

[1] Liu B., Hsu W., and Ma Y. “Integrating classification and association rule mining”. Proceedings of the KDD. New York, NY 1998, pp. 80-86.

[2] Jabbar M. A., Deekshatulu B. L., and Chandra P. “Knowledge Discovery Using Associative Classification for Heart Disease Prediction”. Advances in Intelligent Systems and Computing, Volume 182, 2013, pp. 29-39.

[3] HooshSadat M. and Zaiane O. “An Associative Classifiers for Uncertain Datasets”. Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKFF 2012), Malaysia, 2012, pp. 342-353.

[4] Abdelhamid N., Ayesh A., Thabtah F., Ahmadi S., and Hadi W. “MAC: A multiclass associative classification algorithm”. Journal of Information and Knowledge Management, Volume 11, issue 2, 2012.

[5] Merz C. and Murphy P. “UCI repository of machine learning databases”. Irvine, CA, University of California, Department of Information and Computer Science, 1996.

 ISBN: 978-1-61208-364-3

[6] Frank, E. and Witten, I. "Generating accurate rule sets without global optimization". In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Madison, Wisconsin, pp. 144–151,1998.

[7] Choen W. and Singer Y. "Context-Sensitive Learning Methods for Text Categorization". ACM Transactions on Information Systems, Volume 17, No. 2, pp. 141–173, 1999.

[8] Quinlan, J. "C4.5: Programs for machine learning". *San Mateo, CA: Morgan Kaufmann the KDD. New York, NY.*,1998 pp. 80-86.

[9] Thabtah F., Mahmood Q., McCluskey L., and Abdel-jaber H . "A new Classification based on Association Algorithm". Journal of Information and Knowledge Management, World Scientific, Volume 9, No. 1, pp. 55-64, 2010.

[10] Wang X., Yue K., Niu W., and Shi Z. "An approach for adaptive associative classification". Expert Systems with Applications: An International Journal, Volume 38 Issue 9, pp. 11873-11883, 2011.

[11] Yin, X. and Han, J. "CPAR: Classification based on predictive association rule". Proceedings of the –the SIAM International Conference on Data Mining -SDM, pp. 369-376, 2003.

[12] Thabtah F., Hadi W., Abdelhamid N., and Issa A. " Prediction Phase in Associative Classification". Journal of Knowledge Engineering and Software Engineering. WorldScinet ,Volume: 21, Issue: 6, pp. 855-876, 2011.

[13] Abdelhamid N., Thabtah F., and Ayesh A. " Phishing detection based associative classification data mining". *Expert systems with Applications Journal*. 41 pp.5948–5959, 2014.

[14] Thabtah, F., Cowling, P., and Peng, Y. "MCAR: Multi-class classification based on association rule approach". Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications. Cairo, Egypt 2005, pp. 1-7.

[15] Li, W., Han, J., and Pei J. "CMAR: Accurate and efficient classification based on multiple-class association rule". Proceedings of the IEEE International Conference on Data Mining –ICDM, pp. 369-376, 2001.

[16] Veloso A., Meira W., Gonçalves M., and Zaki. M . "Multi-label Lazy Associative Classification". Proceedings of the Principles of Data Mining and Knowledge Discovery - PKDD, pp. 605-612, 2007.

 ISBN: 978-1-61208-364-3

A Concept-based Feature Extraction Approach

Ray R. Hashemi
Department of Computer Science
Armstrong State University
Savannah, GA, USA
Ray.Hashemi@gmail.com

Azita Bahrami
IT Consultation
Savannah, GA, USA
Azita.G.Bahrami@gmail.com

Nicholas R. Tyler
Department of Biology
Armstrong State University
Savannah, GA, USA
Romtinian@gmail.com

Matthew Antonelli
Department of Computer Science
Armstrong State University
Savannah, GA, USA
Mattr.Antonelli@gmail.com

***Abstract*—A concept has a perceived property and a set of constituents. The goal of this investigation is about extraction of meaningful relationships, if any, between the perceived property and the constituent's attributes. Such meaningful relationships (features) may be used as a prediction tool. The presented methodology for extracting the features is based on the concept expansion. To the best of our knowledge, feature extractions based on a concept expansion approach, for use in data mining, has not been reported in the literature. The goal was met by introducing the b-concept, conceptualizing a universe of objects using b-concept, and generating the complete gamma-expansion (CGE) of the b-concepts. The features were extracted from CGEs as anchor prediction (AP) rules. The AP rules were crystalized by a sequence of horizontal-vertical reductions. The prediction powers of the AP rules and their crystalized version were investigated by: (i) using 10 pairs of training and test sets, and (ii) comparing their performances with the performance of the well-known ID3 approach over the same training and test sets. The results revealed that the AP rules and ID3 have similar performances. However, the crystallized prediction rules have a superior performance over the AP rules and ID3. The average of the correct prediction is up by 17%, the average of the false positive is down by 13%, and the average of false negative is up by 3%. In addition, the number of test objects that cannot be predicted is down by 7%.**

***Keywords-b-concept; Concept expansion; Concept Analysis; Data Mining; Prediction Systems; and Crystallizing Prediction Rules*.**

I. INTRODUCTION

A concept is an abstract object possessing a perceived property [1]. For example, a "carcinogen agent" is a concept and its perceived property is that it causes, say, liver cancer. The constituents of a concept are a set of concrete objects described by their own set of attributes. Since a concept has a perceived property, it is considered a proper vehicle for investigation of the possible relationship between its perceived property and its constituents' attributes. Such a feature extraction is more successful when the relationships among the concepts are also established. Building super-concepts and sub-concepts are a part of this effort. Several concepts may create a super-concept and a given concept may serve as a sub-concept of one or more super-concepts [2][3][4]. We introduce the *complete γ-expansion (CGE)* of a concept and provide a methodology to identify a new relationship between a super-concept and its sub-concept(s) using CGE. A concept has a *CGE*, if every sub-expansion of the concept satisfies a given condition set, γ.

The goal of this research effort is three-fold: (i) introducing b-concept, conceptualizing a large dataset of concrete objects (chemical agents) using the new b-concept, and if it is applicable, building the CGE of the concepts, (ii) Extracting the features from the CGEs as the prediction rules and crystalize them by horizontal and vertical reductions, and (iii) compare the prediction power of the prediction rules and crystalized version of the prediction rules against the prediction power of the decision tree approach of ID3 [5].

The remaining organization of this paper is as follows. The Related Works is covered in Section 2. The Methodology is introduced in Section 3. The Empirical Results are discussed in Section 4. The Conclusions and Implications for Future Research is the subject of Section 5.

II. RELATED WORKS

The concept-based analysis is done primarily for building the internal conceptual structure of a given body of objects [6][7], conducting data mining [3][8], and performing image understanding [9]. As a result, the formal concepts [1][4], rough concepts [10][11], fuzzy concepts [12][13] and other forms of concepts [9] have been developed. There are some efforts in learning from the concepts for the purpose of performing a prediction process [3][8]. However, in such efforts number of generated

 ISBN: 978-1-61208-364-3

concepts are limited and so the number of perceived properties. One may argue that every object can be considered as a concept with its own perceived property. Thus, it makes more sense that every possible perceived properties participate in the process of extracting features and not a limited number of them. The proposed methodology supports the total inclusion of all the possible perceived properties (inclusion trait).

The concept expansion has been heavily investigated in information retrieval for the purpose of query expansions to retrieve more relevant or pseudo-relevant documents (objects) from a corpus of documents [14][15]. In general, the concept expansion is done by changing the "bag of terms (features)" that are relevant to the query to a new larger bag of features that seems more relevant. In fact, such expansion tries to include more relevant features to improve the retrieval of more relevant objects. Such methodology does not have any application in mining data for prediction. In contrast, the concept expansion that we propose includes more relevant objects to improve the extraction of more relevant features (inducing trait) and it has a great potential to serve as a prediction approach.

To the best of our knowledge, there is not any existing concept-based prediction methodology that supports both inclusion and inducing traits.

III. METHODOLOGY

First, some terminologies need to be defined. Second, the expansion of the super-concepts is presented. Third, the extraction of features is explored, and finally, the crystallization of the extracted features is investigated.

Definition 1: Let U be a universe of objects and c = {O_1, . . ., O_s, . . ., O_n} ⊂ U. The subset c makes a b-concept, if $f(A_j^i, A_j^s) \leq b$, (for i =1 to n, i≠ s, and j =1 to m) where, A_j^i is the j-th attribute of O_i, m is the number of attributes for O_i, $f(A_j^i, A_j^s) = \sqrt{(A_j^i - A_j^s)^2}$, and b is a constant value. O_1, . . ., O_s, . . ., O_n are the members of the b-concept c, c = b-concept(O_s), and O_s is the concept's anchor—G(c) = O_s.

Definition 2: If G(c_k) ∈ c_m, then c_k is a sub-concept of c_m ($c_k \leq c_m$, where ≤ is a binary relation) and c_m is a super-concept of c_k.

Definition 3: Let L be all the concepts of U. L is a partial ordered set with binary relation of ≤ and (L, ≤) is a compete lattice.

As an example, let us consider the set of objects in Table 1 and b = 3. Using definition 1, the concept c_4 with the anchor of O_4 is composed of the following objects c_4 = {O_4, O_5, O_6, O_7}. The concept c_6 with the anchor of O_6 includes objects of O_4, O_6, and O_7, c_6 = {O_4, O_6, O_7}. Since the anchor of c_6 is a member of c_4, then c_6 is the sub-concept of c_4 and c_4 is the super-concept of c_6—using definition 2.

TABLE I. A SET OF OBJECTS.

Object	A1	A2	A3	A4	A5
O_1	-1	1	2	3	5
O_2	2	-2	4	6	2
O_3	5	-3	2	3	3
O_4	6	3	5	1	8
O_5	7	4	3	-2	10
O_6	5	4	2	2	7
O_7	6	4	2	2	8

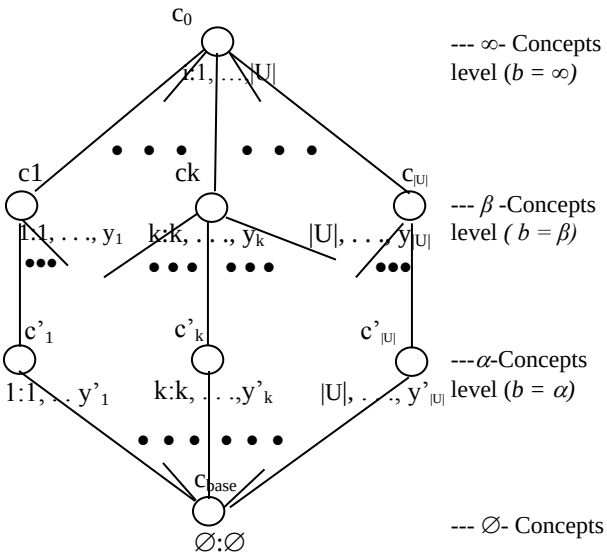


Figure 1. A lattice of concepts for the universe of object U and the two values of b = β and b = α.

The constant b is in the range of [0, +∞]. Let us build the lattice for those b-concepts of U generated for two different values of b (α and β, where α < β)—using definition 3. The lattice has four levels. The first level contains apex, concept c_0. The second level includes all the concepts for which b = β, (β-concepts). At the third level, all the concepts for which b = α (α-concepts) are included. The last level contains the base. At each level, there are |U|, not necessarily distinct, concepts, such that every object of U serves as the anchor of one concept; see Figure 1.

Each concept has a concept name, c_i, anchor, G(c_i), and members. The notation G(c_i): O_i, . . ., O_y is used to display the anchor and members of the concept c_i. The concept at the apex, c_0, includes all the objects of the universe as its members (b = ∞). Thus, any member can be designated as the anchor of the concept of c_0. Therefore, G(c_0) = O_i. The concept at the base includes no objects.

Reader needs to keep in mind that because of the huge range of values for constant b, the resulting lattice may have infinite number of levels. Building such an extremely large lattice is unnecessary because: (i) once the value of b reaches to the point that forces all the objects of U into one concept, then all the levels of the lattice beyond that b value have exactly the same concepts and (ii) turning the entire

 ISBN: 978-1-61208-364-3

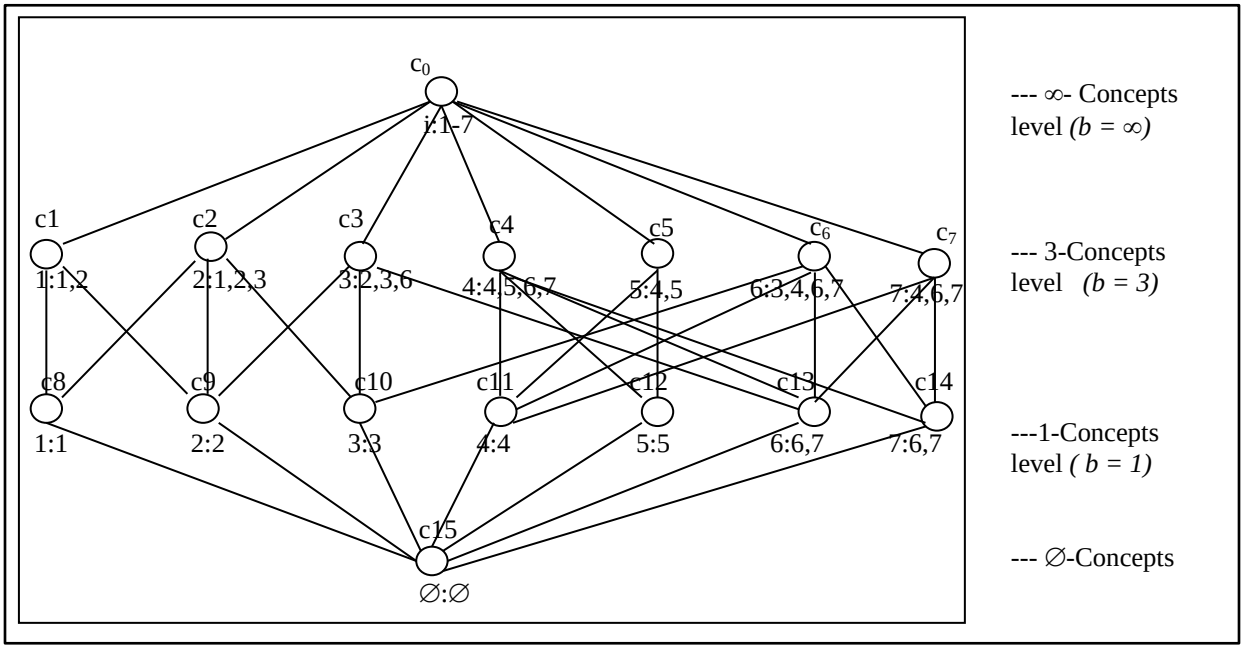


Figure 2. The lattice for the objects of Table I, α=1 and β=3.

TABLE II. THE γ-EXPANSION OF C4: (a) C4 AND ITS MEMBERS AND (b) THE COMPLETE γ-EXPANSION OF C4.

Concept	Members	α-concepts of Member	Cover of α-concept
c_4	O_4	c_{11}	c_4, c_5, c_6, c_7
	O_5	c_{12}	c_4, c_5
	O_6	c_{13}	c_3, c_4, c_6, c_7
	O_7	c_{14}	c_4, c_6, c_7

(a)

Concept	Members	α-concepts of Member	Cover of α-concept
c_4	O_4	c_{11}	c_4, c_5, c_6, c_7
	O_5	c_{12}	c_4, c_5
	O_6	c_{13}	c_3, c_4, c_6, c_7
	O_7	c_{14}	c_4, c_6, c_7
	O_3	c_{10}	c_2, c_3, c_6

(b)

Using the same analogy, the partial expansion for c_{12} and c_{14} do not change c_4 either. However, the $c_4 \cup$ cover(C_{13}) changes c_4 by adding a new object O_3 to c_4, Table 2.b. The new c_4 is the total γ-expansion of the original c_4. Because of the object O_3, only one new α-concept of c_{10} is added to the list of α-concepts of c_4 which has the cover of $\{c_2, c_3, c_6\}$.

objects of U into one concept clearly makes the conceptualization process of U a moot one.

As an example, the lattice for the set of objects of Table 1, for b = α = 1 and b = β = 3 is shown in Figure 2.

A. *Super-Concept Expansion*

A sub-concept within a lattice of concepts may have several super-concepts that are collectively referred to as the *cover* of the sub-concept. For example, the cover for the α-concept of c_{11} in Figure 2 is: cover(c_{11}) = $\{c_4, c_5, c_6, c_7\}$

Definition 4: Let γ be a set of conditions that is used to discriminate against the β-concepts. Let also c_j be a β-concept satisfying γ. In addition, let c_j have q sub-concepts of (α_1-concept, . . , α_q-concept). Furthermore, let c_j be expanded by all the covers of one of its sub-concepts (α_p-concept), $c_j = c_j \cup$ Cover(α_p-concept). If the expanded c_j also satisfies γ, then the new c_j is the *partial γ-expansion* of c_i over α_p-concept. The concept c_j is *totally γ-expanded* over its members when all the possible partial γ-expansions of c_j are done. A totally γ-expanded c_j may have a new set of sub-concepts (α-concepts). The concept c_j reaches its *complete γ-expansion (CGE)* when it cannot have any more partial expansions.

As an example, let us assume that concept c_4 satisfies the condition set of γ. (The condition set of γ is explained in detail in the next subsection.) The c_4 includes objects O_4, O_5, O_6, and O_7. The α-concepts of c_4 along with their covers are shown in Table 2.a. The $c_4 \cup$ cover(c_{11}) is the first partial γ-expansion of c_4. Let us assume that the expanded c_4 also satisfies γ. The partial expansion does not add to the members of c_4. That is, the cover of c_{11} includes concepts c_4, c_5, c_6, and c_7 that collectively contain objects of O_4, O_5, O_6, and O_7 which is the same as the objects in c_4 prior to expansion.

Further expansion of c_4 for creation of its CGE starts with the $c_4 \cup \text{cover}(c_{10})$. This partial expansion changes c4 by adding object O_2 to c_4. Let us assume that the expansion does not satisfy γ. As a result, the CGE of c_4 includes the objects of O_3, O_4, O_5, O_6, and O_7.

After the CGE of a β–concept is obtained the following two steps take place:

a. Removing those concepts along with all the dangling edges from the lattice that their anchors are found in the complete γ-expansion of the β-concept.

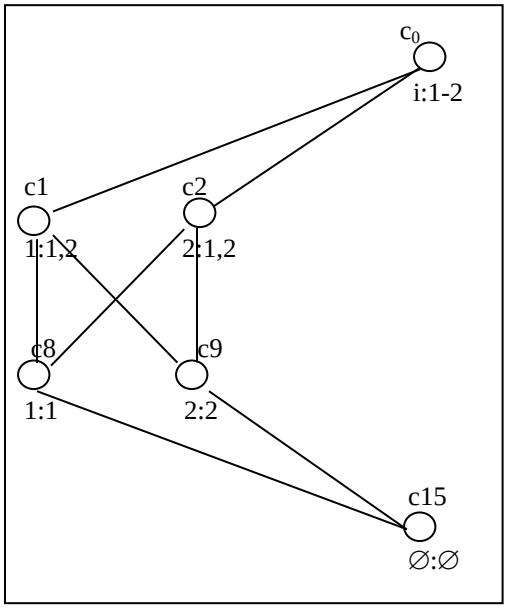


Figure 3. The adjusted lattice after removing the CGE of c4.

b. Deriving a new object as the anchor of the β–concept. The attribute A_i of the new object has the value equal to the average of all the A_i values of its members. Since the expanded β–concept is different from its original, calculation of the new anchor is necessary.

The result of lattice reduction using the CGE of c_4 is shown in Figure 3. Considering Table I, the new anchor for the CGE of c_4 has the attribute values of: (5.8, 2.4, 2.8, 1.2, 7.2).

Following the same process, another complete-γ-expansion is produced, from Figure 3, that belongs to c_1. The reduction of the lattice using the CGE of c_1, causes the deletion of the entire lattice. This indicates the end of the process of the concept expansion. The attribute values for the anchor of the CGE of c_1 are: (0.5, -05, 3, 4.5, 3.5).

B. Feature Extraction

The driven forces behind the feature extraction from the universe of objects, U, are the conceptualization of U, and γ-condition set. The former one contributes to the size, depth and cost of building the lattice and the later one contributes to the expansion of the qualified concepts.

The lattice size and depth are influenced by the number of objects in U the values for b, respectively. Since b values are too many, the building of the lattice is prohibitively expensive. Thus, building of only a two-level lattice (excluding apex and base levels) is preferred. The small values for b are more attractive because they relax the forced conceptualization of objects in U. As a result, the concepts are more organic and so their internal characteristic (features).

The extracted features are influenced by the γ-condition set. Let us assume that we are after extracting features that can be used for prediction (a set of prediction rules). To complete such extractions, a γ-condition set is introduced for the purpose of discriminating against concepts such that the concepts with a weak set of features be filtered. To explain it further, a decision attribute is assigned to every object in U. This attribute does not participate in conceptualization of U. The prediction of the value of the decision attribute for a set of new objects is the ultimate goal. If the minimum 2/3 of the members of a concept have the same decision as the anchor of the concept, the concept satisfies the γ-condition set—(i.e., the concept is a qualified one). Therefore, the concept's members collectively own a set of internal characteristics that support the decision of the concept's anchor with the strength of 2/3 out of one.

During the complete γ-expansion of a concept, the γ-condition set filters the unwanted partial γ-expansions to protect the strength of the internal characteristics of the concept. Since the anchor of a complete γ-expansion of a concept represents the entire members of the concept, so it represents their internal characteristics of them too.

The extracted features from CGE of a concept in form of prediction rules are referred to as the *anchor prediction (AP) rules* and presented in the production rules format. The AP rules extracted from the objects of Table I, are shown below using the two new anchors of the CGEs for concepts of c_4 and c_1. Let us assume that the decision for the anchors of c_4 and c_1 are d_1 and d_2, respectively. The AP rules are:

$(A_1 = 5.8, A_2 = 2.4, A3 = 2.8, A_4 = 1.2, A_5 = 7.2) \rightarrow d_1$
$(A_1 = 0.5, A_2 = -0.5, A_3 = 3, A4 = 4.5, A_5 = 3.5) \rightarrow d_2$

C. Crystallization of the Extracted Features

Let us assume that the extracted set of AP rules is applied on a given test set, TE, and the quality of the prediction outcome is measured, Q. The crystallization of the AP rules is started by applying first the horizontal and then the vertical reductions. The details for both reductions are covered in the following two subsections.

1. *Vertical Reduction of the AP Rules:* The goal of this reduction is to remove as many rules as possible from the set of AP rules without lowering the prediction ability of the set. To meet the goal, one rule is randomly removed from the set, the new set of AP rules is applied against the

 ISBN: 978-1-61208-364-3

test set of TE, and the quality of the prediction is measured, Q'. If Q' ≥ Q then: (a) the new set of AP rules replaces its predecessor and (b) Q' becomes the new Q. This process continues until no more rules can be removed from the set. There is a chance that none of the rules can be removed. This means the prediction rules cannot be vertically reduced.

2. *Horizontal Reduction of the AP Rules:* The goal of this reduction is to make the list of attributes for the entire AP rule set as short as possible. To meet the goal, one attribute, A_i, is kept in the AP rule set and the rest of the attributes are removed. The new AP rules are applied on the TE and the quality of the prediction is measured. This process is repeated for every attribute and at the end the attribute with the highest prediction quality, Q', is the winner. If Q' ≥ Q, then the winner attribute is the smallest subset of attributes representing the horizontal reduction of the set. If this is not the case, then another attribute is added to the winner attribute and the quality of prediction is checked for the pair. This process is repeated for every possible pair and at the end the pair with the highest prediction quality, Q", is the winner. If Q" ≥ Q, then the winner pair is the smallest subset representing the horizontal reduction of the set. If the condition of Q" ≥ Q is not true, the winner pair grows to three attributes and this process continues until the minimum subset of the attributes is found with the prediction quality, at least, as good as Q. There is a chance that such subset cannot be found. This means the prediction rules cannot be horizontally reduced.

IV. EMPIRICAL RESULTS

An object set describing the properties of 1018 chemical agents was provided by a team of bio-chemists. Each chemical agent had eight attributes. One of the attributes is the decision and indicates whether the agent is carcinogen or not. The ten percent of the objects with decision zero along with the ten percent of the objects with decision one are randomly selected to make the test set. One may create 10 different test sets randomly such that the test sets do not have any objects in common. Let us consider one of the test sets. After creating the test set the remaining records are used as a training set. However, the training set must include equal number of objects for both decisions and include the largest number of objects as possible. As a result, we created 10 pairs of training and test sets such that the test sets did not have any objects in common.

For each pair, (i) the conceptualization of the training set was done for $b = \alpha = 0$ and $b = \beta = 1$ and (ii) the AP rule set was generated and used to predict the decision for the objects of the test set and the quality of predictions was measured. We compared the prediction performance of the AP rule set, and the reduced AP rule set. The comparisons are shown in Table 3. All the training sets had both horizontal and vertical reductions. We have used both sequence of horizontal-vertical reductions and vertical-horizontal reductions of the AP rule set and the former one produced better prediction results and they are the ones shown in Table 3.

TABLE III. THE COMPARISON OF THE PREDICTION POWER OF THE ID3, AP RULE SET, AND REDUCED AP RULE SET FOR 10 PAIRS OF TRAINING AND TEST SETS.

Pair	Method	% correct prediction	% False (+)	% False (-)	% Not predicted
1	ID3	66.7	24	0	9.6
	AP Rules	66.7	28.5	4.8	0
	Crystalized AP Rules	90.5	2.4	7	0
2	ID3	76.2	19	0	4.8
	AP Rules	76.2	17	7	0
	Crystalized AP Rules	88.1	2.4	9.5	0
3	ID3	66.7	26	1	7
	AP Rules	81.0	9.6	4.8	4.8
	Crystalized AP Rules	92.9	0	7	0
4	ID3	71.4	17	1	12
	AP Rules	78.6	9.6	12	0
	Crystalized AP Rules	88.1	0	12	0
5	ID3	71.4	21	1	7
	AP Rules	61.9	24	12	2.4
	Crystalized AP Rules	83.4	0	17	0
6	ID3	76.2	12	2	12
	AP Rules	69	9.6	14	7
	Crystalized AP Rules	88.1	2.4	7	0
7	ID3	69	14	0	17
	AP Rules	69	12	12	0
	Crystalized AP Rules	90.5	4.8	4.8	0
8	ID3	66.7	24	0	9.6
	AP Rules	66.7	17	17	0
	Crystalized AP Rules	88.1	2.4	9.5	0
9	ID3	64.3	24	0	24
	AP Rules	66.7	17	9.6	7
	Crystalized AP Rules	85.7	0	14.3	0
10	ID3	73.8	12	2	14
	AP Rules	69	14	17	0
	Crystalized AP Rules	88.1	2.4	9.5	0
Avg	**ID3**	**70.2**	**12**	**2**	**11.7**
	AP Rules	**71.2**	**15.8**	**11.02**	**2.12**
	Crystalized AP Rules	**88.4**	**1.7**	**9.8**	**0**

One may raise the question of how good is the performance of AP rules in reference to other algorithms used for prediction. One of the well-known algorithms used for classification and prediction is ID3. We use ID3 to extract rules from each training set and measure the quality of the extracted rules in predicting the decision for the objects of the corresponding test set. The results are also shown in Table 3.

 ISBN: 978-1-61208-364-3

V. CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

The results presented in Table 3 revealed that the AP rule set performs as good as the ID3 algorithm. That means the proposed concept analysis has a high potential to serve as a prediction tool. The Crystalized AP rule set has a superior performance in compare with both ID3 and AP rule set.

We have observed that the anchor of a concept represents all of the concept members. Based on this observation, one may replace a large set of objects with a much smaller set of their anchors. Reduction of the size of the universe of objects may be useful in reducing the size of a Big Data. By doing so, one may be in a better position to analyze a very large object set. As part of the future research, this investigation is in progress.

Through the entire conceptualization process, we assumed that all the attributes of an object have the same strengths. This assumption is quite true for some universe of the objects such as the one used for obtaining the results shown in Table 3. However, the assumption is false for some other universe of objects. As another part of the future research, we revisit the creation of the b-concepts using varying strengths for the attributes of the objects.

REFERENCES

[1] R. Wille, "Restructuring Lattice Theory: An Approach based on Hierarchies of Concepts, Ordered Sets," Reidel Dordecht-Boston Publisher, 1982.

[2] B. Ganter, B. and R. Wille, "Formal Concept Analysis: Mathematical Foundations," Springer-Verlag Publishing, Berlin, 1999.

[3] J. S. Deogun, V. V. Raghavan, and H. Sever, "Association Mining and Formal Concept Analysis," Proc. the Joint Conference in Information Science, Research Triangle Park, NC, 1998, pp. 335-338.

[4] R. Hashemi, L. LeBlanc, and T. Kobayashi, "Formal Concept Analysis in Investigation of Normal Accidents," the International Journal of General Systems, vol 33, no. 5, October 2004, pp. 469-484.

[5] J. R. Quinlan, "Induction of Decision Trees," Machine Learning", vol. 1, no. 1, 1986, pp. 81-106.

[6] V. Evans, "Lexical concepts, cognitive models and meaning-construction," Journal of Cognitive Linguistics, vol. 17, 2006, pp. 491-534.

[7] L. Barsalou, "Continuity of the conceptual system across species," Journal of Trends in Cognitive Sciences, vol. 9, 2005, pp. 309-311.

[8] R. Hashemi, L. Le Blanc, A. Bahrami, M. Bahar, and B. Traywick, "Association Analysis of the Alumni Giving: A Formal Concept Analysis," the International Journal of Intelligent Information Technologies, vol. 5, no. 2, April-June, 2009, pp. 17-32.

[9] R. Hashemi, L. Sears and A. Bahrami, "An Android Based Medication Reminder System: A Concept Analysis Approach," Proc, the 19th International Conference on Conceptual Structures (ICCS'11), Sponsored and proceedings published by Springer-Verlag as Lecture Notes in Artificial Intelligence, (LNAI 6828) Derby, UK, July 2011, pp. 315-322.

[10] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," The ACM Journal of Machine Learning Research, vol. 5, Dec. 2004, pp.1205-1224.

[11] Y. Lei and M. Luo, "Rough concept lattices and domains," Annals of Pure and Applied Logic, Joint Workshop Domains VIII: Computability over Continuous Data Types, Y. L. Ershov, K. Keimel, U. Kohlenbach and A. Morozov (eds.), Novosibirsk, Rusia, Published by Elsevier, vol. 159, no. 3, Sept. 2009, pp. 333-340.

[12] R. Dietz and S. Moruzzi (editors), "Cuts and clouds. Vagueness, Its Nature, and Its Logic." Oxford University Press, 2009.

[13] A. Markusen, "Fuzzy Concepts, Scanty Evidence, and Policy Distance: The Case for Rigor and Policy Relevance in Critical Regional Studies," In: Regional Studies, vol. 37, no. 6-7, 2003, pp. 701-717.

[14] D. Metzler and W. B. Croft, "Latent Concept Expansion Using Markov Random Fields", SIGIR'07, Amsterdam, The Netherlands, July 2007, pp. 311-318.

[15] Y. Qui and H. P. Frei, "Concept Based Query Expansion", Proc. of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'93), 1993, pp. 160-169.

 ISBN: 978-1-61208-364-3

A Multi-Factor HMM-based Forecasting Model for Fuzzy Time Series

Hui-Chi Chuang, Wen-Shin Chang, Sheng-Tun Li
Department of Industrial and Information Management
Institute of Information Management
National Cheng Kung University
No.1, University Road, Tainan City, Taiwan (R.O.C.)
e-mails:{r78021017, r36024037, stli}@mail.ncku.edu.tw

***Abstract*—In our daily life, people are often using forecasting techniques to predict weather, stock, economy and even some important Key Performance Indicator (KPI), and so forth. Therefore, forecasting methods have recently received increasing attention. In the last years, many researchers used fuzzy time series methods for forecasting because of their capability of dealing with vague data. The followers enhanced their study and proposed a stochastic hidden Markov model, which considers two factors. However, in forecasting problems, an event can be affected by many factors; if we consider more factors for prediction, we usually can get better forecasting results. In this paper, we present a multi-factor HMM-based forecasting model, which is enhanced by a stochastic hidden Markov model, and utilizes more factors to predict the future trend of data and get better forecasting accuracy rate.**

Keywords-fuzzy time series; forecasting; hidden Markov model (HMM).

I. INTRODUCTION

In the information explosion era, forecasting is a useful methodology for enterprises or governments to predict future trends. The more precise the forecasting result, the more appropriate the behavior conducted by managers. In general, many data are present with crisp value, but others are vague and ambiguous instead, such as stock monitoring indicators, signals, and so on. With the purpose of forecasting with vague data, Song and Chissom [9] first proposed the fuzzy time series and suggested seven steps for forecasting. Therefore, numerous fuzzy time series forecasting models have been proposed after them.

The seven steps proposed by Song and Chissom [9] are: (1) define the universe of discourse; (2) partition the universe into several intervals; (3) define fuzzy sets on the universe; (4) fuzzify the historical data; (5) establish fuzzy relations; (6) calculate the forecasted outputs; (7) defuzzification. The literature used Alabama University enrollment data to carry out one factor time relational, invariant fuzzy time series forecast model. Sullivan and Woodall [4] improved the results by using Markov's matrix- based probability statistics method [9] to establish one-factor one-order time invariant forecast model.

The above introduced methods implemented one factor to forecast, but in realistic circumstances, there is more than one factor that affects the forecasted data, such as the temperature and cloud density. Therefore, based on this idea, Lee, Wang, Chen, and Leu [6] proposed a two-factor high-order forecast model, which many people debate problematic in deciding the high-order. Later, Li and Cheng [13] proposed the deterministic fuzzy time series model that uses the concept of state transition diagram to backtrack to the most stable state. Wang and Chen [8] proposed two-factor fuzzy time series model and used automatic clustering techniques to partition intervals. Joshi and Kumar [1] first extended the idea of fuzzy sets into Intuitionistic Fuzzy Sets (IFS) and used it to construction a method for determining the membership degree.

Although fuzzy time series has been developed for decades, there are still some problems. The matrix computing for two-factor fuzzy time series models are too complex. Furthermore, the process of fuzzy time series forecasting is only concerned with whether or not the fuzzy rule is adopted, while ignoring the importance of frequency. Therefore, Sullivan and Woodall [4] used the Markov model to improve the traditional fuzzy time series model. But traditional HMM is unable to solve the two-factor problem. If we can analyze more factors, we can get higher accuracy and fully utilize all the available information. In our study, we enhance the study of Li and Cheng [12] and expand the model for forecasting based on multiple factors.

The fuzzy time series methods are good methodologies, which are suited for data composed of linguistic values, but the fuzzy relationship ignores the importance of the frequency. Hidden Markov models are powerful probability models which use categorical data include linguistic labels, and allow handling of two-factor forecasting problem. But in real world situations, multiple factor data tends to appear, rendering this model ineffective in making accurate assessments.

In this study, we want to combine the benefit of both methodologies, so that we can deal with the problem with many factors and fully utilize the feature of fuzzy theory to get higher accuracy result. There are five sections in this study, which is organized as follows. In Section I, we introduce the background and motivation. In Section Ⅱ, we introduce the basic definition of fuzzy time series and hidden Markov model. Section Ⅲ presents the model development. We propose a fuzzy time series model to deal with the multi-factor forecasting problem and we use HMM to build the fuzzy relationship matrix. In Section Ⅳ, we present the experiment result. We demonstrate the proposed model step by step and make a comparison with other existing methods. In Section Ⅴ, we present the conclusions by discussing the results as explained in Section Ⅳ and proposes future work expecting to improve the proposed method.

 ISBN: 978-1-61208-364-3

II. LITERATURE REVIEW

A. Fuzzy Time Series

Although there are many statistical methods that can be used to solve the problem of time series, it is impossible to resolve this problem when the historical data is linguistic value. It was not until 1993, when Song and Chissom [9] first proposed the fuzzy time series, that this problem has a solution. Because of its easiness to use and comprehend, there are many scholars who have dedicated their time and energy to the field of fuzzy time series making it more complete and accurate. The following is the basic definition of fuzzy time series.

1) Definition 1: Fuzzy Time Series

Let $\{x_t \in R, t = 1,2,...n\}$ be a fuzzy time series, U is the universe of discourse. Let $\{L_i(t), i = 1,2,...,l\}$ be the ordered linguistic variables. Each X_t in the universe of discourse can be the denoted as follows:

$$F(t) = \frac{\mu_1(X_t)}{L_1} + \frac{\mu_2(X_t)}{L_2} + ... + \frac{\mu_l(X_t)}{L_l} \quad (1)$$

where $\mu_k \in [0,1]$ and $\sum_k^l \mu_k(x_t) = 1, \ \forall t = 1,2,...,n \cdot \mu_k(X_t)$ is the membership value of the linguistic variable L_k. After transforming, we can say $F(t)$ is the fuzzy time series of X_t.

2) Definition 2: One-order Fuzzy Relation Equation

Let $F(t)$ be a fuzzy time series, the relationship between $F(t)$ and $F(t\text{-}1)$ is denoted as below:

$$F(t) = F(t-1) \circ R(t, t-1) \quad (2)$$

"$\circ$" is a composition operator, and where $R(t,t\text{-}1)$ which is composed of R_{ij} is a one-order fuzzy relation. The relationship shows below:

$$R(t, t-1) = \cup_{ij} R_{ij}(t, t-1) \quad (3)$$

where $R_{ij}(t, t-1)$ is the fuzzy relation between $F_j(t)$ and $F_i(t\text{-}1)$.

3) Definition 3: One-order Fuzzy Logic Relationship

In order to reduce the complexity of matrix computation, Chen [11] proposed Fuzzy Logical Relationship (FLR) combined with simple arithmetic operation to replace matrix computation. If $F(t\text{-}1)$ is transformed into linguistic variable A_i, $F(t)$_is A_j, then the relationship between A_i and A_j can be shown below:

$$F(t-1) \to F(t) \text{ or } A_i \to A_j \quad (4)$$

The equation on the left side of the arrow is called the Left Hand Side (LHS); the right side is called Right Hand Side (RHS). After fuzzifying the historical data and establishing the fuzzy logical relationship, then we can group the entire data together to establish a logical relationship group. The following shows the fuzzy logical relationship group:

$$\begin{aligned} A_{i1} &\to A_{j1}, A_{j2}, ..., A_{jn} \\ A_{i2} &\to A_{j1}, A_{j2}, ..., A_{jn} \\ &\vdots \\ A_{ik} &\to A_{j1}, A_{j2}, ..., A_{jn} \end{aligned} \quad (5)$$

B. Forecasting Model of Fuzzy Time Series

Song and Chissom [9] first proposed a complete fuzzy time series forecasting model and divided it into seven steps. Later, many scholars have continuously revised this framework in order to get better forecasting accuracy.

1) Define Universe of Discourse U & Partition the universe

Song and Chissom [9] defined the universe of discourse U as follows:

$$U = [D_{\min} - D_1, D_{\max} + D_2] \quad (6)$$

where D_{min} and D_{max} are the minimum and the maximum in the training date set and D_1 and D_2 are the two proper positive integers decided by the analyst. After defining U, we partition the universe into several intervals. One is equal length interval, and another is unequal length interval.

2) Define Fuzzy Set and Linguistic Values

After fuzzyfying historical data we then begin to decide linguistic values like rare, few, plenty and so on. Each interval A_i (i=1,…,n) will have its own membership to get linguistic value u_k (k=1,…,n), which represents the degree of belonging to each interval of each. They used their own experience to formulate the membership degree of each element in each fuzzy set; the example is shown in (7).

$$\begin{aligned} A_1 &= \{u_1/1, u_2/0.5, u_3/0, u_4/0, u_5/0, u_6/0, u_7/0\} \\ A_2 &= \{u_1/0.5, u_2/1, u_3/0.5, u_4/0, u_5/0, u_6/0, u_7/0\} \\ A_3 &= \{u_1/0, u_2/0.5, u_3/1, u_4/0.5, u_5/0, u_6/0, u_7/0\} \\ A_4 &= \{u_1/0, u_2/0, u_3/0.5, u_4/1, u_5/0.5, u_6/0, u_7/0\} \\ A_5 &= \{u_1/0, u_2/0, u_3/0, u_4/0.5, u_5/1, u_6/0.5, u_7/0\} \\ A_6 &= \{u_1/0, u_2/0, u_3/0, u_4/0, u_5/0.5, u_6/1, u_7/0.5\} \\ A_7 &= \{u_1/0, u_2/0, u_3/0, u_4/0, u_5/0, u_6/0.5, u_7/1\} \end{aligned} \quad (7)$$

3) Fuzzify Historical Data

If the historical data is not ambiguous, we need to fuzzify it first, for example hot, cold and linguistic values. The process of fuzzification usually uses triangular fuzzy sets. The triangular fuzzy set of equal length intervals are represented in Figure 1, the unequal length intervals are represented in Figure 2.

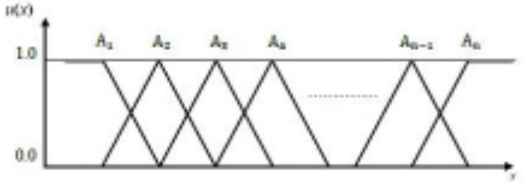


Figure 1. Linguistic values and triangular fuzzy set of equal length intervals.

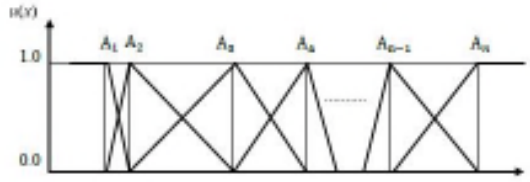


Figure 2. Linguistic values and triangular fuzzy set of unequal length intervals.

4) Establish Fuzzy Relation

Building up fuzzy relations is the most important part in the fuzzy time series forecasting model. Among many scholars, Song and Chissom proposed the time-variant and time-invariant fuzzy relation matrix.

Let us introduce Song and Chissom's method. First, we need to observe the fuzzy relation R_j by historical data. Second, find the relation matrix R which is composed of R_j. By using following equation, we can get the relation matrix R, where R_j comes from first step and N means the total number of fuzzy relations.

$$\mathbf{R} = \bigcup_{j=1}^{N} R_j \tag{8}$$

Finally, we can establish the forecasting model. There are many models for forecasting models and the following formula is one of composition to forecasting data, where "$\circ$"is max-min operator.

$$A_i = A_{i-1} \circ R \tag{9}$$

Then, we can get the forecasting result.

5) Forecasting and Defuzzification

After establishing fuzzy relation matrix or fuzzy logical relationship, we begin to forecast linguistic value and then defuzzify it into precise one. However, there are some scholars who replace the process of forecasting linguistic value with a series of arithmetic operations, directly restoring the value into precise. For example, Wong, Bai, and Chu [14] proposed adaptive time-variant fuzzy time series; they used dynamic analysis window combined with series of heuristic rules to calculate the forecasting result.

The Proposed Defuzzification Method of Song and Chissom is summarized as follows: First, we standardized the calculated membership degree of linguistic value, and then we followed the three criteria which are listed below to defuzzify. (1) If there is only one maximum membership degree, then, we use the corresponding midpoint of the interval to be forecast value. (2) If the membership of an output has two or more consecutive maximums, then select the midpoint of the corresponding conjunct intervals as the forecasted value. (3) Otherwise, standardize the fuzzy output and use the midpoint of each interval to calculate the centroid of fuzzy set as the forecast value.

C. Markov Process

Markov property, named after the Russian mathematician Andrey Markov, refers to the memoryless property of a stochastic process. A stochastic process has the Markov property if the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state, not on the sequence of events that preceded it.

D. Hidden Markov Model (HMM)

Under realistic situation, there are usually multiple factors that influence the behavior and outcome of an event. For example, when we are trying to predict the temperature today, we may want to know the density of cloud, the amount of rainfall and also the weather yesterday. With the observation today and some previous information, we can predict today's weather. We might obtain the better forecast by combining the information of previous data and the present observation. HMM [15] is exactly the tool that can deal with two factors forecasting problem.

HMM is a statistical model to deal with symbols or signal sequences that are assumed to be a Markov process and have found a lot of applications in many areas like speech processing, weather, economy, population growth, stocks, etc.

The hidden Markov process is based on two important and essential assumptions: (1) The next state is dependent only upon the current state. (2) Each state-transition probability does not vary in time, i.e., it is a time-invariant model.

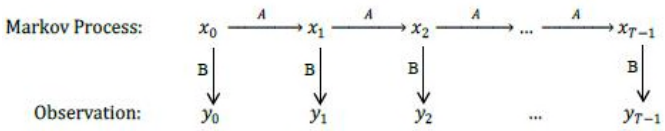


Figure 3. Hidden Markov Model.

The next three matrices of probability describe the common calculations that we would like to be able to perform on a HMM.

(1) Initial state vector $\pi = \{\pi_i\}$

$$\pi_i = \mathrm{P}(x_0 = s_i)$$

(2) State transition matrix A

$$A = \{a_{ij}\} = \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0,N-1} \\ a_{10} & a_{11} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{N-1,0} & \cdots & \cdots & a_{N-1,N-1} \end{bmatrix} \tag{10}$$

$$a_{ij} = P(state\ j\ at\ t+1 | state\ i\ at\ t)$$

(3) Confusion matrix B

$$B = \{b_{ij}\} = \begin{bmatrix} b_{00} & 01 & \dots & b_{0,M-1} \\ b_{10} & b_{11} & & \vdots \\ \vdots & & \ddots & \vdots \\ b_{N-1,0} & \cdots & \cdots & a_{N-1,M-1} \end{bmatrix} \quad (11)$$

$$b_{ij} = P(observation\ j\ at\ t | state\ i\ at\ t)$$

Therefore, we use $\lambda = (A, B, \pi)$ to present the overall HMM model.

III. MODEL DEVELOPMENT

There are many studies of fuzzy time series, but the complexity of matrix computing still remains in multiple factors forecasting. Even though we can use HMM method to get the result faster and better, traditional HMM just can solve two-factors problem only. If we can analyze more factors, we can get higher accuracy and fully utilize all the information we got.

In this study, we want to preprocess data with fuzzification and use the transition of hidden Markov model to forecast outcome.

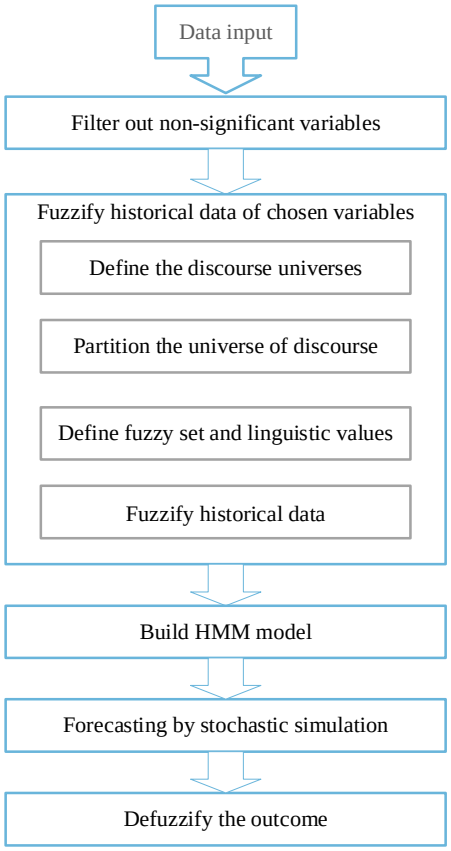


Figure 4. The framework of this study.

A. *Fuzzify historical data of chosen variables*

We follow the methodology of fuzzy time series forecasting to fuzzify the historical data. In this study, we conduct equal length interval method to partition universe of discourse into *n* intervals.

First, we need to define the discourse universes of historical data. Let $U^S, U^{O_1}, U^{O_2} \dots U^{O_K}$ be the discourse universes of hidden state and observation variables respectively, and assume that the number of observation variables is *k*.

In general, the discourse universes are defined as follows:

$$\text{Hidden variable: } U^S = [D^S_{min} - D^S_1, D^S_{max} + D^S_2]$$
$$\text{Observation variable 1: } U^{O_1} = [D^{O_1}_{min} - D^{O_1}_1, D^{O_1}_{max} + D^{O_1}_2]$$
$$\vdots$$
$$\text{Observation variable K: } U^{O_K} = [D^{O_K}_{min} - D^{O_K}_1, D^{O_K}_{max} + D^{O_K}_2]$$

where $D^S_{min}, D^S_{max}, D^{O_1}_{min}, D^{O_1}_{max}, D^{O_K}_{min}$ and $D^{O_K}_{max}$ are the respective minimal and maximal values of historical data of hidden and observation variables, and $D^S_1, D^S_2, D^{V_1}_1, D^{V_1}_2, D^{V_K}_1$ and $D^{O_K}_2$ are proper positive values decided by the analyst.

Second, the discourse universes are then partitioned into several equal lengthy intervals. Let us use $u_1, u_2, \dots, u_n$ for each interval. Therefore, we have fuzzy set $A_1, A_2, \dots, A_n$. When a value approaches the center of linguistic value A_i, it means the greater degree belongs to the linguistic value A_i, thus, the membership degree is closer to 1; on the contrary, if it is closer to the two bounds of A_i, the membership degree which belongs to A_i will be closer to 0.

In traditional study, each value has two linguistic values and corresponding membership degrees in most case, but the past literatures ignored the smaller one. However, even though our study is smaller, we think that it still has some valuable information. Therefore, we use triangle membership function to fuzzify historical data, and the fuzzy sets $A_1, A_2, \dots, A_n$ are defined as seen in Figure 5:

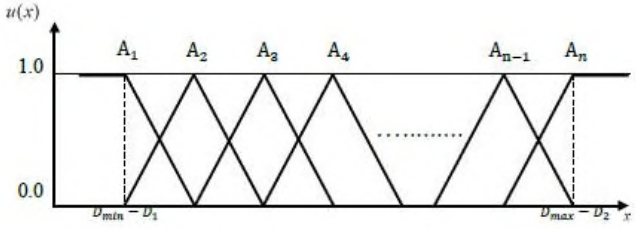


Figure 5. The fuzzy sets.

Finally, let us fuzzify the historical data and find the degree of each data belongs to each A_i $(i = 1,2,\dots,n)$. Then, we assume that if the maximum membership of that data is under A_k , then we treat this data as A_k .

B. *Build HMM model*

The objective of a multi-factor hidden Markov forecasting model is to estimate the probability of hidden state with given observations, so that we can predict the sequence of state that best explains the observed data. The following notations of HMM will be used in this paper:

$N = number\ of\ state\ in\ the\ model$

$M_i = number\ of\ state\ for\ variable\ i, i = 0,1,\dots,K-1$

K = number of variables
T = lengh of sequence
$S = \{s_0, s_1, \ldots, s_{N-1}\}$ = distinct states of the Markov process
$O = \{o^0, o^1, \ldots, o^{K-1}\}$ = distinct observation
$o^i = \{o_0^i, o_1^i, \ldots, o_{M_i}^i\}$ = distinct observation for variable i, $i = 0,1, \ldots, K-1$
$X = \{x_0, x_1, \ldots, x_{T-1}\}$ = state sequences with lengh T
$Y = \{y_0, y_1, \ldots, y_{T-1}\}$ = observation sequences with lengh T
$y_t = \{y_t^0, y_t^1, \ldots, y_t^{K-1}\}$ = all observations for each variable in time t

$t = 0,1, \ldots, T-1$

Markov process:	x_o	$\xrightarrow{A}$	x_1	$\xrightarrow{A}$	x_2	$\xrightarrow{A} \ldots \xrightarrow{A}$	x_{T-1}
Observations 1 (B^0):	y_0^0		y_1^0		y_2^0	...	y_{T-1}^0
Observations 2 (B^1):	y_0^1		y_1^1		y_2^1	...	y_{T-1}^1
	⋮		⋮		⋮		⋮
Observations K (B^{K-1}):	y_0^{K-1}		y_1^{K-1}		y_2^{K-1}	...	y_{T-1}^{K-1}

Figure 6. A Multi-factor Hidden Markov Model.

In our study, we expand the research to multiple variables HMM and there are three important and essential assumptions of our proposed model: (1) The next state is dependent only upon the current state. (2) Each state-transition probability does not vary in time, i.e., it is a time-invariant model. (3) The observations are independent to each other.

The first and second assumptions come from the assumptions of HMM. The need for a third assumption is because our proposed method deals with multiple factors; in order to simplify the process, we ignore the impact of correlation between observations.

Therefore, the multiple observations HMM can be characterized by the following matrices:

$$\pi = \{\pi_i\}, \quad where\ \pi_i = P(x_0 = s_i)$$
$$A = \{a_{ij}\}, \quad where\ a_{ij} = P(x_t = s_j | x_{t-1} = s_i)$$
$$B^0 = \{b_{i,j}^0\}, \quad where\ b_{i,j}^1 = P(y_t^0 = o_j^0 | x_t = s_i)$$
$$\vdots$$
$$B^{K-1} = \{b_{i,j}^{K-1}\}, \quad where\ b_{i,j}^{K-1} = P(y_t^{K-1} = o_j^{K-1} | x_t = s_i)$$

π is a vector with the probability of initial state. A is the state-transaction matrix which provides information about the relation of two contiguous hidden states. B^v is the confusion matrix which is the relation between observation v and hidden state.

Let us defined those parameters which are estimated by relative frequencies:

First, the initial state $\pi = \{\pi_i\}$ is a $1 \times n$ matrix and is defined as

$$\pi_i = \Pr(x_0 = s_i) = \frac{Count(x_0 = s_i)}{N_1}$$
$$N_1 = \sum_{i=0}^{N-1} Count(x_0 = s_i) \quad (12)$$

where $Count(x_0 = s_i)$ is the number of initial state s_i in the data set, and N_1 is the sum of initial state.

Second, the state transition matrix $A = \{a_{ij}\}$ is a $n \times n$ matrix and is defined as

$$a_{ij} = P(x_t = s_j | x_{t-1} = s_i) = \frac{Count(x_t = s_j, x_{t-1} = s_i)}{Count(x_{t-1} = s_i)}$$
$$\text{where, } \forall a_{ij} \geq 0 \text{ and } \sum_{j=0}^{N-1} a_{ij} = 1, i = 0,1, \ldots, N-1 \quad (13)$$

Finally, the confusion matrix $B^v = \{b_{i,j}^v\}$ is a $n \times m$ matrix represented as follows:

$$b_{i,j}^v = P(y_t^v = o_j^v | x_t = s_i) = \frac{Count(y_t^v = o_j^v, x_t = s_i)}{Count(x_t = s_i)}$$
$$\text{where, } \forall b_{ij} \geq 0, \sum_{j=1}^{m} b_{ij} = 1, i = 0,1, \ldots, N-1, v = 0,1, \ldots, K-1 \quad (14)$$

We construct an HMM model $\lambda = (\pi, A, E)$ and use this model to forecast.

C. Forecasting and defuzzificationwihwui

There are many alogrithms that can compuate the probability of the observations and we can also estmate the next state by getting maximal probability. Our study only focuses on forecasting, so the proposed method just uses dynamic programming to calculate maximum likelihood.

This study assumes that the variables of observation are independent of each other. Based on the concept of statistical independence, we calculate the probability directly to present the probability of observations.

$$b_{i,j} = \prod_{v=0}^{K-1} b_{i,j}^v \quad (15)$$

Based on the dynamic programming method, we construct the following equation:

$$\begin{aligned} P(Y | \lambda) &= \sum_X P(Y, X | \lambda) \\ &= \sum_X P(Y | X, \lambda) P(X | \lambda) \\ &= \sum_X \pi_{x_0} b_{x_0,y_0} a_{x_0,x_1} b_{x_1,y_1} a_{x_1,x_2} \ldots a_{x_{T-2},x_{T-1}} b_{x_{T-1},y_{T-1}} \\ &= \sum_X \pi_{x_0} b_{x_0,y_0} \prod_{i=0}^{T-2} a_{x_i,x_{i+1}} b_{x_{i+1},y_{i+1}} \end{aligned} \quad (16)$$

According to the notation of our study, $y_t = \{y_t^0, y_t^1, \ldots, y_t^{K-1}\}$, we then edit the model as follows:

$$P(Y \mid \lambda) = \sum_{X} \pi_{x_0} b_{x_0,y_0} \prod_{i=0}^{T-2} a_{x_i,x_{i+1}} b_{x_{i+1},y_{i+1}}$$

$$b_{i,j} = \prod_{v=0}^{K-1} b_{i,j}^{v} \quad (17)$$

$$P(Y \mid \lambda) = \sum_{X} \pi_{x_0} \prod_{v=0}^{K-1} b_{x_0,y_0}^{v} \left[\prod_{i=0}^{T-2} a_{x_i,x_{i+1}} \left(\prod_{v=0}^{K-1} b_{x_{i+1},y_{i+1}}^{v} \right) \right]$$

According to (18), we obtain the probability of hidden state with given observations. Therefore, following the sequence of maximal probability, we can reach the forecasting sequence of hidden state. However, the sequence we estimated is fuzzy time series. Finally, we need to defuzzy the outcome.

There are several defuzzificaton methods that can be chosen. We use the most popular one, namely, centroid method, which standardize the fuzzy output and use the midpoint of each interval to calculate the centroid of fuzzy set as the forecast value. This method is expressed as:

$$\text{Centroid Method} = \frac{\sum_{i=1}^{N} \mu_A(x_i) C_i}{\sum_{i=1}^{N} \mu_A(x_i)} \quad (18)$$

N: the amount of fuzzy set;

$\mu_A(x_i)$: the x_i membership degree belonging to fuzzy se tμ_A;

C_i: the i^{th} midpoint of interval corresponding to the i^{th} linguistic value.

IV. Experiment Results

The present section demonstrates the application of the proposed method and compared the accuracy of its forecasted results with those results obtained by one factor only. In order to evaluate the superiority of proposed model, we use four evaluation indices to evaluate the performance, such as MAE (Mean Absolute Error), PMAD (Percent Mean Absolute Deviation), MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error).

In this experiment, we conduct the proposed model to forecast the average temperature with other three observations. With the forecasting results, the following Figure 7 shows the respective performances of the Alishan weather forecasting by implementing the proposed model.

This time series data are presented by month and contains four main factors: (1) average temperature (2) average humidity level (3) number of rainy days (4) total sunshine duration. In this experiment, we conduct the proposed model to forecast average temperature with three other factors. In order to confirm the performance of the proposed model, we also exhibit the forecasting result which is estimated by other methods we have mentioned previously. The forecasting result is displayed in Figure 7.

	MAE	RMSE	PMAD	MAPE
Proposed Model	1.0861	1.5100	0.0933	0.1097
Chen(1996)	1.3159	1.6584	0.1145	0.1316
Hsu et al. (2003)	1.2397	1.5269	0.1079	0.1251
Li & Cheng (2007)	2.0443	2.7463	0.1779	0.2078
Chen (2011)	1.0738	1.4074	0.0923	0.1065

Figure 7. The Forcasting result For Weather Data.

All the evaluation indices that are smaller will be favorable. We can discover that the proposed model has better prediction accuracy than most methods except Chen's newest model. Even though Chen's model has better prediction accuracy than the proposed one, the forecasting results are quite similar between two models. In Chen's model, high order information, which requires more computation, was considered. The proposed model uses four factors one order to forecasting, and has the similar performance with the model with one factor high orders. As a result, we can realize the power of multiple factors.

V. Conclusion and Futrue Work

The drawback of traditional forecasting models is that they cannot forecast with multiple factor data and waste the obtained information. However, this proposed model solves the problem and demonstrates the indication that "predicting with more factors can improve the forecasting result".

There is a point can be focused in the future work. In this model, we assume the relations between the observed factors are independent. However some realistic data cannot satisfy with this limitation. Therefore, we need to consider the impact of coefficient between observed factors. But it may make the model to be more complicated in calculation. In the future, we may make our effort on adjusting the model with consideration of coefficient more efficiently.

References

[1] B. P. Joshi and S. Kumar, "Intuitionistic Fuzzy Sets Based Method For Fuzzy Time Series Forecasting." Cybernetics and Systems, 43(1) (2012), pp. 34–47.

[2] C. H. Cheng, Y. S. Chen, and Y. L. Wu, "Forecasting innovation diffusion of products using trend-weighted fuzzy time-series model." Expert Systems with Applications, 36(2) (2009), pp. 1826–1832.

[3] H. K. Yu, "Weighted fuzzy time series models for TAIEX forecasting." Physica A: Statistical Mechanics and its Applications, 349(3) (2005), pp. 609–624.

[4] J. Sullivan and W. Woodall, "A comparison of fuzzy forecasting and Markov modeling." Fuzzy Sets and Systems, 64(1994), pp. 279–293.

[5] K. Huarng and T. H. K. Yu, "The application of neural networks to forecast fuzzy time series." Physica A: Statistical Mechanics and its Applications, 363(2) (2006), pp. 481–491.

[6] L. W. Lee, L. H. Wang, S. M. Chen, and Y. H. Leu, "Handling forecasting problems based on two-factors high-order fuzzy time series." IEEE Transactions on Fuzzy Systems, 14(3) (2006), pp. 468–477.

 ISBN: 978-1-61208-364-3

[7] M. Shah, “Fuzzy based trend mapping and forecasting for time series data.” Expert Systems with Applications, 39(7) (2012), pp. 6351–6358.

[8] N. Y. Wang and S. M. Chen, “Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series.” Expert Systems with Applications, 36(2) (2009), pp. 2143–2154.

[9] Q. Song and B. Chissom, “Fuzzy time series and its models.” Fuzzy sets and Systems, 54(1993), pp. 269–277.

[10] S. M. Chen, “Forecasting Enrollments Based On High-Order Fuzzy Time Series.” Cybernetics and Systems, 33(1) (2002), pp. 1–16.

[11] S. M. Chen and C. D. Chen, “Handling forecasting problems based on high-order fuzzy logical relationships.” Expert Systems with Applications, 38(4) (2011), pp. 3857–3864.

[12] S. T. Li and Y. C. Cheng, “A stochastic HMM-based forecasting model for fuzzy time series. IEEE transactions on systems, man, and cybernetics.” Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 40(5) (2010), pp. 1255–1266.

[13] S. T. Li and Y. C. Cheng, “Deterministic fuzzy time series model for forecasting enrollments.” Computers & Mathematics with Applications, 53(12) (2007), pp. 1904–1920.

[14] W. K. Wong, E. Bai, and A. W. C. Chu, “Adaptive time-variant models for fuzzy-time-series forecasting.” IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 40(6) (2010), pp. 1531–42.

[15] Y. Wang, X. Hao, X. Zhu, and F. Ye, “An approach of software fault detection based on HMM.” In 2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, 2012, pp. 644–647.

 ISBN: 978-1-61208-364-3

The Group Strategic Knowledge Mining Model for Telecom Power Infrastructure

Sheng-Tun Li and Wei-Chien Chou
Department of Industrial and Information Management
Institute of Information Management
National Cheng Kung University
No.1, University Road, Tainan City, Taiwan (R.O.C.)
stli@mail.ncku.edu.tw, sandy.chou@msa.hinet.net

***Abstract*—The telecom industry faces many changes of competitive environments and challenges of technological innovations. Many companies make efforts to solidify their technological superiority and maintain the market share. Making the best use of limited resources and efficient resource allocation are the most important tasks. Among all resources, the power infrastructure of telecom rooms plays an increasingly critical role in broadband infrastructure and it is the heart of telecom resources. This study develops a group strategic knowledge mining model based on knowledge discovery and mining to clearly elicit the criteria and evaluate the strategic values of the alternatives. Furthermore, by uncovering the clusters knowledge rules and generating the decision tree, the decision makers' knowledge will be mined. The contributions of the proposed model may support the planners and managers to develop more effective power management strategies for telecom rooms.**

Keywords-telecom power infrastructure; fuzzy modified Delphi method; multiple criteria decision making; fuzzy aggregation method.

I. INTRODUCTION

Taiwan has one of the more advanced telecom networks in the region of Asia Pacific. Since the privatization and liberalization of Taiwan's telecom markets in 2005, the telecom industry faced many changes of competitive environments and challenges of technological innovations. Telecom companies need to find new ways to obtain competitive advantages. According to the administrative plan of 2013 proposed by the National Communications Commission (NCC), enhancing the monitoring mechanism of high-speed broadband network, and amending guidelines for fixed-line network service quality are part of the telecom supervision plan in Taiwan [1]. Chunghwa telecom (CHT) company is the largest telecom operator in Taiwan, and no one can compare with its rich talented personnel, facilities, fund and network construction [2]. The CHT's 2014 guidance also reported "Facing market saturation and competition on broadband business, we will continue to support growth through offering attractive convergence plans and facilitating higher-speed migration. Lastly, we plan to launch ground-breaking 300Mbps speed services, helping us solidify our technological superiority and maintain our market leading position." [3]. Therefore, the telecom infrastructure plays an important role in CHT's broadband strategic management.

The telecommunications room (telecom room) gathers all connectivity from customers or business, such as broadband networks, data communications, and fixed line. Due to the needs of customers and the market share, the telecom infrastructure needed to provide more diversification of services. The telecom rooms play an increasingly critical role in broadband infrastructure, among these all, power infrastructure is the heart of telecom resources. In terms of management and operation, CHT has some inborn inferiorities because it has been a monopoly for 40 years [2]. To make the best use of limited resources and efficient resource allocation, CHT needs to develop more effective management strategies for power infrastructure to meet increasing demand for broadband network performance in the business market.

With the advent of the Internet and the development in the telecom industry, in selection of telecom alternatives for the power system protection and control applications, greater emphasis is usually placed on reliability than cost [4]. Additionally, the management of telecom power infrastructure based on knowledge discovery from domain experts were however less discussed. Therefore, this study developed a group strategic knowledge mining (GSKM) model for telecom power infrastructure. The objectives of this paper are to clearly elicit the criteria and evaluate the strategic values of the alternatives. Furthermore, by uncovering the clusters knowledge rules and generating the decision tree, the decision makers' knowledge will be mined. The contributions of the proposed model may support the planners and managers to develop more effective power management strategies for telecom rooms. The existing methodologies in the GSKM model include: (1) Eliciting criteria for knowledge discovery by using the modified Delphi method; (2) Evaluating the weights and alternatives to get strategic values by using fuzzy MCDM method; (3) Integrating the group experts' opinions by using two fuzzy aggregation method; (4) Clustering the telecom rooms for mining the experts' knowledge and advice appropriate management strategies of power infrastructure.

II. BACKGROUND

A. Knowledge elicitation

The Delphi method is a set of procedures and methods for formulating a group judgment toward a subject matter in which precise information is lacking, and relies on soliciting individual (often anonymous) answers to written questions by survey or other type of communication [5]. A series of iterations provide each individual with feedback on the responses of the others in the group. The final responses are

 ISBN: 978-1-61208-364-3

evaluated for variance and means to determine which questions the group has reached consensus about, either affirmatively or negatively.

Murry and Hammons defined that the modified Delphi method is a technique to arrive at a group consensus regarding an issue under investigation [6]. It was used to rate the indicators. This process consisted of one round of anonymous ratings of the indicators by the panel, a face-to-face panel discussion, and a second round of anonymous ratings immediately after the panel discussions. It is a structured approach to expert panel deliberations that does not require consensus.

Iggland applied fuzzy Delphi method in coupling of customer preferences and production cost information [7]. Ishikawa et al. implemented experts' judgments with group fuzzy integration based on fuzzy Delphi method (FDM) [8]. Applying fuzzy set theory to the Delphi group decision method seems attractive as demonstrated by Cheng who utilized fuzzy Delphi method to adjust the fuzzy rating of each expert [9]. Therefore, the fuzzy modified Delphi process was used to develop consensus in the proposed model.

B. Getting the Fuzzy Weights for criteria

Because an evaluator always perceives the weight with their own subjective evaluation, an extra or precise weight for a specified criterion was not given. This led to the use of the fuzzy weights of criteria. In decision analysis, pairwise comparison of alternatives is widely used [10]. Usually, decision makers express their pairwise comparison information in two formats: multiplicative preference relations and fuzzy preference relations. The analytic hierarchy process (AHP) with multiplicative preference relations has been applied extensively in telecom fields [11][12][13][14]. However, the decision makers may also use fuzzy preference relations to express their preference due to their different cultural and educational backgrounds, personal habits, and the vague nature of human judgment. There are some common research issues between multiplicative preference relations and fuzzy preference relations, as both are based on pairwise comparison. Therefore, research progress in multiplicative preference relations can benefit research in fuzzy preference relations [15].

The fuzzy preference relations method provides some advantages, including a consistency indicator, simplicity of computation, high precision and preservation of ranks. The method constructs the decision matrices of pairwise comparisons using an additive transitivity. Only comparisons are required to ensure consistency for a level with criteria. The method is simply and practically provides ranking choices in decision-making problems [15][16][17].

Current approaches for group decision-making analysis support different preference formats, but their computational procedures are very complicated. Usually, they consist of three steps: (1) uniform the preference information given by decision makers through a transformation function, (2) aggregate the uniformed preference information into a collective one by means of the aggregation operators, and (3) rank alternatives or select the most desirable alternatives by the selection methods [18].

The multi-granular linguistic methodology permits the unification of the different linguistic domains to facilitate the calculus of consensus degrees and proximity measures on the basis of experts' opinions. The consensus degrees assess the agreement amongst all the experts' opinions, while the proximity measures are used to find out how far the individual opinions are from the group opinion [17]. Therefore, this study assumed that there exist several experts who may have different background and knowledge to solve a particular problem and, therefore, different linguistic term sets (multi-granular linguistic information) could be used to express their opinions.

C. Aggregating linguistic labels into a group opinion

1) FLOWA (Fuzzy Linguistic Ordered Weighted Average): At present, many aggregation operators have been developed to integrate information. The aggregation function, ranking method, and consensus measure are the main problems to be solved for a fuzzy group decision-making issue. The existing main aggregation operators can be briefly classified into the following three categories: (1) One contribution of the methodology presented herein is that the result of this aggregation approach is a collection of linguistic labels with a calculated degree or membership function, presenting a more informative aggregation. (2) A two phase model developed by Herrera and Herrera-Viedma [19]. (3) The operators, which can only be used in situations where the arguments are exact numeric variables, such as Linguistic Ordered Weighted Averaging (LOWA) operators that is based on the OWA and the convex combination of linguistic labels [20]. Ben and Chen proposed a new linguistic-label aggregation operator incorporated fuzzy set theory into LOWA that call the fuzzy-LOWA (FLOWA) operator. FLOWA organized OWA and LOWA aggregation algorithms and individual linguistic opinions into a group opinion. These aggregation methods operate directly on the linguistic labels and allow each expert to represent an optimistic or pessimistic predilection [21].

2) EFWA (Efficient Fuzzy Weighted Average): When the environment is vague, the rating criteria and the weights of their corresponding importance are often evaluated as a fuzzy number. In order to obtain the weighted sum of those criteria evaluated by fuzzy numbers in terms of rating and importance, this study use the fuzzy weighted average for the calculation. There has been some research involved in the field of fuzzy weighted average [22][23][24]. Lee and Park proposed an efficient algorithm, named the Efficient Fuzzy Weighted Average (EFWA), to compute a fuzzy weighted average, which was an improvement over the previous methods by reducing the number of comparisons and arithmetical operations [22]. The computational algorithm of EFWA is based on the α-cut representation of fuzzy sets and interval analysis. The managerial meaning of

 ISBN: 978-1-61208-364-3

α-value can be explained as a confidence level [23]. Because of the above-mentioned advantages, this study will adopt the EFWA algorithm to aggregate decision makers' opinion.

III. THE GROUP STRATEGIC KNOWLEDGE MINING MODEL

This section proposes the GSKM model. It includes three phases, as shown in Figure 1. Phase I is the process of criteria elicitation and weights assessment for telecom knowledge discovery. Phase II is to evaluate and aggregate the alternatives, namely telecom rooms, to get strategic values and ranks. Phase III is the process of clustering the telecom rooms and mining the decision tree for the power infrastructure.

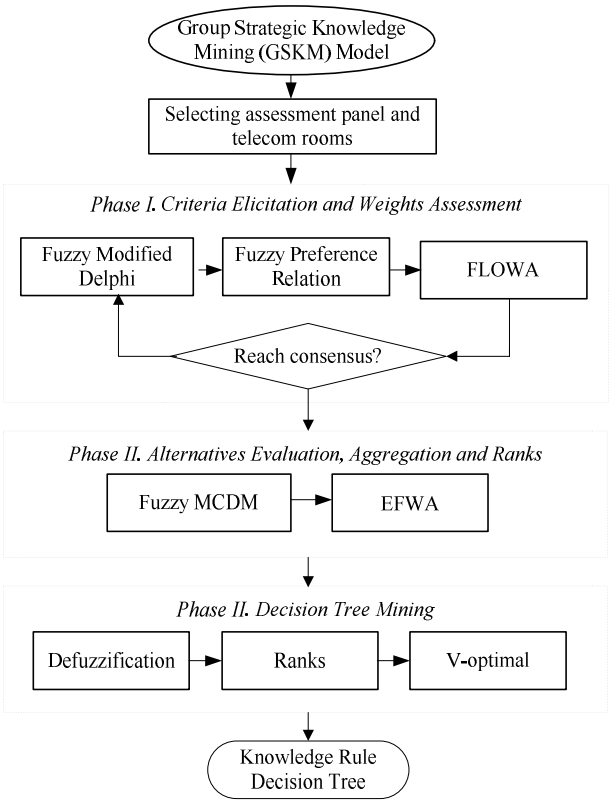


Figure 1. The GSKM model.

A. Phase I. Criteria Elicitation and Weights Assessment

To identify the important criteria for the operation strategies in power infrastructure of telecom rooms, the fuzzy modified Delphi method is used to elicit implicit knowledge of decision makers and assess criteria for deciding weights. The decision makers are invited to take part in a series of rounds to identify, clarify until reaching consensus on this issue.

This process consisted of the first round of anonymous ratings of the indicators and a face-to-face discussion, and then generated an impact assessment form for power infrastructure of telecom rooms. The decision makers investigated that MCDM problem with preference relations could assess the relative importance weights of criteria. Moreover, human thoughts are uncertain, fuzzy theory could be used to express their preference relations in linguistic terms. To integrate the decision makers' opinions, the FLOWA method was used to aggregate all criteria. This way, the decision makers' valuable domain knowledge can be acquired and shared in an effective and efficient manner. In the following rounds, a series of discussion are provided feedback to the decision makers, the appropriate thresholds are also restricted to achieve the goals by cognitive subjective view until consensus is reached. Finally, criteria and weights are decided in this phase.

B. Phase II. Alternatives Evaluation and Aggregation

To evaluate the alternatives with respect to criteria, the MCDM method consists of two steps: (1) collecting the individual confidence level, define the linguistic terms and corresponding triangular fuzzy numbers, and (2) aggregating all decision makers' opinions and integrating the scores of weighing the criteria for each alternative.

The first step should utilize fuzzy set theory to deal with the uncertain problem of linguistics, and the linguistic terms are replaced by suitable triangular fuzzy numbers that are used for arithmetical operations. Each decision maker defined their different confidence levels and corresponding triangular fuzzy numbers in terms of VH (very high), H (high), M (medium), L (low), VL (very low) to score the importance of alternatives that collected with respect to the criteria elicited. When the scores finished assessing completely, the strategic values will be obtained, and then a decision matrix is established for the decision makers.

In the second step, in order to obtain the weighted sum of those criteria evaluated by fuzzy numbers in terms of rating and importance, we used EFWA method to aggregate the weighted scores that are from individual decision-makers to become the group results. Therefore, the alternatives can be ranked according to the group scores.

C. Phase III. Decision Tree Mining

This phase included defuzzification and data analysis. By clustering for discovering the knowledge rules and feedback to the decision makers, two algorithms are employed. One is center of gravity for defuzzifing and ranking the alternatives, and another one is V-optimal for uncovering the clusters knowledge rules and generating the decision tree. Therefore, the results may help decision makers to more effectively manage the telecom power infrastructure, and thus obtain competitive advantages.

IV. A CASE ILLUSTRATION

In order to illustrate the practicability and usefulness of the proposed model, we implemented it in the power infrastructure of telecom rooms for the largest telecom company in Taiwan. We convened three decision makers who are the most knowledgeable in power technologies and services. The processes of evaluating the strategies can be expressed as follows.

 ISBN: 978-1-61208-364-3

A. Identification of selecting criteria and deciding weights

In the first phase, the fuzzy modified Delphi method was applied to reach consensus on the importance of each of the identified criteria for evaluating operation strategies toward the power infrastructure of telecom rooms. During the first round, the decision makers were provided the fault research, cause analysis, accident causation and process and the choice of UPS of the telecom rooms' electric power system to understand practical implementation, planning and management. Based on the decision makers' deep domain knowledge about the related geographical information, facilities, personnel allocation and future investment plan of the telecom rooms, they were elicited the criteria and suggested criteria through brainstorm technique. After discussing in the first round, this study analyzed the priority ranking of criteria is as follows: staffing > room features > initial load > power supply > growth forecast > maintaining support, and included 266 telecom rooms in Southern Taiwan.

The decision makers used fuzzy preference relations to assign the strategic values for criteria. The value represents the important degree of the preference for the first criteria with respect to the second one. The decision matrix, which is based on Saaty's 9 point scale, is constructed. Therefore, the decision makers used the fundamental 1–9 scale defined by Saaty to assess the priority score [21]. To combine the scores of fuzzy preference relation that each decision maker assessed, the FLOWA method is applied to get final weights of each criterion. Herein, the matrix and the weight factors of 6 evaluation criteria for group opinions are shown in Table I.

TABLE I. THE WEIGHTS OF CRITERIA

	Staffing	Room features	Initial load	Power supply	Growth forecast	Maintaining support	Weight
Staffing	0.50	0.57	0.71	0.75	0.87	1.00	0.24
Room features	0.43	0.50	0.63	0.68	0.79	0.93	0.22
Initial load	0.29	0.37	0.50	0.54	0.66	0.79	0.18
Power supply	0.25	0.32	0.46	0.50	0.62	0.75	0.16
Growth forecast	0.13	0.21	0.34	0.38	0.50	0.63	0.12
Maintaining support	0.00	0.07	0.21	0.25	0.37	0.50	0.08

In Table I, the criteria "maintaining support" was "extreme unimportance" and ranked last. In the next round, the decision makers decided to give up it for satisfying the threshold condition, and then regenerated the matrix of five evaluation criteria. According to Table II, the weight set on C_1: staffing, C_2: room features, C_3: initial load, C_4: power supply, and C_5: growth forecast respectively were 0.3, 0.26, 0.29, 0.16 and 0.1.

B. Assessment of telecom rooms

In phase II, the three decision makers, who have the basic knowledge about fuzzy theory are then introduced to the basic concepts of common linguistic term set (LTS) by the facilitator. After a brief introduction, the decision makers defined their linguistic terms and corresponding triangular fuzzy numbers according to their subjective judgments within a scale of 0-10. The LTS {VH, H, M, L, VL} indicates very high, high, medium, low, and very low, respectively. See Table III.

TABLE II. THE WEIGHTS OF CRITERIA

	Staffing	Room features	Initial load	Power supply	Growth forecast	Weight
Staffing	0.50	0.60	0.78	0.84	1.00	0.30
Room features	0.40	0.50	0.68	0.74	0.90	0.26
Initial load	0.22	0.32	0.50	0.56	0.72	0.19
Power supply	0.16	0.26	0.44	0.50	0.66	0.16
Growth forecast	0.00	0.10	0.28	0.34	0.50	0.10

TABLE III. THE SUBJECTIVE PERCEPTION OF DECISION MAKERS OF THE FIVE LEVELS OF LINGUISTIC VARIABLES.

Linguistic Variables	**Fuzzy Numbers**		
	Decision maker 1	***Decision maker 2***	***Decision maker 3***
Very high	(8, 10, 10)	(9, 10, 10)	(8, 10, 10)
High	(5, 8, 10)	(5, 9, 10)	(5, 8, 10)
Medium	(3, 5, 9)	(1, 5, 9)	(2, 5, 8)
Low	(0, 3 ,5)	(0, 1, 5)	(0, 2, 5)
Very low	(0, 0, 3)	(0, 0, 1)	(0, 0, 2)

Each decision maker used their linguistic terms to evaluate the 266 telecom rooms with room features, staffing, initial load, growth forecast and power supply on power infrastructure of telecom rooms for the importance of operation strategies. To aggregate the evaluation results from the three decision makers, the EFWA method was used to calculate the weighted scores of the criteria. Through repeating the computational procedure of EFWA, the interval for α = 0, in which each point is corresponding to the end points of the triangle representing the membership function. The process is repeated for α =1, which corresponds to the center of the triangle. Consequently, with the intervals for α = 0 and α = 1, the aggregation results of EFWA for partial telecom rooms are shown in Table IV.

TABLE IV. AGGREGATE THE EVALUATION RESULTS USING EFWA

Room #	C_1			C_2			C_3			C_4			C_5		
	C_1L	C_1M	C_1R	C_2L	C_2M	C_2R	C_3L	C_3M	C_3R	C_4L	C_4M	C_4R	C_5L	C_5M	C_5R
1	5.33	10.00	10.00	8.44	5.33	10.00	10.00	8.44	7.33	9.33	10.00	8.89	6.00	2.67	6.33
51	0.00	1.00	5.67	2.22	0.00	1.00	5.67	2.22	0.00	1.00	5.67	2.22	0.00	1.00	5.67
101	6.00	5.00	5.67	5.56	6.00	5.00	5.67	5.56	6.00	2.67	6.33	5.00	0.00	2.00	5.00
151	6.00	2.67	6.33	5.00	6.00	2.67	6.33	5.00	0.00	1.00	5.67	2.22	6.00	1.67	7.00
201	6.00	5.00	5.67	5.56	6.00	2.67	6.33	5.00	0.00	1.00	5.67	2.22	0.00	1.00	5.67
251	6.00	2.67	6.33	5.00	6.00	1.67	7.00	4.89	6.00	2.67	6.33	5.00	0.00	1.00	5.67

C. Generating and exploring classification rules

The aggregation matrix, which is a linguistic term, should be transferred to non-fuzzy values; herein the Center of Gravity (COG) method are used to produce the Best Non-fuzzy Performance (BNP) value in given fuzzy sets and corresponding membership degrees. According to the BNP

 ISBN: 978-1-61208-364-3

value, 266 telecom rooms will be ranked. After determining the ranked result, the V-optimal histogram, which is suitable for clustering ordinary data, and based on the concept of minimizing the quantity of the weighted variance, clustered the telecom rooms into five groups. The numbers of each type of telecom room for class A, B, C, D, and E are 31, 68, 72, 32, and 63, respectively.

PolyAnalyst decision tree algorithm can generate classification rules to help the decision makers uncover knowledge for developing effective management strategies. The decision tree can grow each branch just deeply enough to perfectly classify the examples, as shown in Figure 2.

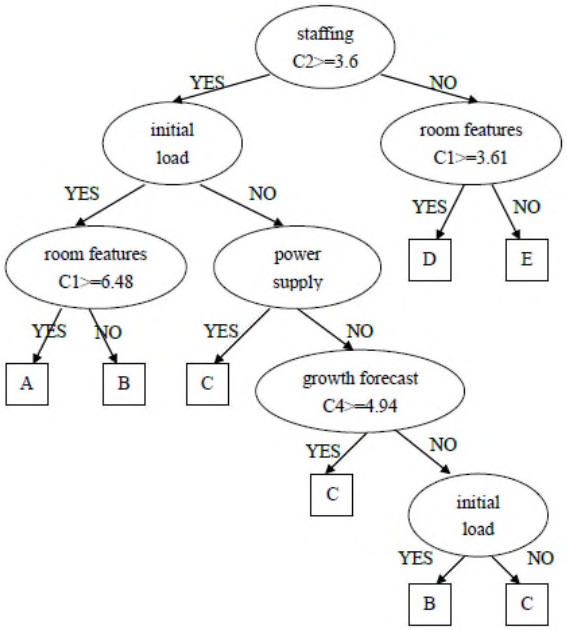


Figure 2. The decision tree.

D. Administrative operational communication

This study implemented and explored if it will enhance readability and simplify the decision-making task for the differences between the management modes and the types of power system. The features and strategies of each class of telecom rooms are shown in Table V.

TABLE V. THE FEATURES AND STRATEGIES OF EACH CLASS OF TELECOM ROOMS

Class	Location	Customer Type	Strategy
A	Metropolitan, commercial and industrial areas.	Enterprise, and leased lines with value-added potential customers	Provide a high standard of telecomm equipment for customer services.
B	Center of township district.	Many competitors are eager to enter such markets.	Keep the long-term customers and create revenue.
C	Center of small township district.	The customer revenue is declining.	Keep the long-term customers and raise value-added services
D	Low growth areas.	The low customer revenue growth	Need to fulfill the social responsibility
E	Negative growth or non-economic region.	The average telephone charges per user are not very high	Deploy the basic equipment on the business investments.

V. CONCLUSIONS

The privatization, liberalization and competition of telecom markets have created a really strong competition. In the new era of the digital economy, telecom resources lack effective management [25][26]. Many telecom companies and DGT pay more attention to the problem. If organizations can achieve performance of implementation of strategy, invest in telecom equipment accurately, and reduce operating costs, they will make optimal use of company resources. According to the classes of the telecom rooms (Table VI), CHT can apply the management modes to the planning stages of engineering works, particularly the initial stages or re-planning of power equipment to advance the effective use and resource development and utilization.

TABLE VI. OPERATIONAL MODES OF POWER PLANNING ADVISED FOR EACH CLASS OF TELECOM ROOM

Class	DC reserve capacity	Automatic Generation	Monitoring Equipment	Maintenance Personnel
A	Design capacity with more than three hours	Full power capacity, and two sets of automatic generation.	Installation and use of the automatic monitoring equipment to connect with main monitoring centrer.	Full-time and first-level maintenance personnel.
B	Design capacity with three hours	Full power capacity and one set of automatic generation.		Full-time and second-level maintenance personnel.
C	After blackout, the power backup system can be maintained two hours.			When failure occurs, second level maintenance personnel will take over the jobs.
D	After blackout, the power backup system can be maintained one hour	Basic power capacity and one set of automatic generation.		Centralized management by regional maintenance unit.
E		Automatic generation.		When failure occurs, delegate technicians will support based on the fault level.

This study proposed the GSKM model for power infrastructure to support maintenance strategies for telecom rooms. Group decision making with structured process is to improve the quality of decision-making and the results can be better judged. Followed by interviewing the decision makers, ascertain opinions by conducting knowledge discovery and drawing up evaluation criteria - staffing, office features, initial load, power supply and growth forecast. The GSKM model was based on knowledge discovery and mining generated the classification rules and decision tree through MCDM evaluation to develop an effective telecom

 ISBN: 978-1-61208-364-3

power infrastructure strategies. It will enhance readability and simplify the decision-making task for the differences between management modes and types of power system in each telecom room.

There are generally two main objectives in constructing GSKM model. The first objective concerns the improvement of the quality of the decisions taken. During the planning stages of engineering works, particularly the initial stages or re-planning of power equipment, power maintenance strategies can be integrated into the design-lifecycle to advance the effective use and resource development and utilization. The second objective of a formalized decision study is to supply technical documentation in support of decisions both in front of authorities and of public opinion.

Theoretically, most of scholars and experts studied the electric power transmission system, discussed how to solve the power distribution system, and trending econometric analysis. Because of different nature of each telecom room, this study may not be applied to plan telecom power infrastructure. Finally, the model will apply to other domain and compare with other group model in future work.

REFERENCES

[1] National Communications Commission, "National Communications Commission Administration Plan 2013," National Communications Commission web site, retrieved 24 Feb. 2014 <http://www.ncc.gov.tw/english/files/13050/306_130508_1.pdf>

[2] C. J. Chang, C. Y. Chen, and I-Ting Chou, "The design of information and communication technologies: telecom MOD strength machines," Journal of Vibration and Control, vol. 19, July 2013, pp. 1499-1513, doi: 10.1177/1077546312449644.

[3] Chunghwa Telecom, "Chunghwa Telecom Reports 2014 Guidance," Chunghwa Telecom web site, retrieved 24 Feb. 2014 <http://www.cht.com.tw/en/aboutus/messages/msg-140128-171656.html>

[4] A. J. Conejo, R. Garcia-Bertrand, and M. Diaz-Salazar, "Generation maintenance scheduling in restructured power systems," Power Systems, IEEE Transactions on, vol. 20, May 2005, pp. 984-992, doi: 10.1109/TPWRS.2005.846078.

[5] G. Rowe and G. Wright, "The Delphi technique as a forcasting tool: Issues and analysis," International Journal of Forecasting, vol. 15, Oct. 1999, pp. 353-375, doi: 10.1016/S0169-2070(99)00013-8.

[6] J. W. Murry and J. O. Hammons, "Delphi: A versatile methodology for conducting qualitative research," The Review of Higher Education, vol. 18, 1995, pp. 423-436.

[7] B. Iggland, "Coupling of customer preferences and production cost information," Technology Management : the New International Language, Oct. 1991, pp. 250-253, doi: 10.1109/PICMET.1991.183626.

[8] A. Ishikawa, M. Amagasa, T. Shiga, G. Tomizawa, R. Tatsuta, and H. Mieno, "The max-min Delphi method and fuzzy Delphi method via fuzzy integration," Fuzzy Sets and Systems, vol. 55, May 1993, pp. 241-253, doi: 10.1016/0165-0114(93)90251-C.

[9] C. H. Cheng, "A simple fuzzy group decision making method" IEEE International Conference on Fuzzy Systems, vol. 2, Aug. 1999, pp. 910-915, doi: 10.1109/FUZZY.1999.793073.

[10] R. E. Bellman and L. A. Zadeh, "Decision-Making in a Fuzzy Environment" Management Sciences, vol. 17, Dec. 1970, pp. 141-164, doi:10.1287/mnsc.17.4.B141.

[11] H. C. Yu, Z. Y. Lee, and S. C. Chang, "Using a fuzzy multi-criteria decision making approach to evaluate alternative licensing mechanisms, "Information and Management, vol. 42, May 2005, pp. 517-531, doi: 10.1016/j.im.2002.12.001.

[12] C. Yang, G. L. Fu, and G. H. Tzeng, "Creating a Win-Win in the Telecommunications Industry: The Relationship between MVONs and MNOs in Taiwan," Canadian Journal of Administrative Sciences, vol. 22, Dec. 2005, pp. 316-328, doi: 10.1111/j.1936-4490.2005.tb00377.x.

[13] Y. F, Kuo and P. C. Chen, "Selection of mobile value-added services for system operators using fuzzy synthetic evaluation," Expert Systems with Applications, vol. 30, May 2006, pp. 612-620, doi: 10.1016/j.eswa.2005.07.007.

[14] G. Isiklar and G. Buyukozkan, "Using a multi-criteria decision making approach to evaluate mobile phone alternatives," Computer Standards & Interfaces, vol. 29, Feb. 2007, pp. 265-274, doi: 10.1016/j.csi.2006.05.002.

[15] E. Herrera-Viedma, F. Herrera, F. Chiclana, and M. Luque, "Some Issues on Consistency of Fuzzy Preference Relations," European Journal of Operational Research, vol. 154, Apr. 2004, pp.98-109, doi: 10.1016/S0377-2217(02)00725-7.

[16] F. Chiclana, F. Herrera, and E. Herrera-Viedma, "Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations," Fuzzy Sets and Systems, vol. 122, Sep. 2001, pp. 277-291, doi: 10.1016/S0165-0114(00)00004-X.

[17] E. Herrera-Viedma, L. Martínez, F. Mata, and F. Chiclana, "A consensus support system model for group decision-making problems with multigranular linguistic preference relations," IEEE Transactions on Fuzzy Systems, vol. 13, Oct. 2005, pp.644-658, doi: 10.1109/TFUZZ.2005.856561.

[18] Z. P. Fan, J. Ma, Y. P. Jiang, Y. H. Sun, and L. Ma, "A goal programming approach to group decision making based on multiplicative preference relations and fuzzy preference relations," European Journal of Operational Research, vol. 174, Oct. 2006, pp. 311-321, doi: 10.1016/j.ejor.2005.03.026.

[19] M. Delgado and F. Herrera, "A communication model based on the 2-tuple fuzzy linguistic representation for a distributed intelligent agent system on Internet.," Soft Computing, vol.6, Jan. 2002, pp. 320-328, doi: :10.1007/s00500-002-0185-7.

[20] M. Delgado, J. L. Verdegay, and M. A. Vila, "On aggregation operations of linguistic labels," International Journal of Intelligent Systems, vol. 8, 1993, pp. 351-370, doi:. 10.1002/int.4550080303.

[21] D. A. Ben and Z. Chen, "Linguistic-Labels Aggregation and Consensus Measure for Autocratic Decision Making Using Group Recommendations," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 36, May 2006, pp. 558-568, doi: 10.1109/TSMCA.2005.853488.

[22] D. H. Lee and D. Park, "An efficient algorithm for fuzzy weighted average," Fuzzy Sets and Systems, vol. 87, Apr. 1997, pp. 39-45, doi: 10.1016/S0165-0114(96)00027-9.

[23] W. M. Dong and F. S. Wong, "Fuzzy weighted average and implementation of the extension principle," Fuzzy Sets and Systems, vol. 21, Feb. 1987, pp. 183-199, doi: 10.1016/0165-0114(87)90163-1.

[24] T. S. Liou and M. J. Wang, "Fuzzy weighted average : An improved algorithm," Fuzzy Sets and Systems, vol. 49, Aug. 1992, pp. 307-315, doi: 10.1016/0165-0114(92)90282-9.

[25] G. N. Ericsson, "Classification of Power Systems Communications Needs and Requirements: Experiences From Case Studies at Swedish National Grid," IEEE Transactions on Power Systems Delivery, vol. 17, Apr. 2002, pp. 345-347, doi: 10.1109/61.997896.

[26] Ericsson, "Enhancing Telecom Management," Ericsson web site, retrieved 1 July. 2005 <http://www.ericsson.com/products/white_papers_pdf/telecom_management.pdf>

 ISBN: 978-1-61208-364-3

What Grammar Tells About Gender and Age of Authors

Michael Tschuggnall and Günther Specht
Databases and Information Systems
Institute of Computer Science, University of Innsbruck, Austria
{michael.tschuggnall, guenther.specht}@uibk.ac.at

***Abstract*—The automatic classification of data has become a major research topic in the last years, and especially the analysis of text has gained interest due to the availability of huge amounts of online documents. In this paper, a novel style feature based on grammar syntax analysis is presented that can be used to automatically profile authors, i.e., to predict gender and age of the originator. Using full grammar trees of the sentences of a document, substructures of the trees are extracted by utilizing pq-grams. The mostly used patterns are then stored in a profile, which serve as input features for common machine learning algorithms. An extensive evaluation using a state-of-the-art test set containing thousands of English web blogs investigates on the optimal parameter and classifier configuration. Finally, promising results indicate that the proposed feature can be used as a significant characteristic to automatically predict the gender and age of authors.**

***Index Terms*—Author Profiling; Text Classification; Grammar Trees; Machine Learning.**

I. Introduction

With the advent of the internet in general and recently especially with social media, users frequently use the numerous possibilities to compose and post text in various ways. Considering current statistics [1] estimating 70 billion pieces of content shared via Facebook or 190 million short messages posted on Twitter every day, the amount of shared textual information is huge. Although the authors of the latter examples are generally known, the information is most often restricted to a user name. Moreover, there also exist cases like anonymized blogs where every information concerning the originator is intentionally hidden.

In contrast to traditional authorship attribution approaches [2] that try to assign one of several known candidate authors to an unlabeled document, the author profiling problem deals with the extraction of useful meta information about the author. Often this information includes gender and age of the originator [3][4][5], but also other demographic information like cultural background or psychological analyses are examined in recent approaches [6][7]. Where the mining of such information can be applied very well to commercial applications by knowing the percentages of gender and age commenting on a new product release for example, it is also of growing importance in juridical applications (*Forensic Linguistics*) [8], where, e.g., the number of possible perpetrators can be reduced. Moreover especially nowadays in the area of cybercrime [9], recent approaches investigate the content of e-mails [10], suicide letters or try to automatically expose sexual predators from chat logs [11].

In this paper, a novel style feature to automatically extract the gender and age of authors of text documents is presented.

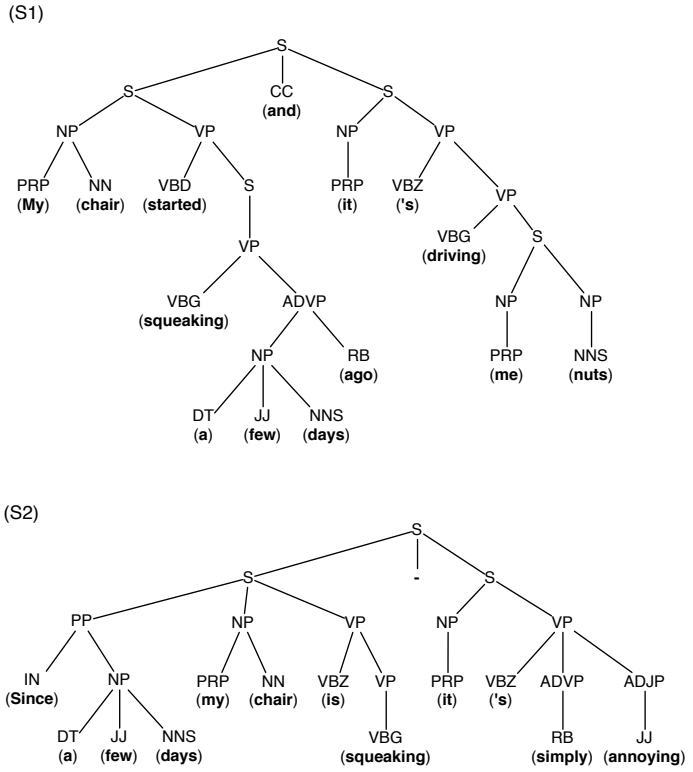


Fig. 1. Grammar Trees of the Semantically Equivalent Sentences (S1) and (S2).

Using results of previous work in the field of intrinsic plagiarism detection [12] and authorship attribution [13], the assumption that individual authors have significantly different writing styles in terms of the syntax that is used to construct sentences has been reused. For example, the following sentence extracted from a web blog: *"My chair started squeaking a few days ago and it's driving me nuts."* (S1) could also be formulated as *"Since a few days my chair is squeaking - it's simply annoying."* (S2) which is semantically equivalent but differs significantly according to the syntax as can be seen in Figure 1. The main idea of this work is to quantify those differences by calculating grammar profiles using pq-grams of full grammar trees, and to evaluate how reliable a prediction of an authors meta information is when solely this grammar feature is used. Given the grammar profiles, the prediction of gender and age, respectively, is finally examined by utilizing modern machine learning approaches like support vector machines, decision trees or Naive Bayes classifiers.

 ISBN: 978-1-61208-364-3

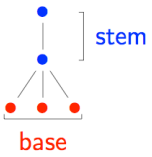


Fig. 2. Structure of a pq-gram Consisting of Stem $p = 2$ and Base $q = 3$.

The rest of this paper is organized as follows: Section II recaps the concept of pq-grams as a fundamental basis of this work, while Section III explains the profiling process in detail. An extensive and promising evaluation using a comprehensive test set of web blogs is presented in Section IV. Finally, related work is summarized in Section V and conclusions including future work are discussed in Section VI.

II. Preliminaries: pq-grams

Similar to n-grams that represent subparts of given length n of a string, pq-grams extract substructures of an ordered, labeled tree [14][15]. The size of a pq-gram is determined by a stem (p) and a base (q) like it is shown in Figure 2. Thereby p defines how much nodes are included vertically, and q defines the number of nodes to be considered horizontally. For example, a valid pq-gram with $p = 2$ and $q = 3$ starting from `PP` at the left side of tree (S2) shown in Figure 1 would be `[PP-NP-DT-JJ-NNS]` (the concrete words are omitted).

The pq-gram index then consists of all possible pq-grams of a tree. In order to obtain all pq-grams, the base is shifted left and right additionally: If then less than p nodes exist horizontally, the corresponding place in the pq-gram is filled with $\star$, indicating a missing node. Applying this idea to the previous example, also the pq-gram `[PP-IN-*-*-*]` (no nodes in the base) is valid, as well as `[PP-NP-*-*-DT]` (base shifted left by two), `[PP-NP-*-DT-JJ]` (base shifted left by one), `[PP-NP-JJ-NNS-*]` (base shifted right by one) and `[PP-NP-NNS-*-*]` (base shifted right by two) have to be considered. As a last example, all leaves have the pq-gram pattern `[`*leaf_label*`-*-*-*-*]`.

Finally, the pq-gram index is the set of all valid pq-grams of a tree, whereby multiple occurrences of the same pq-grams are also present multiple times in the index.

III. Profiling Authors Using pq-gram Indices

The number of choices an author has to formulate a sentence in terms of grammar structure is rather high, and the assumption in this approach is that the concrete choice is made mostly intuitively and unconsciously. Evaluations shown in Section IV reenforce that solely grammar syntax represents a significant feature that can be used to categorize authors.

Basically, the profiling of a given text using pq-grams works as follows:

1) At first the text is parsed and split into single sentences using common sentence boundary detection algorithms, which is currently implemented with *OpenNLP* [16]. Each sentence is then anaylzed by its grammar, i.e., a full syntax tree is calculated using the *Stanford Parser* [17]. For example, Figure 1 depicts the grammar trees resulting from analyzing sentences (S1) and (S2), respectively. The labels of each tree correspond to a part-of-speech (POS) tag of the Penn Treebank set [18], where e.g *NP* corresponds to a noun phrase, *DT* to a determiner or *JJS* to a superlative adjective. In order to examine the building structure of sentences only like it is intended by this work, the concrete words, i.e., the leafs of the tree, are omitted. In case of ambiguity of grammar trees, i.e., if there exist more than one valid parse tree for a sentence, the tree with the highest probability estimated by the parser is chosen.
2) Using the grammar trees of all sentences of the document, the pq-gram index is calculated. As shown in Section II all valid pq-grams of a sentence are extracted and stored into a pq-gram index. By combining all pq-gram indices of all sentences, a pq-gram profile is computed which contains a list of all pq-grams and their corresponding frequency of appearance in the text. Thereby the frequency is normalized by the total number of all appearing pq-grams. As an example, the three mostly used pq-grams using $p = 2$ and $q = 3$ of a sample document are shown in Table I. The profile is sorted descending by the normalized occurrence, and an additional rank value is introduced that simply defines a natural order and is used in the evaluation (see Section IV).

TABLE I
Example of the Three Mostly Used pq-grams of a Sample Document.

pq-gram	Occurrence [%]	Rank
`NP-NN-*-*-*`	2.68	1
`PP-IN-*-*-*`	2.25	2
`NP-DT-*-*-*`	1.99	3

3) Finally, the pq-gram profiles including occurrences and ranks are used as features that are applied to common machine learning algorithms. This step is explained in detail in Section IV.

IV. Evaluation

Basically, the prediction of gender and age of the author of a text document is made by machine learning algorithms. Independent of the classifier used (see Section IV-D), the input consists of a large list of features with appropriate values and a corresponding classification class. The class is used to train the algorithms if the document is part of the training set, as well as for evaluating if the document is part of the test set. Details on the usage of training and test sets, respectively, and on the test corpus in general are explained in Section IV-C.

A. Features

The features that have been used as input for the classifiers consist of the pq-gram profiles described previously. Thereby, each pq-gram represents a feature. To examine the significance of the concrete percentage of occurrence compared to the plain rank, a pq-gram-rank feature has been added additionally.

 ISBN: 978-1-61208-364-3

A small example of a feature list including the correct gender and age classification is depicted in Table II. If a document does not contain a specific feature, i.e., a pq-gram, the feature value for the pq-gram as well as for the corresponding rank is set to -1. For example, the author of document C didn't use the structure `[PP-IN-*-*-*]` to build his/her sentences, so therefore the according feature values are set to -1.

TABLE II
EXAMPLE OF A FEATURE LIST SERVING AS INPUT FOR CLASSIFICATION ALGORITHMS.

Feature	Doc. A	Doc. B	Doc. C
`NP-NN-*-*-*`	2.68	1.89	2.84
`NP-NN-*-*-*--RANK`	1	6	2
`PP-IN-*-*-*`	2.25	0.24	-1
`PP-IN-*-*-*-RANK`	2	153	-1
`NP-DT-*-*-*`	1.99	2.11	1.23
`NP-DT-*-*-*-RANK`	3	2	11
...	...	...	...
correct gender	male	female	male
correct age	20s	10s	30s

Depending on the evaluation setup shown subsequently the number of attributes to be handled by the classification algorithms range between 7,000 and 20,000.

B. Evaluation Setup

The computation of the feature list is an essential part of the approach. Basically, it depends on the assignment of p and q, respectively, that is used for the extraction of pq-grams from sentences. For example, by using $p = 1$ and $q = 0$ the pq-grams would be reduced to single POS tags. Nevertheless, based on results in previous work such configurations have been excluded as they led to insufficient results. The range of both stem and base of pq-grams has been evaluated in the range between 2 and 4, conforming to the size of n-grams that are used in efficient approaches in information retrieval (e.g. [19]).

Considering the huge amount of possible features, especially if $p + q > 6$, the maximum number of sentences per text sample (s_{max}) as well as the maximum number of pq-grams in a profile (pq_{max}) have been limited. Accordingly, only the first 200 sentences of each document have been processed. The final pq-gram profile has then been sorted descending by the rank and limited to the 500 mostly used patterns.

Finally, three different feature sets have been used as input for the machine learning algorithms: the percentage of occurrence of each pq-gram, the rank of each pq-gram, and a combination of both occurrence-rate and rank.

An overview of all settings that have been evaluated can be seen in Table III.

C. Test Set

The approach has been evaluated extensively using a state-of-the-art test set created by Schler et. al [5], containing thousands of freely accessible English web blogs. For this evaluation, a subset of approximately 8,000 randomly selected blogs have been used, whereby for each blog entry the gender as well as the age of the composer is given.

TABLE III
PARAMETER SETUP USED FOR THE EVALUATION.

Parameter	Assignment
p, q	$2 - 4$
s_{max}	200
pq_{max}	500
input feature set	occurrence-rate, rank, combined

Regarding the latter, the ages are clustered into three distinct groups, as defined by the original test set [5]: 13-17 (=10s), 23-27 (=20s) and 33-42 (=30s). The five-year gap between each group is thereby added to gain higher distinguishability. The corpus is fairly balanced with respect to gender, but has a majority in the 20s group and a minority in the 30s group. A detailed information about the class distribution is shown in Table IV. Because of the fact that simply predicting the majority class in all cases would lead to an accuracy of, e.g., 53% for `male`, the baseline which should be exceeded is set accordingly to 53% for gender, 46% for age and 25% for gender+age profiling, respectively.

TABLE IV
TEST DATA DISTRIBUTION.

	female	male	sum
10s	18%	19%	37%
20s	21%	25%	46%
30s	8%	9%	17%
Sum	47%	53%	

Each blog consists of at least 200 English words and has been textually cleaned in the original test data, i.e all unnecessary whitespace characters and HTML tags etc. have already been removed. Hyperlinks have been replaced by the word '`urlLink`'. Nonetheless, because this approach depends on the calculation of grammar trees, the latter tags have been manually removed for the evaluation, as the computation of grammar trees would be falsified.

D. Classifiers

Besides the parameter settings the accuracy of the profiling process depends on the classification algorithm that is used in combination with the set of features that are applied. Therefore, to determine the best working algorithm for this approach, several commonly used methods have been tested. Using the WEKA toolkit as a general framework [20], the following classifiers have been utilized: Naive Bayes classifier [21], Bayes Network using the K2 classifier [22], Large Linear Classification using LibLinear [23], support vector machines using LibSVM with nu-SVC classification [24], k-nearest-neighbours classifier (kNN) [25] using $k = 1$ and a pruned C4.5 decision tree [26].

E. Results

All possible settings, i.e., combinations of assignments of p and q with classifiers, have been evaluated on the test set

 ISBN: 978-1-61208-364-3

using a 10-fold cross validation. For each classifier the best results for predicting the gender, age and both gender and age combined are shown in Table V. The detailed results for each feature set is depicted, as well as the concrete sub results for the individual classes. Note that the average value is weighted, i.e., adjusted to the test data distribution.

In general, the results could significantly exceed the corresponding baselines, which manifests that solely the grammar of authors - analyzed with syntax trees and pq-grams - serves as a distinct feature for author profiling.

Despite of the class to predict, the support vector machine framework *LibSVM* and the large linear classification *LibLinear* worked best, whereas the kNN classifier and the C4.5 decision tree produced worse results. Also, the combined feature set using the occurrence-rate and the rank is always inferior to the isolated subsets, which may possibly be correlated to the large amount of features employed with this set (double the size of the other sets).

1) Gender Results: The best result using $p = 2$ and $q = 3$ could be achieved with *LibSVM*, leading to an accuracy of 69%. It utilizes the occurrence-rate feature set, whereby males could be identified with 71%. Although the prediction rate is a little worse than those of other approaches (e.g. [5] achieves 80% over the full test set using several style and content features), the result is promising as it uses and evaluates only the proposed feature and the baseline of 53% could be surpassed clearly.

2) Age Results: Using an almost identical setting, the maximum accuracy of 63% results again from using *LibSVM* and the occurrence-rate feature set (but with $p = 3$ instead of $p = 2$). In general the accuracy for the prediction of the age groups `10s` and `20s` are very solid, but all classifiers have problems predicting the `30s` group. For example, the best configuration achieved a rate of 70% for `10s` and 68% for `20s`, respectively, but could only predict 5% correctly in the eldest group. While the other algorithms could profile the latter class at a higher accuracy, interestingly the Naive Bayes classifier even missed it totally.

A reason for this may be the unbalanced distribution of the test data, which contains only a small amount of `30s` text samples compared to the other groups. It might be the case that the classifiers would have needed more samples to construct a proper prediction model. Even though the unbalanced test set is an immediate consequence of the original test data distribution ([5]), future work should try to create a smaller, but equally distributed test set in order to examine the source of the problems occurring in the `30s` classification.

As with gender, the age results also significantly exceed the baseline of 43%. Like it can be assumed, by taking also other features into account, a higher accuracy can be achieved (e.g. [3] could reach 77% for age profiling).

3) Gender+Age Results: For this problem, the combinations of gender and age, i.e., six classes, had to be predicted. The baseline coming from the majority class `male-20s` is 25% and could also be surpassed using the *LibLinear* classifier. With relatively large structure fragments resulting from the assignments $p = 4$ and $q = 3$, an accuracy of 39% could be achieved using the rank feature set.

TABLE VI
CONFUSION MATRICES OF THE BEST RESULTS FOR GENDER AND AGE PROFILING.

	classified as [%]	
	female	**male**
female	**30.8**	16.1
male	15.0	**38.1**

(a) Gender

	classified as [%]		
	10s	**20s**	**30s**
10s	**25.3**	11.6	0.4
20s	7.8	**37.5**	0.9
30s	1.7	14.3	**0.5**

(b) Age

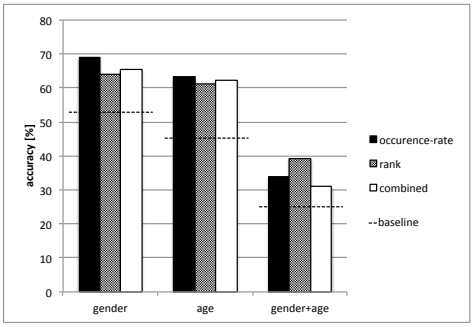


Fig. 3. Summarizing Evaluation Results Using Different Feature Sets.

Due to visibility reasons the details for the individual sub results have been omitted in the table. Nonetheless the experimental data shows that the combined gender and age classification also suffers from predicting the male/female classes of the `30s` age group correctly.

4) Confusion Matrices: A detailed analysis of the best working classifications are shown in the confusion matrices in Table VI. When predicting the gender, the number of false-positives for `male` as well as for `female` are approximately the same. On the other side, the classification of age groups had massive problems concerning the `30s` group, where only 0.5% have been labeled correctly. The majority of this group has been predicted as `20s`, which represents also the majority group of the test set.

As already mentioned, a possible explanation might be the unbalanced test set. This is reinforced by the fact that mostly all false-positives of the `10s` group have also been labeled as `20s`. But what also seems plausible is the hypothesis that the grammar of 13-17 (`10s`) year olds differs significantly from that of 23-27 (`20s`) year olds, where on the other hand the grammatical style of the latter is similar to 33-42 (`30s`) year olds. Intuitively this seems reasonable when looking at sample documents, but future work should investigate further to verify or falsify this assumption.

Summarizing Figure 3 illustrates the evaluation results for all three classification problems using the different feature sets. As can be seen, all baselines could be exceeded.

V. RELATED WORK

Since the advent of the world wide web, offering a huge amount of publicly available documents, the automatic clas-

 ISBN: 978-1-61208-364-3

TABLE V
EVALUATION RESULTS IN PERCENT FOR PROFILING GENDER, AGE AND GENDER+AGE.

			Feature Set									
			Occurrence-Rate			Rank			Combined			
Classifier	**p**	**q**	**female**	**male**	**w. avg**	**female**	**male**	**w. avg**	**female**	**male**	**w. avg**	**Max**
Naive Bayes	4	4	65.9	66.4	66.2	66.4	67.2	**66.8**	65.8	66.1	66.0	66.8
BayesNet	2	4	66.5	67.2	66.8	67.5	68.2	**67.8**	67.4	67.7	67.6	67.8
LibLinear	3	2	61.4	65.7	**63.7**	59.2	64.2	61.8	60.0	64.4	62.4	63.7
LibSVM	2	2	66.5	71.1	**69.0**	61.5	66.3	64.0	62.5	67.8	65.3	**69.0**
kNN	2	2	54.6	62.9	**59.2**	48.6	53.6	51.2	47.7	57.9	53.2	59.2
C4.5 tree	4	2	55.3	61.1	58.4	58.1	62.0	**60.2**	56.4	61.7	59.2	60.2

(a) Results for Gender Prediction.

			Feature Set											
			Occurrence-Rate				Rank				Combined				
Classifier	**p**	**q**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**Max**
Naive Bayes	3	3	39.4	64.6	0.0	52.7	38.2	63.9	0.0	51.9	40.0	65.1	0.0	**53.2**	53.2
BayesNet	2	4	67.6	48.1	39.7	**53.4**	66.7	48.0	39.7	53.1	67.5	47.5	40.3	**53.4**	53.4
LibLinear	2	2	62.3	58.7	24.7	54.8	61.9	55.6	26.1	53.0	63.7	59.1	25.6	**56.6**	56.6
LibSVM	3	2	70.1	68.4	5.0	**63.2**	67.0	66.1	19.9	61.1	68.2	67.5	18.0	62.4	**63.2**
kNN	3	3	54.4	53.3	25.1	48.9	51.2	56.8	27.2	49.8	53.5	56.8	26.5	**50.5**	50.5
C4.5 tree	2	4	52.9	52.3	24.8	48.2	56.1	53.5	26.4	**50.2**	56.9	51.6	24.8	49.3	50.2

(b) Results for Age Prediction.

			Feature Set			
Classifier	**p**	**q**	**Occurrence-Rate**	**Rank**	**Combined**	**Max**
Naive Bayes	4	2	34.8	**35.7**	35.1	35.7
BayesNet	2	4	36.1	**36.4**	36.0	36.4
LibLinear	4	3	33.9	**39.1**	30.9	**39.1**
LibSVM	4	2	**37.2**	34.5	25.8	37.2
kNN	2	2	**32.6**	25.6	25.5	32.6
C4.5 tree	3	3	31.1	**28.2**	27.1	31.1

(c) Results for Combined Gender+Age Prediction.

sification of text has gained more and more interest in the information retrieval field. An often applied concept in order to categorize documents into predefined classes is the utilization of different machine learning algorithms [27], like it is used in this paper. Thereby the problem types are differentiated between single-label and multi-label classification problems, respectively, where the first type assigns only one label for a document (e.g. the gender or age of the author) and the latter type is allowed to assign more labels (e.g. the content type of an article: sports, religion, science, etc.) [28].

Within the single-label text categorization problem the gender and age of the author of a text document has been analyzed frequently. Based on the work of [29] that analyzes the gender of the author and also automatically distinguishes between fiction and non-fiction documents, the web blog corpus used in this approach has been created to classify gender and age based on many style and content features [5]. Here, also blogwords (neologisms) like *'lol'*, *'haha'* or *'ur'* as well as the frequency of hyperlinks have been analyzed. An extension that additionally attempts to classify the language and personality (e.g. neuroticism or extraversion) of a writer has been proposed in [3] by utilizing taxonomies of POS tags combined with other style and content-specific features. Two new feature sets using POS tag patterns are proposed in [30] to enhance current state-of-the art classification approaches.

An interesting approach that also analyzes web blogs is presented in [6]. Besides commonly used features in the field of text categorization the focus has been laid on blog-specific features such as the usage of background colors, emoticons, punctuation marks or fonts. It is shown that the prediction of gender can be enhanced by using these features.

English emails have been profiled into ten classes including gender, age, geographic origin or level of education as well as into five psychological traits in [31]. The authors use several character-level, lexical and structural features and report a similar accuracy for gender classification as the outcome presented in this paper, but show a worse result for age classification (note that emails are typically significantly shorter than blogs).

With the recent rise of social media networks, also content such as chat lines, Facebook postings or tweets have been analyzed and automatically profiled. It is shown (e.g. in [4] or [32]) that a well-defined set of style and content features can be used to expose meta information of chat logs, also in other languages such as Spanish. Nevertheless, the authors in [33] show that the application of common text categorization techniques using natural language processing is challenging - but possible - when facing highly limited data sets. It is demonstrated that even for text samples containing only approximately 12 tokens, the classification of gender and age is feasible.

A related problem in the field of forensic linguistics has recently been investigated in a scientific workshop [11]: Given

 ISBN: 978-1-61208-364-3

the task to automatically expose sexual predators from chat logs, several approaches showed promising results.

The analysis of grammar trees with pq-grams has also been used in previous work, where it has been shown that the grammar of authors is also a feasible criteria to intrinsically expose plagiarism [12] and to correctly attribute authors to unlabeled text documents [13].

VI. Conclusion and Future Work

In this paper, a novel feature that can be used to automatically profile the author of a text document is presented. Based on full grammar trees, it utilizes substructures of these trees by using pq-grams. State-of-the-art machine learning algorithms are finally applied on pq-gram profiles to learn and predict the gender and age of the originator. An extensive evaluation using a state-of-the art test set shows that pq-grams can be used as significant features in text classification, whereby gender and age can be predicted with an accuracy of 69% and 63%, respectively. With respect to the fact that the experiment in this paper solely uses the presented feature, the results are promising.

Evaluation results showed that the approach has problems predicting the `30s` age group. Although hypothesis explaining the problem have been stated, they should be verified or falsified in detail by utilizing a different test set.

In order to build a reliable text classification approach, the grammar feature should be combined with other commonly used style and content feature sets in future work. Besides the utilization of common lexical, syntactic or complexity features, the usage of vocabulary or neologisms should be considered, especially when analyzing online content. Moreover it should be investigated whether the proposed feature is also applicable to shorter text samples such as chat logs or even single-line Twitter postings.

Research should finally also examine whether pq-gram profiles are also exploitable to other languages, especially as syntactically more complex languages like German or French may lead to even better results due to the higher amount of grammar rules available.

References

[1] "Statistic Brain Research Institute," http://www.statisticbrain.com/social-networking-statistics, visited February 2014.

[2] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically Profiling the Author of an Anonymous Text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[4] L. Flekova and I. Gurevych, "Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media," *Notebook Papers of CLEF 13 Labs and Workshops*, 2006.

[5] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of Age and Gender on Blogging," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 199–205.

[6] X. Yan and L. Yan, "Gender Classification of Weblog Authors," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 228–230.

[7] J. Noecker, M. Ryan, and P. Juola, "Psychological Profiling Through Textual Analysis," *Literary and Linguistic Computing*, 2013.

[8] J. Gibbons, *Forensic Linguistics: An Introduction to Language in the Justice System*. Blackwell Pub., 2003.

[9] S. Nirkhi and R. Dharaskar, "Comparative Study of Authorship Identification Techniques for Cyber Forensics Analysis," *International Journal*, 2013.

[10] E. E. Abdallah, A. E. Abdallah, M. Bsoul, A. F. Otoom, and E. Al-Daoud, "Simplified Features for Email Authorship Identification," *International Journal of Security and Networks*, vol. 8, no. 2, pp. 72–81, 2013.

[11] G. Inches and F. Crestani, "Overview of the International Sexual Predator Identification Competition at PAN-2012," in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[12] M. Tschuggnall and G. Specht, "Using Grammar-Profiles to Intrinsically Expose Plagiarism in Text Documents," in *NLDB*, 2013, pp. 297–302.

[13] ——, "Countering Plagiarism by Exposing Irregularities in Authors Grammars," in *EISIC, European Intelligence and Security Informatics Conference, Uppsala, Sweden*, 2013, pp. 15–22.

[14] N. Augsten, M. Böhlen, and J. Gamper, "The pq-Gram Distance between Ordered Labeled Trees," *ACM Transactions On Database Systems (TODS)*, vol. 35, no. 1, p. 4, 2010.

[15] S. Helmer, N. Augsten, and M. Böhlen, "Measuring Structural Similarity of Semistructured Data Based on Information-theoretic Approaches," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 21, no. 5, pp. 677–702, 2012.

[16] The Apache Software Foundation, "Apache OpenNLP," http://incubator.apache.org/opennlp, visited February 2014.

[17] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03, Stroudsburg, PA, USA, 2003, pp. 423–430.

[18] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, pp. 313–330, Jun. 1993.

[19] E. Stamatatos, "Intrinsic Plagiarism Detection Using Character n-gram Profiles," in *CLEF (Notebook Papers/Labs/Workshop)*, 2009.

[20] M. Hall et al., "The WEKA Data Mining Software: an Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[22] G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks From Data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library For Large Linear Classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[25] D. Aha and D. Kibler, "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[26] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning, 1993.

[27] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[28] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[29] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.

[30] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," in *Proceedings of the 2010 Conference on Empirical Methods in NLP*. Association for Computational Linguistics, 2010, pp. 207–217.

[31] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson, "Author Profiling for English Emails," in *Proceedings of the 10th Conference of the Pacific Assocation for Computational Linguistics*, 2007, pp. 263–272.

[32] M. Meina et al., "Ensemble-Based Classification for Author Profiling Using Various Features," *Notebook Papers of CLEF*, 2013.

[33] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting Age and Gender in Online Social Networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011, pp. 37–44.

 ISBN: 978-1-61208-364-3

Semantic Mining and Analysis of Heterogeneous Data for Novel Intelligence Insights

Rick Adderley, Patrick Seidler
A E Solutions (BI) Ltd.
Badsey, UK
rickadderley@a-esolutions.com
patrickseidler@a-esolutions.com

Atta Badii, Marco Tiemann
University of Reading
Reading, UK
atta.badii@reading.ac.uk
m.tiemann@reading.ac.uk

Federico Neri, Matteo Raffaelli
Synthema srl
Pisa, Italy
federico.neri@synthema.it
matteo.raffaelli@synthema.it

***Abstract*—This paper describes the implementation of a Data Mining (DM) system under the EU FP7 Security Research Project Multi-Modal Situation Assessment & Analytics Platform (MOSAIC). The system aims to enable the part-automatic detection and recognition of crime threats in uncertain environments. It facilitates the automatic retrieval of intelligence data providing deep semantic information access and dynamic classification features for distributed data sources, such as Policing legacy databases, Police text documents and free text database fields. A specific pipeline of linguistic processors that share a common knowledge base on crime patterns has been created to retrieve entities and events from text documents and websites. Structured and unstructured data retrieved from the individual data sources are integrated in a semantically query-able unified data representation using specific ontological models. A domain specific entity resolution module ensures the resolution of conflicting and misleading identities to enable data retrieval and fusion from disparate data sets. As criminal network analysis depicts a major part of the intelligence process, specific measures and algorithms have been developed to support analysts in retrieving, analysing, and disrupting criminal networks.**

Keywords-Data mining; text mining; entity recognition and resolution; social and criminal network analysis; semantic interoperability.

I. INTRODUCTION

Despite huge progress in Data Mining (DM) in the last decade, a gap remains between DM technologies and the actions that are taken upon knowledge creation based on them [1]. The most labour intensive and at the same time most expensive parts of mining projects are generally concerned with data pre-processing, i.e., with preparing data in such a way to be able to further examine data for meaningful information [2]. The fact that data pre-processing is often embedded in a large amount of domain knowledge might explain the slow progress in the area which is to be retrieved repeatedly for each project and is then often encoded in low level system parts such as in Structured Query Language (SQL) statements.

Therefore, the efficiency of any institution still relies heavily on the human factor to close this gap [3], limiting the DM process and the applicability of DM itself. DM can, for example, reveal all the data to create an offender profile, but the existing systems are often not able to sufficiently link known profiles with unsolved crimes, i.e., other forensic evidence such as the method of offending [4]. This lack of sufficiently enriched data in some parts of the DM process often creates a knowledge gap that hinders effective and targeted intervention, but leaves analysts with labour-intensive bottlenecks [5].

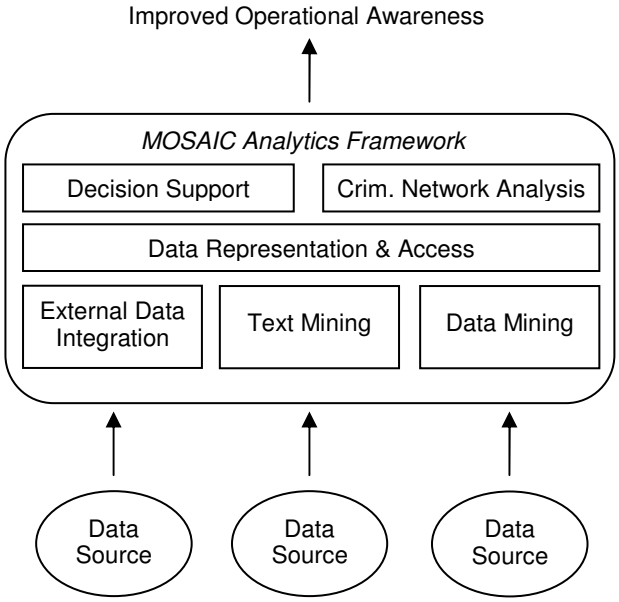


Figure 1. Overview diagram of the MOSAIC Analytics Framework.

The primary objective of the Multi-Modal Situation Assessment & Analytics Platform (MOSAIC) (see Fig. 1) is to improve the targeted surveillance of and intervention into complex systems of criminal behaviour by combining intelligence to provide a decision support system for the relevant authorities. The system facilitates the correlation of data from disparate sources into a semantically operational system to form contextual and valuable information – the information whole being greater than the sum of its parts, and thus to enable targeted surveillance. The system uses a loosely coupled system architecture where sensing and analysis components communicate through Web Services and exchange data through a central system.

 ISBN: 978-1-61208-364-3

Mining and analysis of different kinds of data including data taken from legacy databases and heterogeneous sources of text from sanitised Police reports, from free text database fields, and from WWW public sources allows the user to integrate those data within a unified framework in order to be able to conduct social and criminal network analysis. The framework has been designed to be compatible with existing procedures, tools and legacy systems used by Police forces within the European Union.

In this paper, we begin by describing the MOSAIC data sources. We then outline the system architecture, including the analysis components and the semantic interoperability approach. Finally, we report on a preliminary case study and user experiments.

II. DATA SOURCES

Police repositories hold millions of entries on crimes, offenders, and other intelligence. For the purpose of the project and the mining of those data by the DM Component, a representative MOSAIC legacy database has been created and sanitised using original Police data, representing data on Nominals, Crimes, Intelligence, Automatic Number Plate Recognition (ANPR), and Stop & Search.

The Text Mining (TM) Component takes as input two types of data, whereas anonymised entities within the documents have been reproduced in the MOSAIC legacy databases. The first type is two case studies provided by Police partners spanning a wide variety of Police disciplines. For example, one of the case studies is an 85 pages case document of a missing person investigation, which took place over three months. The second type constitutes long text fields mainly in the Intelligence table. This free text cannot be analysed with standard DM algorithms, but will be passed to the TM Component to retrieve additional information on entities and their relations.

III. DATA PREPARATION

A. Entity Resolution Component

Poor data quality is a constant issue in Policing systems. Many errors are introduced when data is entered into the systems. Moreover, we cannot expect that data are always easily identifiable through global unique identifiers. If an organisation or institution is not able to identify unique objects, suboptimal decisions will be the result.

Using the Apache Lucene framework as a backend indexing tool, an entity resolution module is being developed in order to resolve those issues. A decision engine uses scenarios that contain predefined probability levels for matches on specific database field types which are used to calculate the final probability that a match has been found. Matching fields are thereby compared using the Bayes function using as input the pre-determined probability for each field. Finally, fuzzy string matching is to be introduced providing Metaphone [6] and Soundex [7] string matching options.

Preliminary results on the algorithm's performance and its accuracy in correctly matching entities show a minimum accuracy of 65% for 90% of test runs, whereas several combinations show an accuracy of up to 88% when compared to the Gold Standard. In comparison, compilation of the Gold Standard took the analyst 1 ½ days, involving handcrafting 1002 offender records into sets containing the same individual.

B. Data mining workbench

During their work, analysts iterate through a set of unspecified tasks in no particular order and as needed. The widely used National Intelligence Model (NIM [8]) does not provide a structured approach to those tasks.

We formalise analysis tasks using the Cross Industry Standard Process for Data Mining (CRIPS-DM [9]) model in conjunction with the intelligence cycle. DM algorithms for the preparation and analysis of data are being implemented into an integrated MOSAIC DM workbench to assist analysts in manipulating data without the hassle of having to access disparate systems. Using the workbench, analysts will be able to use data search and linking, exploration, modelling and visualisation capabilities through a process of interconnected nodes. The approach taken accommodates for the various possible working environments and data requirements in which the final system could be applied. The resulting DM processes will be reusable and can be re-run any time taking into account data that have newly arrived in the system.

IV. ANALYSIS COMPONENT

Data analysis inside law enforcement has remained a time-consuming process, with technical support restricted to a large number of unconnected systems tools lacking support in the provision of actionable intelligence. It has further been argued that an operational gap exists between intelligence analysis and operational policing, with advanced technology often used to manage offenders rather than providing insights on criminal behaviour and possible interventions [5].

A. Data analysis algorithms

To fulfil a request for information the analyst performs a query through disparate systems, researching, e.g., names, addresses and telephone numbers. Intelligence logs are individually read by the analyst and recorded into a spreadsheet and/or directly into one of the existing link visualisation tools. This largely manual acquisition and preparation of data is time-consuming and prone to error as the amount of data to search exceeds the humanly comprehensible limit [10].

The MOSAIC system offers DM support that has been tailored to analysts' needs regarding their work tasks, processes, and their needs for actionable operational intelligence. The main focus is put on the creation of such results that are immediately and easily applicable inside the intelligence cycle.

 ISBN: 978-1-61208-364-3

- Offender mining and automatic assignment of priorities to offenders: Track prioritised criminal behaviour and enable law enforcement to allocate responsive actions in order to meet Force priorities.
- Identification of crime series and mapping of known offenders to unsolved crimes: Application of self-organising maps to link spatial, temporal and modus operandi (MO) and overlay of offender data onto clusters of similar crimes thereby suggesting possible involvement.
- Identification of criminal roles: Create offender profiles, group offenders by their profile, and apply a K-means clustering algorithm to determine the prominent group for all offenders.

B. Text Mining Component

Data which have been retrieved from free text database fields, from the document repository and from the World Wide Web (WWW) will be converted into Clean TXT format, processed by the Natural Language Processing (NLP) engine and indexed.

In MOSAIC, TM and entity extraction are going to be applied through a pipeline of linguistic and semantic processors that share a common knowledge and a common ontology. A crime ontology and a domain specific knowledge base with crime patterns, abbreviations, technical terms and terms relationships mainly extracted from the sanitised Police reports are created. This shared ontology and knowledge base guarantees a uniform interpretation layer for the diverse information from different sources.

The TM process is implemented by the following steps:

- Morpho-syntactic or Part-of-Speech (POS) tagging
- Multiword tagger (MWT)
- Word-sense disambiguation (WSD)
- Named-entity recognition (NER)
- Semantic role labelling (SRL)
- Entity Relationship extraction

At the heart of the morpho-syntactic analysis module, which aims at identifying the part-of-speech (POS tagging), is McCord's theory of Slot Grammar [11][12]. The module will analyse each sentence, cycling through all its possible constructions and trying to assign the context-appropriate meaning – the sense – to each word by establishing its context. The parser – a bottom-up chart parser – employs a parse evaluation scheme used for pruning away unlikely analyses during parsing, as well as ranking the final analyses. It will build the syntactical tree incrementally. Multi-word combinations are then identified and ambiguous terms disambiguated depending on the syntactic and semantic context, by considering super-subordinate related concepts.

These two modules are closely related to named-entity recognition. Extensive effort is being spent on the identification of pre/suffixes, specific linguistic patterns and specific data formats for the English language in order to recognise the following entities in texts: dates, addresses, person names, locations, license plate numbers, brands, web entities (web addresses, Internet Protocol (IP) addresses, email addresses, etc.), bank accounts and phone numbers. Entities are reduced to their semantic roles (agent, predicate, theme, recipient, time and location; in simpler terms: *who* does *what* to *whom*, *how*, *when* and *where*), identified as a result of the dependency parsing. The NLP engine will then be able to extract entity relationships from a text. Heuristic algorithms are being implemented in order to extract all kinds of relationships between the entities mentioned above.

C. Social and Criminal Network Analysis and Visualisation Component

Empirical research has shown that people who have a propensity to commit crime rarely work in isolation, but in a group of associates who have differing skills and interests to complement the activities of individuals or sub groups within their criminal network [13][14][15]. As security and law enforcement resources are not unlimited, prioritisation decisions have to be made for policing and investigative effort. It is, therefore, highly desirable to be able to identify, characterise and rank the networks which are operating within an area so as to identify, and prioritise for further investigation, those networks and individuals within them that are most significant in terms of who are causing the most harm.

The aim of the MOSAIC criminal network analysis and visualisation component is to support law enforcement in continuously grasping a full picture of current criminal activity and close the gap towards previsional systems by evaluating beforehand the impact of decisions. Results shall enable agencies to create improved intelligence products on effective ways for effectively disrupting criminal activity.

To create networks from structured data, we use the approach outlined by Adderley et al. [16]. The algorithm identifies all of the criminal networks that are present in a dataset and prioritises those that are causing most harm to the community based on a crime scoring mechanism. We further provide algorithms that combine network topological measures with domain based weighting scores, and enable the identification of criminal roles, sub group and network themes, and the running of network robustness simulation tests against target law enforcement interventions. Visualisation will be achieved by presenting the network structure in a 3D environment with textual statistics and data overlays.

V. SEMANTIC INTEROPERABILITY

When extracting and analysing data from multiple and quite distinct data sources, integrating the gathered and extracted data and information from these sources becomes a significant issue: in current practice, police intelligence analysts need to gather the available information from a multitude of completely separate systems with different output formats and to then manually create unified

 ISBN: 978-1-61208-364-3

representations based on the data gathered - clearly not an efficient procedure. To reap the benefits of automated data analysis on a large scale, data must be made accessible through a single system. And to be able to combine different sources of information in order to find previously "unfindable" connections, the data to be integrated must "speak the same language". The available information must be made semantically interoperable. In MOSAIC, semantic interoperability involves three main aspects: the definition of a semantic domain model which can represent the available information while preserving its meaning; the development of a system that organises the available information using the developed model that makes it accessible; the connection to the individual data sources and to any further "consumers" of the data.

The world model for MOSAIC is being defined as an Ontology Web Language (OWL)-Lite model [17]. This model represents actors, objects, actions and other relevant information types as subject – predicate – object triples that establish object types, their properties and their relations to other object types. Data gathered is added to this model as instances of the defined types with the relevant properties and relations to other instances, thus populating the data model. The data model has been developed in collaboration with police partners using real-world scenarios and refined given the available data in order to retain conceptualisable and groundable concepts only [18].

A semantic data store will be used to manage the processes of creating, reading, updating and deleting instance data within the semantic representation model. MOSAIC uses a data store implementation that stores data triples – a triple store. The MOSAIC data store is based on the Apache Jena project and uses the core Apache Jena components for data storage and access as well as the Apache Fuseki Web Service front end. Data in the MOSAIC data model can be queried and updated using the Simple Protocol and Resource Description Framework (RDF) Query Language (SPARQL) [19], which provides an equivalent to SQL for accessing and updating data that is stored in the form of triples. The data store implementation has been extended with additional MOSAIC-specific features such as the ability to subscribe with queries in order to receive notifications when new relevant data are added to the data store.

Semantic interoperability can only be achieved when the data of interest is adequately integrated into the MOSAIC data store; to this end, data importers have been integrated as data store plugins. These importers provide Web Service endpoints to which source data can be sent in their native formats. The data are then analysed for consistency, converted in terms of terminology and representation via mediator components and added to the MOSAIC data store.

The MOSAIC data representation and data store system allows analysts and operators to query a single data representation for information across information provided by all of the data sources described. The ontology used in MOSAIC extends this by allowing users to make use of the knowledge encoded in the ontology while querying it – a trivial example for this is the ability to query for persons involved in violent crimes without having to enumerate the individual identifiers for violent crimes as might be necessary in a conventional SQL database.

The semantic representation of data can also be used to reason using the world model and to define complex events that may be hard to spot by human operators but that can be defined as sets or sequences of events that taken together either lead to new information or should trigger a specific (re-)action [20]. A reasoning and rule engine that is suitable to work with the triple data representation and can describe groups and sequences of observations entered to be matched and actions to be taken with the help of the MOSAIC system is currently being integrated.

VI. EVALUATION

A preliminary case study has been conducted. The goal for users was to automatically identify the network(s) with the highest police force priority, the most prolific offenders inside the network, as well as appropriate interventions.

The analyst extracts data with the DM workbench and creates a problem profile that will be enhanced as more of the data is understood. To increase data quality, offender identities from the joined data set entries are resolved before starting a DM process. The output contains 995 unique identities compared to 1505 unique ids in the original data set. A DM process was then developed and applied which retrieves police force priority scores, crime roles and travel distances to develop a criminal profile for each offender.

Applying the network generation, 568 networks were identified from the dataset for networks with two degrees of freedom in 3.5 seconds. Respective generation of networks with 3 degrees of freedom took 69 seconds to run, and 165 seconds, while as a rule analysts will use two degrees of freedom in most circumstances as those cover the most common crimes and criminal networks for most Police areas.

Utilising offenders' criminal profiles, the highest ranked network containing 57 unique offenders was identified and further analysed. Topological measures are added to each offender's criminal profile and we retrieve a final prioritised list of offenders (see Table I) which facilitates decision making in targeting the appropriate person(s).

TABLE I. TOP 3 CRIMINAL PROFILES IN DATA SET

Id	Role	Harm	Distance	Inform. Control	Access	Activity	Score
1	Burglary	600	Compact	Controller	Best	Active	30.49
2	Burglary	600	Compact	Some	Best	None	27.49
3	Violence	360	Compact	None	Average	None	17.91

We further evaluate effectiveness of interventions on the network. Based on degree centrality and domain scores, in each step the vertex with the highest overall rank amongst all vertices is selected for sequential removal, compared with a

 ISBN: 978-1-61208-364-3

random removal approach. Testing the network for its robustness based on the largest remaining component [21], results show that by removing only the top two offenders, we are able to disrupt this specific network by 70% (see Fig. 2).

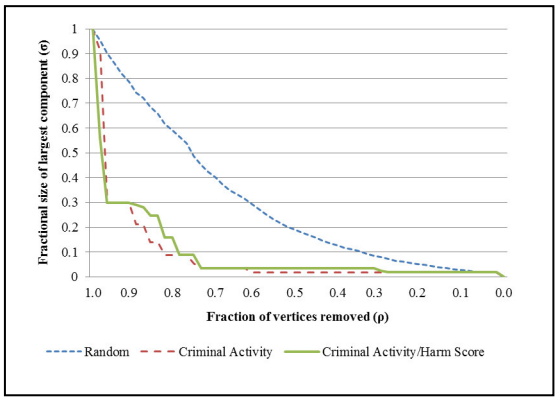


Figure 2. Robustness of criminal network under sequential attacks.

To provide explicit prototype evaluations, additional experiments were conducted involving seven domain experts who were asked to perform several sub tasks under two experimental conditions: 1) automated visualisation and analysis; 2) automated visualisation and manual analysis. The average total score provided by the experts was 13.86 from a maximum of 20, resulting in a 69.3% satisfaction level. Comments regarding how the prototypes could be improved were also provided.

VII. CONCLUSION AND FUTURE WORK

This contribution described interim results of the MOSAIC project. It is in particular concerned with showing how data analysis and information mining techniques are applied in order to extract useful information from large amounts of noisy data, and how the extracted data can be represented and made accessible using a semantic integration system.

Future work in the project will involve the effective presentation of extracted information and reasoning over the extracted data with in order to aid in decision making processes based on the information extraction and analysis processes outlined in this contribution.

ACKNOWLEDGMENT

This work is supported by the European Commission in the 7th Framework Programme, within the Security Research Theme, under Grant 261776 with the acronym MOSAIC.

REFERENCES

[1] P. Domingos, “Toward knowledge-rich data mining,” Data Mining and Knowledge Discovery, vol. 15, no. 1, 2007, pp. 21–28.

[2] L. Pipino and D. Kopcso, “Data Mining, Dirty Data, and Costs,” in Proceedings of the Ninth International Conference on Information Quality, MIT, 2004, pp. 164–169.

[3] Z. T. Kardkovács, “Business Intelligence and Data Mining,” in Research and Development in E-Business through Service-Oriented Solutions, K. Tarnay, S. Imre, and L. Xu, Eds. IGI Global, 2013, pp. 57–70.

[4] R. Adderley, “Exploring the Differences Between the Cross Industry Process for Data Mining and the National Intelligence Model Using a Self Organising Map Case study,” in Business Intelligence and Performance Management, P. Rausch, A. F. Sheta, and A. Ayesh, Eds. Springer London, 2013, pp. 91–105.

[5] P. Seidler and R. Adderley, “Criminal network analysis inside law enforcement agencies – a data mining system approach under the National Intelligence Model,” IJPSM, vol. 15, no. 4, 2013, pp. 323–337.

[6] L. Philips, “Hanging on the Metaphone,” Computer Language, vol. 7, no. 12, 1990, pp. 39-43.

[7] M. K. Odell, "The profit in records management," Systems, vol 20, no. 20, 1956.

[8] National Criminal Intelligence Service, “The National Intelligence Model.” National Criminal Intelligence Service, 2000.

[9] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” Journal of Data Warehousing, vol. 5, no. 4, 2000, pp. 13–22.

[10] Y.-W. Si, S.-H. Cheong, S. Fong, R.P. Biuk-Aghai, and T.-M. Cheong, “A layered approach to link analysis and visualization of event data,” in Seventh International Conference on Digital Information Management, 2012, pp. 181–185.

[11] M. C. McCord, “Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars,” in Proceedings of the International Symposium on Natural Language and Logic, London, UK, 1990, pp. 118–145.

[12] M. C. McCord, “Slot Grammars,” Comput. Linguist., vol. 6, no. 1, 1980, pp. 31–43.

[13] E. Patacchini and Y. Zenou, “Juvenile delinquency and conformism,” Journal of Law, Economic, and Organization, vol. 28, 2012, pp. 1–31.

[14] M. Warr, Companions in Crime. Cambridge Univ Pr, 2002.

[15] D. L. Haynie, “Delinquent peers revisited: does network structure matter,” American Journal of Sociology, vol. 106, 2001, pp. 1013–1057.

[16] R. Adderley, A. Badii, and C. Wu, “The Automatic Identification and Prioritisation of Criminal Networks from Police Crime Data,” in Intelligence and Security Informatics, vol. 5376, D. Ortiz-Arroyo, H. Larsen, D. Zeng, D. Hicks, and G. Wagner, Eds. Springer Heidelberg, 2008, pp. 5–14.

[17] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, “From SHIQ and RDF to OWL: the making of a Web Ontology Language,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 1, no. 1, pp. 7–26, 2003.

[18] A. Jakulin and D. Mladenic, “Ontology Grounding”, in Proc. 8th Intl. Multi-Conf. Information Society, 2005, pp. 170-173.

[19] J. Pérez, M. Arenas, and C. Gutierrez, “Semantics and Complexity of SPARQL,” in The Semantic Web - ISWC 2006, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, Eds. Springer Berlin Heidelberg, 2006, pp. 30–43.

[20] J. Z. Pan, “A Flexible Ontology Reasoning Architecture for the Semantic Web,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, 2007, pp. 246–260.

[21] B. Keegan, M. A. Ahmed, D. Williams, J. Srivastava, and N. Contractor, “Dark Gold: Statistical Properties of Clandestine Networks in Massively-Multiplayer Online Games,” presented at the 2010 IEEE Second International Conference on Social Computing, Los Alamitos, CA, USA, 2010.

 ISBN: 978-1-61208-364-3

Wireless Transmission of Stereo Images and its Disparity Levels

Apurva Naik, Keshav Velhal, Kunal Shah, Pratik Raut, Arti Khaparde
Department of Electronics and Telecommunication
Maharashtra Institute of Technology
Pune, Maharashtra, India
Emails :{apurva.naik, arti.khaparde} @mitpune.edu.in; {keshav.velhal, pratikraut15.11}@gmail.com; {kunds18}@yahoo.com

***Abstract*— One of the promising application of wireless transmission is in Computer Vision. Real-time stereo images are captured using camera and transmitted to another system by using ZIGBEE wireless module. Before transmission, image pixels are grouped to form packets. These packets when received at the receiver end are recovered, and a 3-D image is generated. Also at the transmitter, images are segmented by using DPSO (Darwinian Particle Swarm Optimization). Segmented images are given to the line growing algorithm. Depth levels are estimated with the help of disparity values obtained from the disparity algorithm. These depth levels are transmitted through ZIGBEE module to another system. Depth levels received are used to control a ROBOT. This proposal is a prototype which can be implemented for industrial applications. The present paper deals with the Transmitter-Receiver link for stereo images and movement of ROBOT proportional to estimated depth levels.**

Keywords-ZIGBEE; Darwinian Particle Swarm Optimization; ROBOT.

I. INTRODUCTION

A two-dimensional camera image does not give information about depth levels. However, information about depth is required in several applications such as, satellite imaging, robotic vision, target tracking and automatic map making. Stereo matching is used to extract depth information from images [1]. These estimated depth levels can be used to control the movement of a ROBOT that can be used in robotic vision applications. Until recently, stereography was used either for entertainment purpose or DEM (Digital Elevation Model) for depth analysis of sea bed as, no evidence was found in literature on wireless transmission link for transmission of disparity levels. But this novel approach will help us to control the unmanned vehicle to perform the numerous tasks in medical, mining applications and in volumetric analysis of water reservoirs, etc., which requires the knowledge of depth. For example, one of the applications that can be developed is for computer-aided surgery. Images can be captured with help of stereoscopic endoscope. These images can be transferred to the control room. By doing an analysis and using depth information, the surgeon can instruct a ROBOT to perform certain tasks.

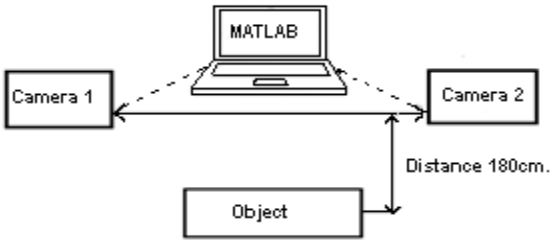


Figure 1. Block diagram of camera setup.

For capture of real-time stereo images, the distance between two webcams should be at least 6cm [7], which is approximately equal to the distance between centers of human eyes, i.e., 2.5inches, as shown in Fig. 1. The camera should be placed at a distance of 1.8m from the object to be captured.

For estimation of the focal length camera calibration, the toolbox of MATLAB [3] has been used. For estimating the focal length of the camera, four images of a chess board, each image having different orientation from the other were taken. A procedure [3] was followed and the focal length obtained was approximately 1300 pixels. Real-time images were captured through camera and processed using image acquisition toolbox of MATLAB. Pixel values of the images were grouped into packets of 2000 pixels and transmitted using AT transmission mode of ZIGBEE. Also, disparity estimation was carried out and depth levels were calculated. These depth levels were also transmitted using ZIGBEE. The maximum value of each depth level received was used to control the ROBOT movement. At the receiver end, a good quality 3-D image was reconstructed.

The rest of this paper is organized as follows. Section II describes the experimental setup. Section III describes the ZIGBEE module and its protocol, and ROBOT control. Section IV gives details of results. Section V concludes the paper.

II. EXPERIMENTAL SETUP

Experimental setup consists of 2 webcams, 2 general purpose PCs, 1 PC controlled wired ROBOT, 2 ZIGBEE modules (one coordinator node and one router node). ZIGBEE modules have been used for transmission of real-time images and depth values. In the present setup, ZIGBEE module of DIGI Company (XBee RF Modules) is used. ZIGBEE standard operates on the IEEE 802.15.4 physical

 ISBN: 978-1-61208-364-3

radio specification and operates in unlicensed bands including 2.4GHz, 900MHz and 868MHz. Each ZIGBEE module is connected to a PC via a Serial to USB Convertor for communication with MATLAB program. MATLAB has been used to encode the image data, and then transmit data to the router nodes [4]. At first, the image at coordinator node is divided into small packets and then these packets are transmitted. The router node receives the image data, in form of packets and then MATLAB is used to decode the image data. This image data is used to generate a 3-D image. Secondly, captured real-time stereo images are segmented using Darwinian Particle Swarm Optimization [2]. The Segmented images are given to the disparity algorithm to estimate the depth values. The coordinator node of ZIGBEE module sends this depth data directly. Another router node receives the depth data, which is then decoded and these decoded values are used to control the ROBOT. Block diagram of system implemented for wireless transmission is shown in Fig. 2.

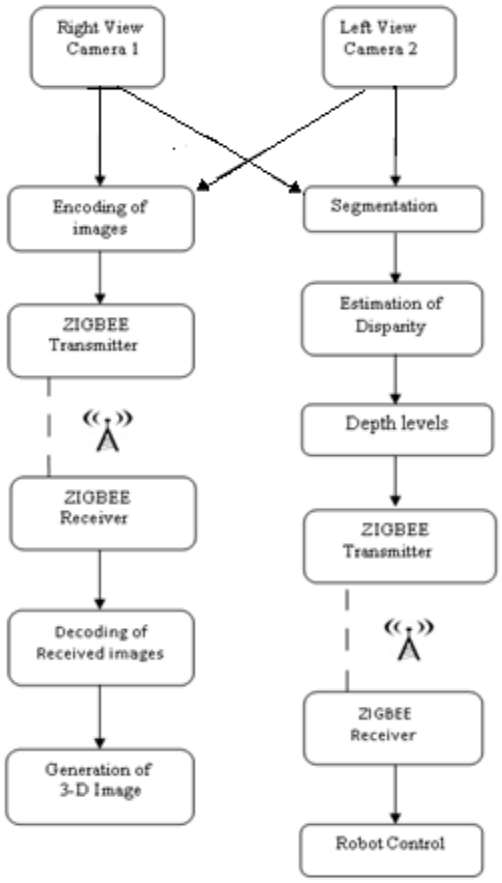


Figure 2. System used for wireless transmission.

III. ZIGBEE

The ZIGBEE Alliance [6] is a consortium of over 90 companies that is developing a wireless network standard for commercial and residential control and automation applications. Wireless communication standards focusing on high speed and long range have been applied for cellular and local area data networks. Transmission of images by using Bluetooth network had been tried, but Bluetooth-based networks can cover the distance up to 10m, while ZIGBEE based networks can be used up to 100m. Bluetooth takes three seconds to join a network while ZIGBEE joins a network in 30 milliseconds [6].

The Alliance has recently released its specifications for a low data rate on wireless network. The design goals for the network have been driven by the need for a Machine-to-Machine (M2M) communication of small simple control packet and sensor data, and a desire to keep the cost of wireless transceivers to a minimum. ZIGBEE is a wireless technology developed as an open global standard to address the unique needs of low-cost, low-power wireless M2M networks, and it currently uses IEEE 802.15.4 MAC and PHY layers, as shown in Fig. 3 [5]. ZIGBEE generally uses a single channel for data transmission. A ZIGBEE module has three nodes, namely, coordinator node, a router node, and an end device node. End-device nodes communicate with each other through a coordinator node. A coordinator node is responsible for starting the network and for choosing certain key network parameters. The end-device nodes not only communicate with the coordinator node but also communicate with every router node. However, the router nodes processing a routing function cannot directly communicate with each other; they can communicate only with coordinator [5]. ZIGBEE network has three modes of transmission, namely, AT (by default), API and API with escape character. In the AT (Transparent Mode), data coming into the Data IN (DIN) pin is directly transmitted over-the-air to the intended receiving radios without any modification. API (Application Programming Interface) mode is a frame-based method for sending and receiving data to and from a serial UART (Universal asynchronous receiver/transmitter). API with escape character is an extended version of API which is used to prevent data loss in noisy environments. Both API and API with escape character are used to insure secure communication. In this setup AT (Transparent Mode**)** mode of transmission has been used as it is easy to configure ZIGBEE in this mode and currently secure communication is not considered in the present prototype.

TABLE I. HARDWARE SPECIFICATIONS

ZIGBEE module

- Operating frequency: 2.4GHz.
- Low cost wireless module.
- Data rate: 250Kbps.
- Operating range: 100ft (30m).

Wireless camera

- Connection Type – Corded USB.
- USB Type –High Speed USB 2.0.

 ISBN: 978-1-61208-364-3

A. ZIGBEE Protocol

ZIGBEE is best described by referring to the 7-layers of the OSI model [8] for layered communication systems. The Alliance specifies the bottom three layers (Physical, Data Link, and Network), as well as Application Programming Interface (API) that allows end developers the ability to design custom applications that uses the services provided by the lower layers. Fig. 3 shows the architecture adopted by the ZIGBEE alliance [5].

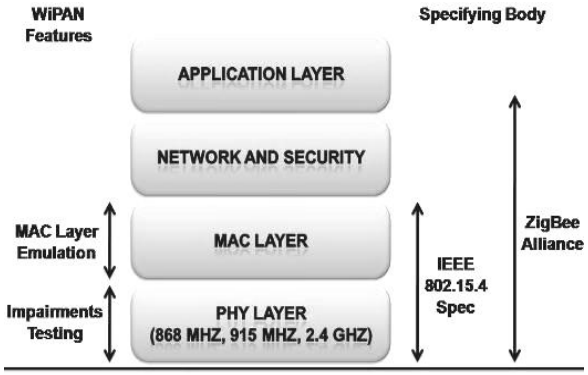


Figure 3. ZIGBEE stack [5].

B. Limitations of ZIGBEE protocol

The 2.4GHz band provides the highest bit rate of 50 Kbps in IEEE 802.15.4 PHY specification. The physical layer supports transfer of only small sized packets which is limited to 127 bytes. Due to overhead at the network, each packet may contain at most 89 bytes for application data. This leads to loss of data during transmission. Therefore, there is a need for fragmentation of bit streams larger than 89 bytes. A flow-control mechanism is also needed to acknowledge and request retransmission of missing fragments above the network layer [5].

C. Transmission of image through ZIGBEE

If a large number of pixel values of an image are transmitted by using ZIGBEE then there is a loss of data in an abrupt manner at the receiving end. For this, the data needs to be fragmented. In this case an image of size 115 X 132 was transmitted using ZIGBEE. An image of size 115 X 132 has 15180 pixel values. The image is fragmented into small packets and each packet contains approximately 2000 pixel values. For a complete transmission of the image, eight packets are required. Since each packet is transmitted separately, there is an increase in time taken for transmission of the complete image.

D. Control of ROBOT by using depth information

The depth levels estimated from disparity data are transmitted through ZIGBEE module. The depth levels received by the receiver are used to control the ROBOT. Binary data for a specific time delay (depending upon the depth levels) is sent from parallel port of the computer to ROBOT, according to which it covers a specific distance. The data given to a parallel port from MATLAB at data pins (D4-D7) goes to the octal buffer IC 74LS244 through db-25 which is used to reduce DC loading. The output of the 74LS244 IC is fed to the motor driver IC L239D which controls the rotation of the motor, which in turn causes the movement of ROBOT. The complete control action of ROBOT is shown in Fig. 4.

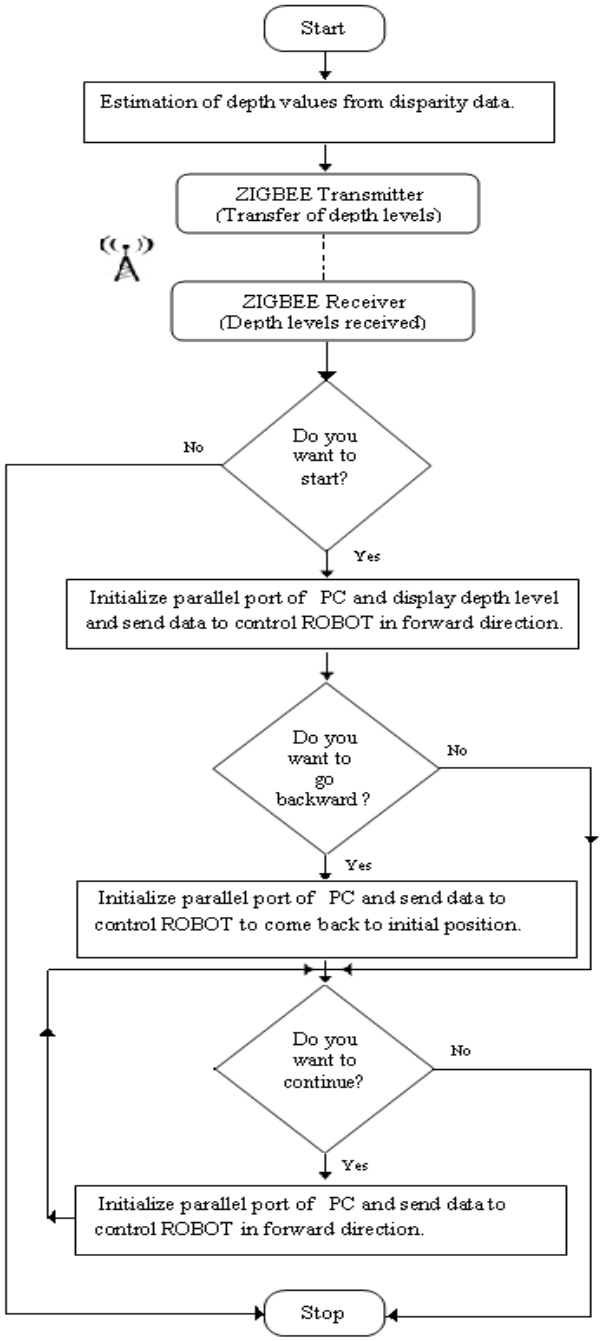


Figure 4. Flowchart for ROBOT control.

IV. RESULTS

The testing of the present setup was done on several images from Middlebury data set [4]. One of the image pair, which was transmitted using ZIGBEE and received at the receiver ZIGBEE module, is shown in Fig. 5 and Fig. 6.

.

Figure 5. Left and Right view of images transmitted.

Figure 6.Left and Right view of images received.

Figure 7. Reconstructed 3-D image.

The time taken to send a complete image was about 60 seconds; but, as there was loss of data, this image was split into eight packets. Therefore, the time taken for transmission of each packet of image data was 30.2857 seconds. Thus, the time taken to transfer the image of size 115 X 132 was 2.42 minutes. If there is loss of data, retransmission is necessary. It was observed that when there was a need for retransmission of packets, the maximum time taken to transfer a complete image was found to be 3.52 minutes. In this case, the ROBOT control mechanism is not synchronized with the type of image, but it moves forward depending upon the number of depth levels in an image. However, in the future, it may be synchronized with real-time industrial control applications. A reconstructed 3-D image at the receiver side is shown in Fig. 7.

Six different images from Middlebury data set [4] were transmitted and received at the receiver. The Peak Signal-to-Noise Ratio (PSNR) values of received images in db were plotted and are shown in Fig. 8.

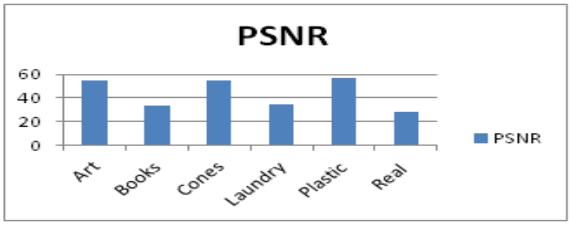


Figure 8. PSNR values obtained for images received at ZIGBEE receiver.

V. CONCLUSION AND FUTURE WORK

A 3-D image was generated at the receiver end. It was observed that there is always a compromise between PSNR and time taken to transmit the image. The time taken for transmitting an image can be reduced by implementing a mesh or star topologies using a set of ZIGBEE modules, which may give rise to loss of data. The transmission time can also be reduced by using image compression techniques at the transmitting end but this will affect the PSNR of generated stereo images, which, in turn, will affect estimation of disparity. In other words, transmission of compressed images obtained by using compression techniques lead to lossy received images. This is due to the fact that transmission of decimal point values requires more time as these data has to be converted into string format which further increases the size of data. In the future, the above algorithms can be implemented on advance microprocessors such as, ARM9 [9] which will facilitate system on chip wireless transmission modules.

REFERENCES

[1] Arti Khaparde, Apurva Naik, Manini Deshpande, Sakshi Khar, Kshitija Pandhari, and Mayura Shewale, "Performance Analysis of Stereo Matching Using Segmentation Based Disparity Map", ICDT 2013:The Eighth International Conference on Digital Telecommunications, 21-26, April 2013, Venice, Italy, pp. 38-43.

[2] Pedram Ghamisi, Micael S. Couceiro, Jón Atli Benediktsson, and Nuno M.F. Ferreira, "An Efficient Method for segmentation of Images based on Fractional Calculus and Natural Selection", Expert Systems with Applications: An International Journal, vol. 39, iss. 16, November 2012, pp. 1207-1217.

[3] http://www.vision.caltech.edu/bouguetj/calib_doc/ [retrieved: December 2nd, 2013]

[4] http://vision.middlebury.edu/stereo/data [retrieved: April 6th, 2014]

[5] Wongsavan Chantharat and Chaiyod Pirak, "Image Transmission over ZigBee Network with Transmit Diversity", 2011 International Conference on Circuits, System and Simulation IPCSIT vol. 7 (2011) © (2011)IACSIT Press, Singapore, pp. 139-143.

[6] www.zigbee.org. [retrieved:January, 2014]

[7] http://www.dashwood3d.com/blog/beginners-guide-to-shooting-stereoscopic-3d/[retrieved:January, 2014]

[8] IEEE Std 802.15.14: Wireless Medium and Physical Layer(PHY) Specification For Low-Rate Wireless Personal Area Networks (LR-WPANs),2003.

[9] www.arm.com[retrieved :January,2014]

 ISBN: 978-1-61208-364-3

Improving Digital Forensics Through Data Mining

Chrysoula Tsochataridou, Avi Arampatzis, Vasilios Katos
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi, Greece
{chrytsoc, avi, vkatos}@ee.duth.gr

Abstract—**In this paper, we reflect upon the challenges a forensic analyst faces when dealing with a complex investigation and develop an approach for handling and analyzing large amounts of data. As traditional digital forensic analysis tools fail to identify hidden relationships in complex modus operandi of perpetrators, in this paper, we employ data mining techniques in the digital forensics domain. We consider as a vehicle the Enron scandal, which is recognized to be the biggest audit failure in the U.S. corporate history. In particular, we focus on the textual analysis of the electronic messages sent by Enron employees, using clustering techniques. Our goal is to produce a methodology that could be applied by other researchers, who work on projects that involve email analysis. Preliminary findings show that it is possible to use clustering techniques in order to effectively identify malicious collaborative activities.**

Keywords-Digital Forensics; Email Analysis;Text Mining; Clustering; Weka; Simple K-means.

I. INTRODUCTION

The incorporation of computer technology in modern life has increased the productivity and the efficiency in several aspects of it. However, computer technology is not only used as a helpful tool that enhances traditional methodologies. In unethical hands, it can be used as a crime committing tool as well. Particularly, technically skilled criminals exploit its computing power and its accessibility to information, in order to perform, hide or aid unlawful or unethical activities. Nowadays, the number of information security incidents is increasing globally. Considering the fact that a big percentage of the total information produced is digital, arises the need of retrieving electronic evidence in a manner that does not affect its value and integrity [1].

Most of the collected digital evidence is often in the form of textual data, such as e-mails, chat logs, blogs, webpages and text documents. Due to the unstructured nature of such textual data, investigators, during the stage of analysis usually employ searching tools and techniques to identify and extract useful information from a text. Textual information represents one of the core data sources that may contain significant information. The amount of textual data, even on a single personal digital device, is usually very large, in the order of thousands of texts or short messages. The analyst, in this context, encounters objective difficulties in data content analysis and in finding important investigational patterns.

In this paper, we propose an approach for handling and analyzing large amounts of textual data using a tool for data analysis and predictive modeling called Weka [2], which provides a collection of machine learning algorithms that perform data mining techniques. Text mining has been proved to be able to profitably support intelligence and security activities in identifying, tracking, extracting, classifying and discovering patterns useful for building investigative scenarios [3]. The data we experiment with are the emails of the Enron corpus. The objective is to develop a method for future investigators, so that they can effectively identify and gain information from a large volume of unstructured textual data. This was accomplished first by parsing the data which were organized in folders, then by storing them into a MySQL database [4] to better manage them and finally by performing data mining techniques to the textual data in order to draw some conclusions about the content of the emails. The proposed method is especially useful in the early stage of an investigation when the researchers may have a little clue of how to begin with. As such, the main contribution of this paper is the demonstration of use of several standard data mining techniques in the domain of digital forensics as the current state of the art in digital forensics is limited in content based searching, rather than identifying contextual relations.

The remainder of this paper is organized as follows. Section II includes the related work and research that was conducted during the previous years by other scientists who dealt with email analysis. Section III introduces the reader to the Enron case. This section describes the Enron corpus of emails, the script that was developed to process the email objects as well as the MySQL database which was designed to store the data retrieved. Section IV describes the data mining techniques and tools that were used to process the textual data and enabled us to further analyze them in Section V. Finally, Section VI draws our conclusions.

II. RELATED WORK

In 2004, Bryan Klimt and Yiming Yang conducted email classification for the Enron dataset [5]. Their goal was to explore how to classify messages as organized by a human. In order to accomplish it, Support Vector Machine (SVM) [6] classifier was used after they had cleaned the data from duplicate messages. Moreover, in 2005, Jitesh Shetty and Jafar Adibi created a MySQL database for the Enron dataset and statistically analyzed it [7]. In addition to this, they

 ISBN: 978-1-61208-364-3

derived a social network from the dataset and presented a graph of it.

Concerning the text mining part, which is the extraction of knowledge from text documents, there were several tools proposed. Some of those were the Email Mining Tool (EMT) and the Malicious E-mail Tracking (MET). Those tools were developed at the Columbia University and employed data mining techniques to perform behavior analysis as well as social network analysis [8]. Furthermore, in the field of text mining, R. Al-Zaidy B. C. Fung, A. M. Youssef and F. Fortin proposed a data mining algorithm to discover and visualize criminal networks from a collection of text documents [9]. This paper also used Enron email corpus as a case study of real-life cybercrime.

III. THE ENRON CASE

The fraud investigated in this paper is the Enron scandal. The scandal was revealed in October 2001 and eventually led to the bankruptcy of the energy company Enron Corporation, based in Houston, Texas [10]. The fall of Enron has been characterized as the greatest failure in the history of American capitalism and its collapse had a major impact on financial markets, since Enron dealt with many financial institutions and organizations. The company's collapse caused investors to lose a large amount of money and employees to lose their jobs, their medical insurances as well as their retirement funds. Additionally, it caused the dissolution of Arthur Andersen LLP, which was the audit company that performed both the internal and external accounting for Enron Corporation [11]. The federal investigations lasted 5 years and revealed the complex and illegal accounting practices that were conducted and encouraged by Enron's former executives. The investigations also came up with 31 terabytes of digital data including data from 130 computers, thousands of e-mails, and more than 10 million pages of documents, culling evidence that helped deliver convictions of the company's top executives, among others [12]. The collection of those emails is used to form our forensic methodology.

A. *The Enron dataset*

A few years after the scandal, a part of the digital collection was published by William Cohen, a professor at Carnegie Mellon University. The collection originally consisted of 1,500,000 emails [13]. However, some of them were withdrawn because of the complaints that former employees made, since they believed that their personal life was violated. The Enron dataset used in our project consists of 519,000 electronic messages both personal and formal, excluding the files that were attached in those mails [14]. The emails of this collection follow the RFC 5322 format and were used as a first material in order to produce a methodology that could be applied by other researchers, who work on projects that involve email analysis [14]. To achieve this goal, a certain procedure was followed.

B. *Directory traversing and processing of email objects*

The first step was to download the previously described dataset which included electronic messages organized in 3,500 files. Each and every one of the 151 employees that participated in this collection was represented by a unique folder. The employee's folder included other subfolders such as "inbox", "sent_items", "deleted_items" etc. Finally, inside those subfolders were the electronic messages. It is obvious that various levels of folders had to be traversed in order to reach the electronic message. This became feasible via the development of a Python script. The script conducted in the first place directory traversing and data parsing (meaning syntactic analysis) soon after the electronic message was reached.

As mentioned earlier, the messages were formulated according to the RFC 5322 format. For the purpose of data

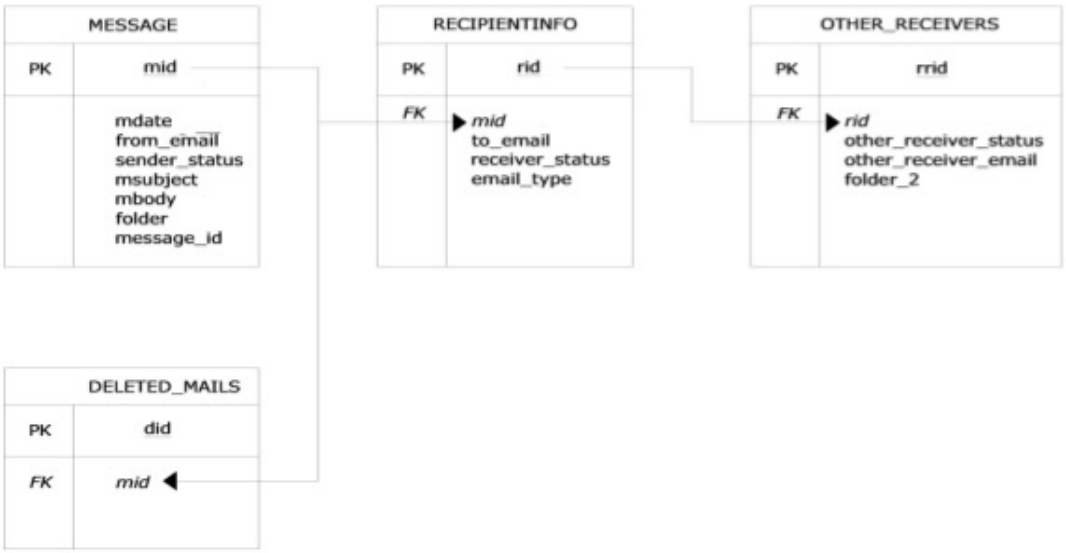


Figure 1. The Database Schema

 ISBN: 978-1-61208-364-3

parsing, the message was separated into header and body. The header of the message included information such as the date, the electronic addresses of the sender and the recipients, the subject and the id of the message. The body part represented the text of the message. After the split of the message into those two parts, the syntactic analysis of the header was performed, line by line, so that the necessary data could be extracted. The size of the data was large causing difficulties in storing and processing them using simple text files. Thus, the creation of a database that would solve those problems seemed reasonable.

C. *Description of the MySQL Database*

The data retrieved from the syntactic analysis of the message were inserted via a custom Python script into the corresponding fields of the database tables. The MySQL database named "enron" consisted of 4 tables. The table "MESSAGE" included fields with records related to the electronic message (Date, Subject, body, etc.), the table "RECIPIENTINFO" included fields with records that referred to the recipients of the message, the table "OTHER_RECEIVERS" included fields with records regarding the recipients of the BCC and CC type emails. Finally, a table named "DELETED_MAILS" was also created to store the data of the messages that employees used to delete. The database schema is shown in Fig. 1.

IV. D ATA MINING TECHNIQUES WITH WEKA

Having stored the necessary data in the MySQL database, the next step is to perform data mining techniques on them, in order to extract some useful information. For this purpose Weka is used, which is a collection of machine learning algorithms for data mining tasks [15]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization [16]. This paper focuses on textual analysis of the electronic messages using clustering techniques. Therefore, a series of filters and algorithms were applied to the subject and the body of the messages. Since Weka's software is written in Java, Weka provides access to SQL databases using Java Database Connectivity (jdbc connector) and can process the result returned by a database query. In this case, msubject and mbody were loaded from the database into Weka via the execution of a query. An example of a query executed follows below.

```
select msubject,mbody from MESSAGE
where from_email="ken.lay@enron.com";
```

Via the execution of the query above, the fields msubject and mbody of table "MESSAGE" are loaded as attributes into Weka. The messages processed and then clustered were the ones that were sent by key executives of the company. The top executives examined were Jeffrey Skilling, Kenneth Lay, Andrew Fastow, Tim Belden, Mark Koenig and Vincent Kaminski. The reason of using the "where from_email=user@enron.com " statement is to specify each time whose messages are going to be loaded.

A. *The StringToWordVector filter*

After the data is loaded, the filter StringToWordVector is applied to the attributes msubject and mbody. Those attributes contain instances which represent the subject and the body of the messages sent by the x executive. The StringtoWordVector filter pre-processes the data so that the clustering algorithm Simple K-means can later be applied. This filter converts string attributes like msubject, mbody into a set of attributes that each of them represents a single word. In other words, StringToWordVector creates a list of words which are actually attributes. These words existed inside the subject and the body of the messages. Before the application of the filter some settings have been set in the pop-up window of StringToWordVector. The settings used are the ones below:

- **TF-IDF Transform, (term frequency–inverse document frequency)**, which is a numerical statistic used as a weighting factor that reflects how important a word is to a message of the collection. The tf-idf value increases proportionally to the number of times a word appears in a message, but is offset by the frequency of the word in the corpus [17]. The weighting factor gets bigger when the word appears frequently in a small amount of messages and smaller when the word appears frequently in the majority of messages [17]. For example, the word "energy", which is going to appear many times in most of the emails- as Enron was originally an energy company- is of little importance despite its common use and is going to get a low weighting factor. Some other words instead, appear many times in a few messages. From the investigational point of view, the smaller the community that discusses a specific matter is, the more suspicious this pattern of behavior becomes. Therefore, these words get a high weighting factor because they might be indicative of malicious activities. It should be noted that TF-IDF is a standard effective transformation in Information Retrieval, however, recent empirical evidence shows that it may not be the best choice for some classification methods, e.g. SVM [18]. Nevertheless, there is no study, in our knowledge, suggesting that the same holds for clustering with K-means as employed in this work; consequently, we hold to traditional methods and use TF-IDF.
- **minTermFreq**, which defines a minimum frequency of a word occurrence. If set on 3 for example, the words that come up after the filter's application are those which appear at least 3 times in the collection.
- **Stemmer**, which removes the endings of the words. In general, stemming is the process of reducing inflected or derived words to their stem, base or root form [16].
- **Stopwords**: This parameter if set to "True", activates the default stopwordlist of Weka. A stopwordlist is a list of words such as because, is, then, often that are commonly used and do not

 ISBN: 978-1-61208-364-3

provide important information about the content of a message.

- **WordsToKeep**: This parameter predetermines the number of the words that are going to appear after we apply the StringToWordVector filter.

Table I shows the settings in the pop-up window of StringToWordVector, when the filter was applied to the string attributes msubject and mbody of Lay's sent messages. It should be noted that the settings - minTermFrequency and words_to_keep in particular - may differ between suspects, depending on the number of messages sent by each of them.

TABLE I. STRING TO WORD VECTOR SETTINGS FOR LAY'S MESSAGES

StringToWordVector settings	Lay
IDF Transform	TRUE
TF Transform	TRUE
minTermFrequency	3
words to keep	1500
stoplist Weka 3-7-1	TRUE
Stemmer	SnowballStemmer
Delimeters	0123456789!#$%^&*( )-=_+\/;"',.><][

B. Simple K-means algorithm

After the application of the filter, the string attributes msubject, mbody are converted into a list of words (dictionary), which are obviously the most frequent words that exist in the messages that sent the x executive (Kenneth Lay in the example above). The next step is to apply the Simple K-means algorithm [19] in order to perform cluster analysis on the messages. Simple K-means is very popular and one of the simplest algorithms which uses unsupervised learning to solve clustering problems. Although it has some drawbacks (strong sensitivity to outliers and noise; works best with hyper-spherical cluster shapes; number of clusters and initial seed value need to be specified beforehand; converges to local optima), other clustering algorithms with better features tend to be much more expensive [16]. Our focus in this study is to investigate and demonstrate how clustering can be useful for forensic analysis of data. The dataset we employ (519,000 electronic messages) can be considered large for many clustering algorithms, but Simple k-means is easy to be implemented efficiently and applied even to large datasets.

Cluster analysis is the task of grouping a set of messages in such way that messages from the same group (called cluster) are more similar to each other than to those in the other clusters. Simple K-means separates the emails into K groups (clusters). The K variable is an integer number and expresses the number of groups into which the messages are divided. The objective of this algorithm is the minimization of the mean squared Euclidean distance of the messages from the centers of their clusters [17]. Apart from separating the messages into K clusters, Simple K-means gives a weighting factor in the words that resulted after we applied the StringToWordVector filter. This weighting factor indicates the correlation of the word to the content of the messages that belong to a cluster. Moreover, the weighting factor depends on the word's frequency in the messages in general. The weighting factor is different for the same word from cluster to cluster. This is because one word can be more representative of the content of the messages in cluster0 than in cluster1. The biggest the factor is, the more representative of the cluster the word is. This function of Simple K-means helps in better understanding the content of clusters of messages. There is also a pop-up window for Simple K-means with a series of settings. The parameters which have been set before the application of the algorithm are:

- **numClusters**, which determines the number of the clusters into which the messages will be separated [17].
- **maxIterations**, which expresses the maximum number of iterations and predetermines the implementation time of the algorithm Simple K-means [17].
- **Seed**: With this setting multiple iterations of the algorithm with different random initial centers can be done [17].

The first 3 rows of Table II show the values that were given to the settings of Simple K-means pop-up window for the cluster analysis of Lay's sent messages, before the algorithm was executed. The other rows indicate Simple K-means behavior after the execution of the algorithm to the data. Similar to StringToWordVector, the settings number_of_clusters and seed may differ between the suspects examined. The clusters for each suspect in the performed experiments were not more than six. This choice was guided by the practical demand of obtaining a limited number of informative groups.

TABLE II. SIMPLE K-MEANS SETTINGS FOR THE CLUSTER ANALYSIS OF LAY'S SENT MESSAGES

SimpleK-means settings	Lay
number of clusters	5
max iterations	30
Seed	8
iterations needed	2
sum of squared errors	1815.91
Attributes	598

The final values given in the parameters above were chosen based on the fact that their combination minimized the sum of squared errors. The distribution of messages into clusters is being described in the next section.

V. RESULT ANALYSIS

In this section, the results that came up after the application of the Simple K-means algorithm will be analyzed. As it was previously stated, the data processed were the subject (msubject) and the body (mbody) of the

 ISBN: 978-1-61208-364-3

messages that were sent by Enron's key "players". The cluster analysis which was conducted resulted in the configuration of five clusters of messages. Table III shows the distribution of Lay's sent messages per cluster.

TABLE III. D ISTRIBUTION OF L AY ' S SENT MESSAGES INTO 5 CLUSTERS

Clustered Instances		
Cluster 0	7566	94%
Cluster 1	255	3%
Cluster 2	16	0%
Cluster 3	70	1%
Cluster 4	127	2%

It is being obvious, that the majority of messages (7566 messages) that Lay sent belongs to cluster0. The second cluster (cluster1) includes 255 messages, the third (cluster2) 16 messages, the fourth (cluster3) 70 messages and finally the fifth one (cluster4) includes 127 messages. With the help of MS Excel, Weka's results are processed. Specifically, every time an executive's messages are being clustered, a table is produced.

Each column of this table represents a cluster of messages and includes the most important words of the cluster. The words for each cluster are being written in a descending order, according to the weighting factor that Simple K-means gave in each word for every cluster. The bigger the factor is, the most important the word for the cluster becomes. Consequently, the most representative words of a cluster's content possess the first positions of each column.

TABLE IV. THE FIRST MOST FREQUENT AND IMPORTANT WORDS IN K ENNETH LAY'S CLUSTERS OF MESSAGES

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Analyst	calendars	AFL	Tuesday	Electric
Billy	Noon	AFLCIO	Open	Embedded
ChairManagement	St	CounselAFLCIO	Hold	Energy
ChairmanSubject	Joannie	DC	October	Gas
CommitteeAssociate	Williamson	DSilvers	Forward	HNG
DepartmentHuman	Note	Damon	Meeting	Natural
LayDepartment	questions	Dsilvers	Monday	Pipeline
Leading	Call	General	Directors	Transco
LeadsProgram	Place	November	Managing	allies
Lemmons	Executive	Press	Bringing	arrival
Office	announced	Release	Quarterly	arrived
Program	Basis	ReleaseKen	Earlier	chairman
ProgramDate	quarter	Silvers	Executive	challenges
ProgramJohn	Directors	SilversAssociate	Announced	changing
RepsProgram	Managing	Street	Basis	commitment
Resources	bringing	aflcio	Quarter	dedication
Sherriff	quarterly	apologies	PMTo	developing
SupervisorsEmbedded	Monday	inadvertantly	Lay	directors
Worlds	meeting	omitted	MessageFrom	endeavors
WorldwideFrom	October	release	Work	energy
Asset	forward	doc	Kenneth	environment
Broadest	Earlier	press	Office	Era
businesses	purpose	Washington	KennethSubject	facing
Campus	Committee	attached	Rosalee	Felt
Cc		fax	Dont	gratitude

 ISBN: 978-1-61208-364-3

Table IV shows the most frequent words that exist in every cluster of Lay's messages. Our aim is to understand the content of the messages that belong to each cluster and then make some assumptions about the concerns of the executives (in this case Lay's) through time. It is worth mentioning that the table below is just a sample. The original table was very large to fit in this paper. For this reason, we chose to show the table with the first 30 most important words for each cluster of messages.

The same process was also applied in the cases of the other executives. The collected emails were processed and clustered separately, thus obtaining four different scenarios, one for each employee. The underlying hypothesis was that email contents can also be characterized by the role the mailbox owner played within the company. Tables V to VIII that follow, report on the results obtained by these experiments. Each table shows the terms that characterize each and every one of the clusters. For each cluster, the most descriptive words between the 30 most frequent words of the cluster are listed.

TABLE V. LAY'S RESULTS

cluster	Most frequent and important words
0	LEADS Program, Program, graduates, worldwide, guide, supervisors, direction, importance, members, philosophy
1	Joannie Williamson, executive, director, meeting, earlier, October, purpose, Committee
2	AFL, AFLCIO, counsel, release, November, Press, Ken, Silvers, apologies
3	Open, meeting, directors, office, Kenneth, Rosalee
4	HNG, Natural, Gas, challenges, changing, commitment, dedication, era, history, ideas, gratitude

TABLE VI. SKILLING'S RESULTS

cluster	Most frequent and important words
0	Axis, analysts, create, Department, find manager, dataProgram, search, Toolbox, profile, qualifications, applications, website, career
1	slipped, busy, running, meeting, Baxter, Whalley
2	Stanford, Survey, favorite, firm, feedback, employees, McKinsey, Globe, marketplace
3	Harvard, Presented, GarvinProfessor, ProfessorDavid, Plan, seminar, drinks, partnership, HBS
4	Amanda, Andrew, attend, plz, Colwell, Donahue, Fastow, Scott, Shapiro, Kaminski, Whalley, Wednesday
5	Skilling, Joannie, dinner, discuss, meeting, questions, company

The analysis of the tables was the part of the process that involved a significant amount of uncertainty. Based on the words of each cluster of messages and on previous research that was conducted on the Enron case, an attempt was made on drawing some conclusions regarding the content of the

TABLE VII. FASTOW'S RESULTS

cluster	Most frequent and important words
0	Brasoil, Brazil, Rio, transaction, accountants, agreement
1	AIG, Highstar, coming, fund, discuss, Louise, date
2	Lay, Cliff, Andy, Jeff, Louise, critical, LJM, didn't, interest, job, loss, lost, market, members

TABLE VIII. BELDEN'S RESULTS

cluster	Most frequent and important words
0	Article, Californiacentric, Pools, PowerExchange, Professor, Wilson, economist, versuscentralized, powerpools, pros, cons, fit, advised, compares, people, market, California
1	Communicating, DianaScholtes, risk, cashflows, headaches, problem, California, term
2	FRR, frequency, pool, construction, hydrological,machines, spin, Ingersoll
3	Communication, Excellence,GlobalMarkets, Improve, Introduce, Influence, Promote, Simplify, application, opportunity, diagnosis
4	information, prices, market, electricity, observations, analysis, demand, power, purchases, FERC, year, month, Ray Alvarez

messages that were sent by some of the top executives of the organization. Even by a superficial inspection of the words in the first column of Table IV, represented as cluster0, we assume that Lay's messages had to do with a program called "LEADS". There is a great chance that Enron had employed graduates from that program. Lay's messages tend to inspire the superior executives so that they help those new employees to adapt easily in Enron's environment. Those assumptions are based on the presence of the words LEADS Program, Program, graduates, worldwide, guide, supervisors, direction, importance, members, philosophy (some words cannot be seen in Table IV, but exist in the original full table and are worth mentioning). After some research from external public sources, it was discovered that the UC LEADS Program is one of the most prestigious fellowships awarded by the University of California system [20]. This program supports up to nine UCLA upper-division undergraduate students in the fields of science, technology, engineering, and mathematics with educational experiences that prepare them to claim positions of leadership in academia, industry, government, and public services following the completion of a doctoral degree. This confirms our initial hypothesis for the content of Lay's message, since we know the target group of employees that Enron use to have fulfills the above criteria. From the investigational point of view, some interesting aspects emerges in Lay's results as there is an interesting cluster (cluster 1) in which the context is referring to a meeting that was scheduled to take place earlier than it was originally planned. The words "purpose" and "Committee" might be an indication of why the meeting had to be rescheduled, something that causes suspicion as we

 ISBN: 978-1-61208-364-3

know that there was a committee that was monitoring and investigating Enron's financial statements and balance sheet.

Skilling's emails do not underline any particular trend. It could be interesting to analyze more in depth cluster 4 where several names of top executives are being noticed. However, Fastow's results seem extremely interesting. Specifically, cluster 0 includes words that express business activities and transactions. After some electronic research, it was found that in September 1999, Enron sold LJM I a 13% stake in a company that was building a power plant in Cuiaba, Brazil. This sale, for approximately $11.3 million, altered Enron's accounting treatment of a related gas supply contract and enabled Enron to realize $34 million of mark-to-market income in the third quarter of 1999, and another $31 million of mark-to-market income in the fourth quarter of 1999 [21]. The offshore company LJM was Fastow's creation and this transaction seemed to be very critical. Cluster 2 also underlines a compact group of emails in which important financial aspects are discussed. It's worth noticing that the name "Louise" exists both in clusters 1 and 2, which indicates that Louise and Fastow had a close professional relationship.

Belden's emails have no particular terms. The only aspect that can be clearly understood is that his position is tightly linked to California business and electricity market. This definitely addresses the need to examine more in depth those emails, as they may lead to the acquisition of valuable information regarding the California crisis and Enron's role in it [22]-[23].

Summarizing, we made an effort to approach the content of the messages of the other clusters in a similar manner. Initially, we examined the words of each cluster of messages in order to formulate an assumption related to the content of the emails. Then, we cross-validated our hypothesis via further literature research confirming some keywords to events relating to suspicious activities as shown above.

VI. CONCLUSIONS

The textual analysis of each cluster of messages was a process that provided useful information to the researcher. However, there are risks in the analysis of the results since this process is cumbersome and difficult because of the unstructured nature of the text documents. It is important that every time cross-validation of the digital evidence with other resources is conducted to reveal the truth of a fact. Concerning the clustering techniques used, the Simple K-means algorithm was efficient and performed cluster analysis in a satisfying level without significant differences in the resulting clusters when the parameters in the algorithm's pop-up window were changed for the same person. The content of the messages that belonged to each cluster was obvious enough to make our assumptions and gain a deeper idea of each executive's concerns. Moreover, neither null clusters nor large squared errors were noticed after the execution of the algorithm. Finally, with respect to the matter of large datasets, it was noticed that the process was time-consuming and inaccurate.

REFERENCES

[1] U.S. Department of Justice, "Electronic Crime Scene Investigation: A guide for First Responders", [Online]. Available: http://www.ncjrs.gov/pdffiles1/nij/219941.pdf. [Accessed 26 05 2014]

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, vol. 11, Issue 1, 2009.

[3] W. Fan, L. Wallace, S. Rich and Z. Zhang, "Tapping the power of text mining". Comm. of the ACM. 49, pp. 76–82 , 2006.

[4] MySQL Database [Online] Available: http://www.mysql.com/

[5] B. Klimt and Y. Yang, "The Enron Corpus: A new dataset for email classification research", [Online]. Available: http://www.bklimt.com/papers/2004_klimt_ecml.pdf. [Accessed 26 05 2014].

[6] C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp.273-297, 1995.

[7] J. Shetty and J. Adibi, "The enron email dataset, database schema and brief statistical report", [Online]. Available: http://foreverdata.com/1009/Enron_Dataset_Report.pdf. [Accessed 26 05 2014].

[8] S. J. Stolfo, S. Hershkop, K. Wang, and O. Nimeskern, "EMT/MET: systems for modeling and detecting errant e-mails", Proceedings of DARPA Information Survivability Conference and Exposition, 2003.

[9] R. Al-Zaidy, B. C. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents", [Online]. Available: http://dmas.lab.mcgill.ca/fung/pub/AFYF12diin.pdf. [Accessed 26 05 2014].

[10] Wikipedia, the free encyclopedia, "Enron scandal", [Online]. Available: http://en.wikipedia.org/wiki/Enronscandal. [Accessed 26 05 2014].

[11] W. W. Bratton, "Enron and the dark side of shareholder value", [Online].Available: http://papers.ssrn.com/sol3/papers.cfm? abstract_id=301475. [Accessed 26 05 2014].

[12] The FBI, federal bureau of investigation, "Digital Forensics: It's a bull market", [Online]. Available: http://www.fbi.gov/news/stories/2007/may/ rcfl050707. [Accessed 26 05 2014].

[13] W. Cohen, MLD, CMU, "Enron Email Dataset",[Online]. Available: https://www.cs.cmu.edu/~enron/. [Accessed 26 05 2014].

[14] E. P. Resnick, "Internet Message Format", [Online]. Available: http://tools.ietf.org/html/rfc5322. [Accessed 26 05 2014].

[15] Machine Learning Group at the University of Waikato, "Data Mining:Practical Machine Learning Tools and Techniques", [Online]. Available:http://www.cs.waikato.ac.nz/~ml/weka/book.html. [Accessed 26 05 2014].

[16] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementation,", Department of Computer Science, University of Waikato, New Zealand, [Online]. Available: http://www.cs.waikato.ac.nz/~ml/publications/1999/99IHW-EF-LT-MH-GH-SJC-Tools-Java.pdf. [Accessed 26 05 2014].

[17] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[18] G. Forman, "BNS feature scaling: an improved representation over tf-idf for svm text classification". In proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08), ACM, New York, NY, USA, pp. 263-270, 2008.

 ISBN: 978-1-61208-364-3

[19] J. A. Hartigan, M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm", Journal of the Royal Statistical Society, Series, vol. 28, no. 1, pp. 100–108, 1979.

[20] Univeristy of California, "UC LEADS", [Online]. Available: http://www.ugresearchsci.ucla.edu/ucleads.htm. [Accessed 26 05 2014].

[21] FindLaw, "Brazil is rising", [Online]. Available: http://news.findlaw.com/wsj/docs/enron/sicreport/chapter6.html.[Accessed 26 05 2014].

[22] Wikipedia, the free encyclopedia, "California electricity crisis,[Online]. Available: http://en.wikipedia.org/wiki /California_ electricity_crisis. [Accessed 26 05 2014].

[23] J. Leopold, "Enron linked to California blackouts", Available:http://www.marketwatch.com/story/enron-caused-california-blackouts-traders-say. [Accessed 26 05 2014]

 ISBN: 978-1-61208-364-3

Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining

Process Mining in the Education Domain

Awatef Hicheur Cairns[1], Billel Gueni[1], Mehdi Fhima[1], Andrew Cairns[1] and Stéphane David[1]
Nasser Khelifa[2]

[1] ALTRAN Research, [2] ALTRAN Institute
Vélizy-Villacoublay, France
e-mails: {awatef.hicheurcairns, billel.gueni, mehdi.fhima, andrew.cairns, stephan.david, nasser.khelifa}@altran.com

***Abstract*—Educational Process Mining constitutes a new opportunity to better understand students' learning habits and finely analyze the complete set of educational processes. In this paper, we investigate further the potential and challenges of Process Mining in the field of professional training. Firstly, we focus on the mining and the analysis of social networks between course units or training providers. Secondly, we propose a two-step clustering approach for partitioning educational processes following key performance indicators.**

Keywords- Process Mining; Educational Data Mining; Curriculum Mining; Key Performance Indicators; ProM.

I. INTRODUCTION

Recently, education and training centers have started introducing more agility into their teaching curriculum in order to meet the fast-changing needs of the job market and meet the time-to-skill requirements. In fact, modern curriculums are no longer monolithic processes. Students can pick the courses from different specialties, may choose the order, the skills they want to develop, the level (from beginner to specialist), the way they want to learn (theoretical or practical aspects) and the time they want to spend. This need for personalized curriculum has increased with the emergence of collaborative tools and on-line training which often supplement and sometimes replace traditional face-to-face courses. The broad number of courses available and the flexibility allowed in curriculum paths could create, as a side effect, confusion and misguidance. Students may be overwhelmed by the offer and blurred on the time required to enter and remain in the job market. Moreover, teachers and educators may lose control of the education process, its end-results and feed-back. The use of information and communication technologies in the educational domain generates large amount of data, which may contain insightful information about students' profiles, the processes they went through and their examination grades. The deriving data can be explored and exploited by the stakeholders (teachers, instructors, etc.) to understand students' learning habits, the factors influencing their performance and the skills they acquired [6] [15]. Rather than relying on periodic performance tests and satisfaction surveys, exploiting historical educational data with appropriate mining techniques enables in-depth analysis of students' behaviors and motivations [6] [8]. *Educational Data Mining* (EDM) is a discipline aimed at developing specific methods to explore educational. EDM methods can be classified into two categories – (1) Statistics and visualization (e.g., Distillation of data for human judgment), and (2) Web mining (e.g., Clustering, Classification, Outliers detection, Association rule mining, Sequential pattern mining and Text mining) [15]. However, most of the traditional data mining techniques focus on data or simple sequential structures rather than on full-fledged process models with concurrency patterns [20] [21]. Precisely, the basic idea of *process mining* [1] is to discover, monitor and improve real processes (i.e., not assumed nor truncated processes) by extracting knowledge from event logs recorded automatically by Information Systems. Our research aims to develop generic methods which could be applied to general education issues and more specific ones concerning professional training or e-learning fields for:

- The extraction of process-related knowledge from large education event logs, such as: process models and social networks following key performance indicators or a set of curriculum pattern templates.
- The analysis of educational processes and their conformance with established curriculum constraints, educators' hypothesis and prerequisites.
- The enhancement of educational process models with performance indicators: execution time, bottlenecks, decision point, etc.
- The personalization of educational processes via the recommendation of the best course units or learning paths to students (depending on their profiles, their preferences or their target skills) and the on-line detection of prerequisites' violations.

In this paper, we focus mainly on (1) process model discovery, deriving from Key Performance Indicators; (2) social network discovery between training courses and training providers. We used the ProM framework (i.e., a "pluggable" environment for process mining) [7] for process discovery and analysis from event logs. For the first time, to our knowledge, the present approach handles a professional training dataset of a consulting company involved in the training of professionals. The rest of this paper is organized

 ISBN: 978-1-61208-364-3

as follows. Section II introduces process mining techniques and their application in the educational domain. Section III presents our approach for social networks mining and process models discovery. Section IV describes briefly the PHIDIAS platform. Finally, section V concludes the paper.

II. EDUCATIONAL PROCESS MINING

The purpose of Process Mining is to develop automated techniques to extract process-related knowledge from *event logs* [1]. An event log corresponds to a set of process instances following a business process. Each recorded event refers to an *activity* and is related to a particular process instance. An event can have a *timestamp* and a *performer* (i.e. a person or a device executing or initiating the activity). Moreover, in such logs, events are assumed to be totally ordered. The scope of Process Mining includes process discovery (i.e., extracting process models from an event log), conformance checking (i.e., monitoring deviations by comparing model and log), social network/organizational mining, automated construction of simulation models, model extension, model repair, case prediction, and history-based recommendations. Educational Process Mining (EPM) or Curriculum Mining refers to the application of process mining techniques in the education domain [20] [21]. Beyond limitations of EDM, EPM enables greater insights into underlying educational processes. To illustrate, process mining techniques were used by Pechenizkiy et al. [13] to investigate the students' behavior during online multiple choice examinations. Southavilay et al. [18] used process model discovery techniques to mine and analyze a collaborative writing process. Analysis techniques were also applied to check the conformance of a set of predefined constraints (e.g., prerequisites) with event logs [21]. However, the application of Process Mining techniques in the education domain faces numerous challenges related to event logs' specificities:

Voluminous Data: event logs in the education domain, particularly those coming from e-learning environments, contain massive amounts of fine granular events and process-related data. Real-life testing showed that most of the current process mining techniques/tools are unable to handle massive event logs [12] [14] [17].

Heterogeneity and complexity: educational processes are complex and flexible by nature reflecting the high diversity of behaviors in students' learning paths. Consequently, traditional process discovery techniques generate intricate models (spaghetti) which are often very confusing and difficult to analyze [22].

Concept drift: in the education domain, subjacent curriculum and trainings may evolve over time and occasionally undergo major changes. Concept drift refers to a situation where the process will change while being analyzed [5].

Interpretation of results by the end users: visualization techniques and notation simplification is a major stake to facilitate interpretation by the end users, using suitable academic notation or lists of recommendations [14].

III. CASE STUDY: ANALIZING TRAINING PROCESSES USING EPM TECHNIQUES

A. Data description and preprocessing phase

In our case study, the dataset encompasses the *employees' profiles*, their *careers* (i.e., their jobs/assignments history) and their *training paths* (performance and satisfaction surveys throughout the different training phases). The data being scattered in several databases, we had first to rebuild a consolidated event log (using an ETL -*Extract, Transform and Load*) containing the following information: *Employee Id, Training Id, Timestamp, Training provider Id, Training Cost* and the *List* of all the *employees' missions* over a three years period. Secondly, we transformed this event log into MXML (Mining eXtensible Markup Language) format by using the *ProM Import* plug-in [7], with the condition that (1) an employee identifier corresponds to a process instance identifier (i.e., an employee training path corresponds to a process instance), (2) a training identifier corresponds to a task identifier tagging 'start' and 'end' events, with various attributes (grades, satisfaction, employee profile, etc.).

B. Social network mining

Social Network Analysis (SNA) refers to the collection of methods, techniques and tools in sociometry aiming at the analysis of the structure and composition of ties in social networks [4]. In our case, we conducted mining and analysis of the key interaction patterns between training providers and training courses, using social mining techniques deriving from the process mining field. These techniques aim to extract social networks from event logs based on the observed interactions between performers and depending on how process instances are routed between these performers [1] [3]. These interactions can be generated following one of these five metrics: (1) *handover of work*, (2) delegation or *subcontracting* of tasks, (3) frequent collaboration (*working together*) (4) similarity in executed tasks and (5) *reassignment* of tasks. In order to generate social networks between training courses, we replaced originator IDs by training IDs of the same events (i.e., trainings) during the event log conversion step in *ProM import*. Social mining plug-ins generate graphs where each node represents a training provider (resp. a training course) where the names have been anonymized for privacy reasons. The oval shape of the nodes in a graph (see Figures 1 and 3) visually expresses the intensity of in and out connections (arrows) between the nodes: a higher proportion of ingoing (outgoing) arcs lead to greater vertical (horizontal) distortion of the oval shape (see Figures 2 and 4). We use different views (a ranking view, a stretch by degree ratio,

 ISBN: 978-1-61208-364-3

etc.) and two key SNA indicators when generating these graphs, depending on the patterns we want to extract. The key SNA indicators [4] we used are: (1) *Degree Centrality* of a node (i.e. the number of nodes that are connected to it): it represents the popularity of a node (e.g., training courses or training providers) in a community (e.g., training paths or curriculums). (2) *Betweenness Centrality* of a node: representative of the enabling power of a node to connect two different groups (i.e., two different training paths or curriculums). A node (i.e., training provider or training course) with high betweenness centrality value means that it performs a crucial role in the network. We apply these five metrics to mine social networks between training providers and training courses, giving the following outcome:

1) Handover of work: within a process instance, there is a handover of work from individual *i* to individual *j* if there are two subsequent activities where the first is completed by *i* and the second by *j*. In our case, this metric allowed us to discover the flow of trainees (specified by the direction of the arrows) between training providers and courses. For instance, in Figure 1, two providers are connected if one performs a training causally followed by a training performed by the other provider. We distinguished two groups of providers strongly related to each other (clustered in cliques) following their causal involvement in training paths. Training providers without arc are those which offer very stand alone trainings without causal dependency with others. In Figure 2, the most important training courses (trainings with ID 4 and ID 1) appeared to play a central roles in training paths. In Figure 3, the size of training courses (with high betweenness) indicates their crucial role as a bridge (i.e., intermediate trainings) between different types of trainings. We can deduce that:

- Training providers or courses with *high degree* are the most popular ones. Training providers or courses with no connection with others represent outliers, providing very specific skills, not involved in training paths.

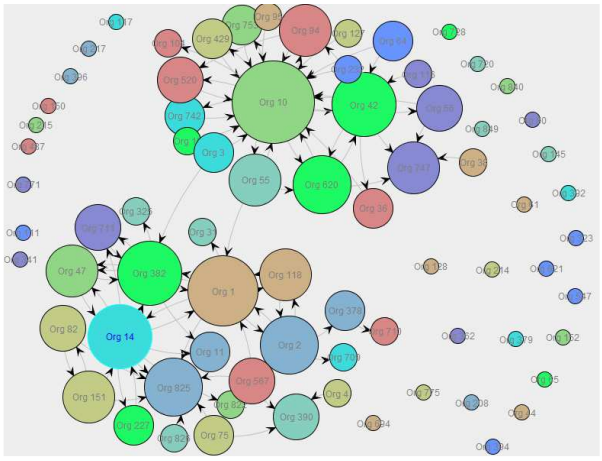


Figure 1. Social network showing handovers between providers of the top 80% of followed training courses using a *size by ranking* view.

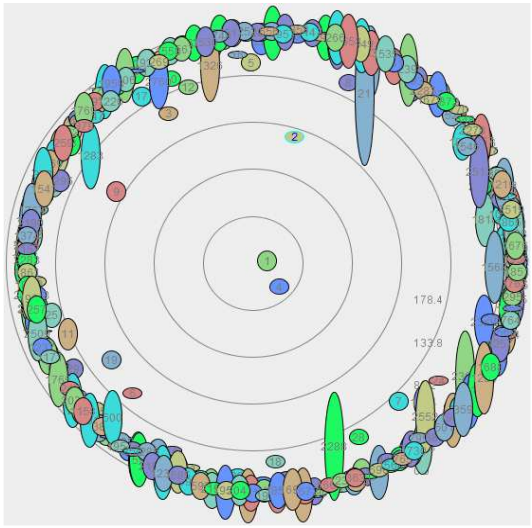


Figure 2. Social network showing *handovers* between training courses using (1) *a rancking view on degree* and (2) a *stretch by degree ratio*.

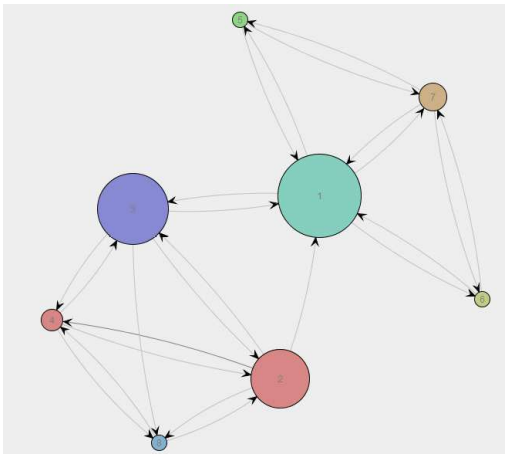


Figure 3. Social network showing *handovers* between the top 60% of followed training courses, using a *ranking on betweenness centrality* and a *size by ranking* view.

- Nodes with no incoming arcs are training providers (or training courses) who only initiate learning processes (i.e., give the basics), while nodes with no outgoing arcs are training providers (or courses) who perform only final trainings (i.e., complete training paths with the most required skills).
- Training courses strongly connected to each other hint popular or typical curriculums (or learning paths). The direction of the edges gives the order of training courses followed by students in such curriculums.
- Training courses or providers with *high betweenness centrality* represent the ones playing a crucial role as a bridge (i.e., offering intermediate trainings) connecting different types of learning paths.

 ISBN: 978-1-61208-364-3

2) Subcontracting metric: A resource *i* subcontracts a resource *j*, when in-between two activities executed by *i* there is an activity executed by *j*. In this case, the start node of an arc represents a contractor and the end node means a subcontractor (see Figures 4 and 5). In this case study, this metric allow us to extract complementary patterns between training courses and providers. Using SNA measures, we deduce that:

- Nodes (i.e., training providers or courses) with a high out-degree of centrality (indicated by a horizontal oval shapes) usually play the role of contractors (the main providers or trainings which give basic skills in these training paths).
- Nodes (i.e., training providers or courses) with a high in-degree of centrality (indicated by a vertical oval shapes) usually act as subcontractors (providers or trainings which give complementary notions or skills allowing to enhance the notions given by contractors in these training paths).

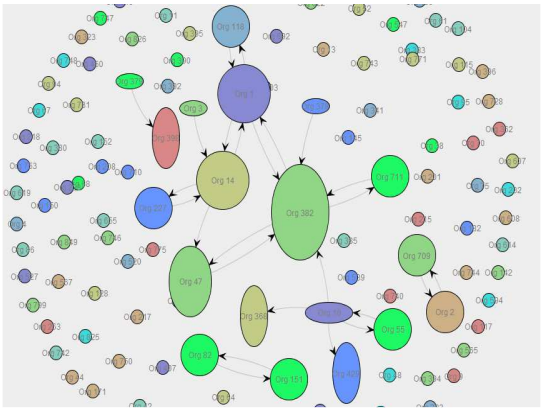


Figure 4. Social network showing *subcontracting* between training providers of the top 90% of followed training courses.

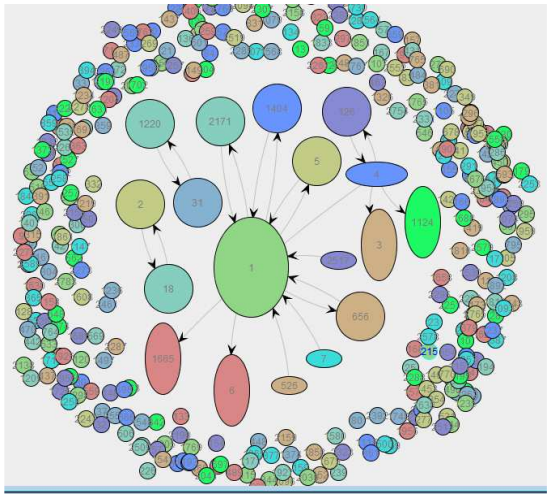


Figure 5. Social network showing *subcontractin*g between the top 80% of followed training courses.

3) Working together metric: This metric ignores causal dependencies but simply counts how frequently two recourses are performing activities for the same case (see Figure 6). We can deduce from this social network the most popular curriculums (training providers or courses that work together i.e., are involved together in training paths). The difference with the handover metric is that the latter gives us the order followed by students in such curriculums.

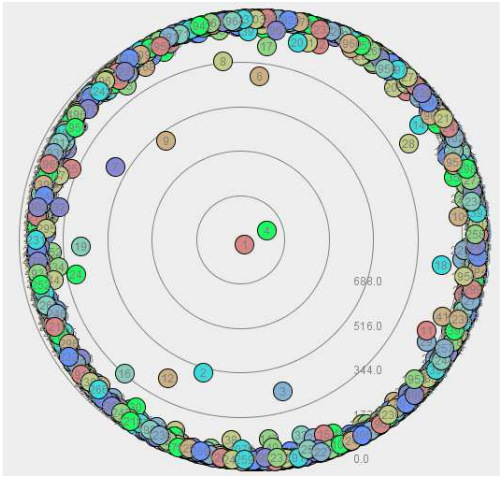


Figure 6. Social network based on *working together* between training courses using *a rancking view on degree*.

4) Similar task metric: This metric determines who performs the same type of activities in different cases. In our case study, this metric makes sense only to generate relationship between training providers. Therefore, it allows us to detect training providers who perform the same kind of trainings in curriculums.

This experience shows that social network analysis based on event logs is a powerful tool for analyzing coordination patterns between training courses and training providers. Such an approach can also be used to mine interesting patterns about students' behaviors in on-line environments based on resources' usage logs and various interaction logs (e.g., in the case of an intelligent tutoring system).

C. Process discovery using Clustering Techniques

Clustering techniques can be used as a preprocessing step to handle large and heterogeneous event logs by dividing an event log into homogenous subsets of cases following their similarity. One can then discover simpler process models for each cluster. For this purpose, several clustering techniques have been developed and implemented in ProM [11], such as the Trace Clustering plug-in [9] [17], the Sequence Clustering plug-in [22] and other clustering approaches based on time [11]. Clustering of event logs still remains a subjective technique. A desired goal would be to introduce some objectivity in partitioning the log into homogenous cases. In this paper, in order to identify the best training paths, we propose a two-step clustering approach where

 ISBN: 978-1-61208-364-3

training paths are firstly partitioned following a performance indicator then they are partitioned following their structural similarity. *The first step* consists of creating clusters of similar trainees' profiles based on a training path performance indicator expressed via two criteria: (1) *employability* (matching degree between skills required by a mission and skills obtained via training) and (2) *duration* between the training end and new job start. Based on trainees' profiles, three clusters are created using the k-means technique. The optimal number of these clusters (three) is determined using a method based on the average silhouette of many clustering where the number of the clusters is varied [16] [19]. For more details on this method we refer the reader to [10]. Figure 7(a)-(b) presents, respectively, the clusters we obtained and the silhouette used to determine the optimal number of clusters.

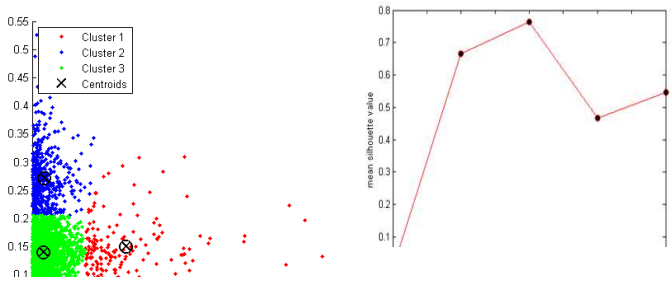


Figure 7. (a) Three clusters obtained (b) silhouette used to determines the optimal number of clusters

Let us note that the first cluster groups trainees with the best employability factor and the shortest duration between trainings' end and new missions. Cluster 2 and Cluster 3 group less optimal training paths regarding employability factor and period of unemployment. We use the fuzzy miner plug-in of ProM (given its robustness to noises) to discover the process model from the training traces of the trainees grouped in the first cluster. We obtain clearly identifiable training paths, as illustrated in Figure 8. Let us note that these training paths correspond to the highest performing ones regarding employability factor and period of unemployment.

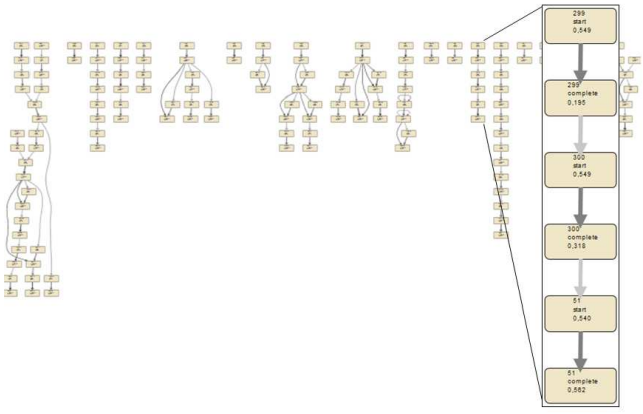


Figure 8. A fragment of the process model showing all the training patterns of cluster 1

In the second step, for further simplification, we group training paths (traces in log events) from each of the clusters discovered in the first step, following their structural similarity using the Sequence clustering technique proposed by Veiga and Ferreira [22]. Each cluster is based on a probabilistic model, namely a first-order Markov chain. The sequence clustering technique is known to generate simpler models than trace clustering techniques developed in [22]. In our example, when we apply the sequence clustering technique on the second group of trainees with an average employability (i.e., the second cluster of the first step), we obtain three more clusters (cluster 2.1, cluster 2.2 and cluster 2.3). Figure 9 shows the training model obtained from the cluster 2.1 obtained above, where only transitions occurring above the threshold of 0.05 are represented.

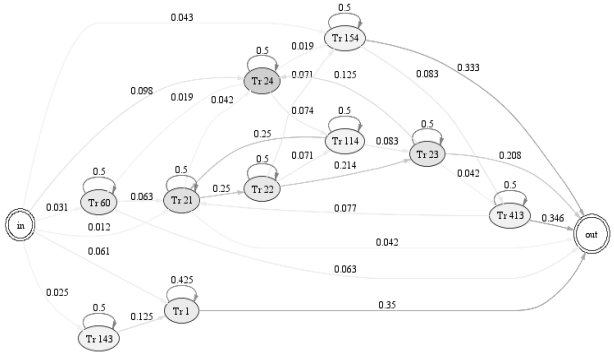


Figure 9. The process model describing the training paths of the first cluster of the second group of trainees (Cluster 2.1), with an edge threshold of 0.05

IV. PHIDIAS: A PLATFORM FOR DISTRIBUTED EDUCATIONAL PROCESS MINING

To implement our approach, we aim to develop an interactive platform tailored for educational process discovery and analysis. This platform will allow different education centers and institutions to load their data and access advanced data mining and process mining services. Such a platform has to address several issues related to: (1) the heterogeneity of the applications and the data sources; (2) the connection to some web portals and desktop applications to allow users dealing with the data and exploiting analysis results; (3) the ability to add new data sources and analysis services; (4) the possibility to distribute heavy analysis computations on many processing nodes in order to optimize and enhance platform response time. To reach these targets we adopt a Service Oriented Architecture using an Enterprise Services Bus (ESB) depicted in Figure 10. This architecture is composed of the following elements: data sources, Enterprise Service Bus, business applications and tools, web services, web portals and connectors. The core of this architecture is the application bus which guarantees the interoperability and integration of the data sources and applications. We have chosen to use ESB architecture in order to have a flexible architecture allowing easily plugging of new applications, data sources and web portals.

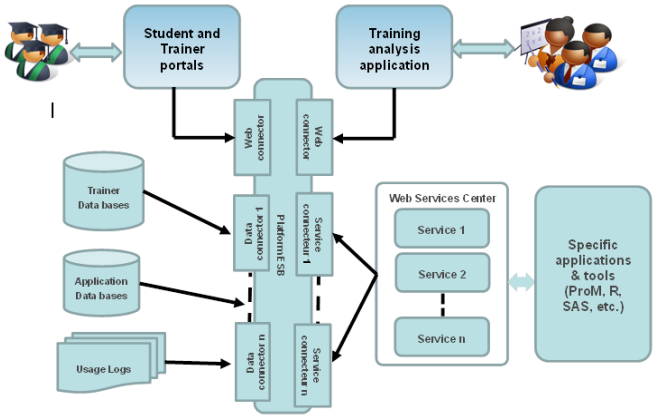


Figure 10. PHIDIAS Architecture

V. CONCLUSION AND FUTUR WORK

In this paper, we showed how social mining techniques can be used to examine interactions between training providers and training courses, involved in students' training paths. We also proposed a two-step clustering approach to extract the best training paths depending on an employability indicator. Our future work will continue in several directions. Firstly, we intend to combine the approach proposed to mine interaction patterns with other mining techniques which allow to discover interaction patterns between students in their collaborative learning tasks, communication actions and online discussions [2]. Secondly, we intend to apply the conformance checking techniques to check if prerequisites, other kinds of constraints and training path templates were indeed always respected. Thirdly, we plan to investigate further clustering techniques in event logs partitioning to extract typical or atypical training paths depending on domain specific performance indicators and/or on a set of predefined training path templates. Finally, the proposed architecture will be implemented and deployed and tested on a *distributed environment* connected to several data sources and applications. We plan also to conduct a case study that would illustrate the feasibility of process mining approaches in an on-line education setting.

REFERENCES

[1] W. M. P. van der Aalst et al. "Process mining manifesto," In BPM 2011 Workshops Proceedings (BPM 2011), Aug. 2011, pp. 169–194, doi:10.1007/978-3-642-28108-2_19.

[2] W. M. P. van der Aalst and Adriy Nikolov, "EMailAnalyzer: An E-Mail Mining Plug-in for the ProM Framework," BPM Center Report BPM-07-16, BPMCenter.org, 2007.

[3] W. M. P. van der Aalst and M. Song, "Mining social networks: Uncovering interaction patterns in business processes," The second International Conference on Business Process Management (BPM 2004), LNCS, vol. 3080, June 2004, pp. 244–260, doi. 10.1007/978-3-540-25970-1_16.

[4] C. Aggarwal, "An Introduction to Social Network Data Analytics," Social Network Data Analytics, Springer, 2011, pp. 1-15.

[5] R. Bose, W. M. P. van der Aalst, I. Zliobaite, and M. Pechenizkiy, "Handling Concept Drift in Process Mining," The 23rd International Conference (CAiSE 2011) LNCS 6741, Springer, June 2011, pp. 391–405, doi:10.1007/978-3-642-21640-4_30.

[6] T.Calders and M. Pechenizkiy "Introduction to The Special Section on Educational Data Mining," SIGKDD Explorations Newsletter, ACM, may 2012, pp. 3-6, doi:10.1145/2207243.2207245.

[7] B. van Dongen, H. Verbeek, A. Weijters, and W. van der Aalst, "The ProM framework: a new era in process mining tool support," The 26th International Conference (ICATPN 2005) LNCS Vol. 3536, June 2005, pp. 444–454, doi:10.1007/11494744_25.

[8] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting Students Drop Out: a Case Study," The 2nd International Conference on Educational Data Mining (EDM 2009), July 2009, pp. 41–50, ISBN 978-84-613-2308-1.

[9] R.P. Jagadeesh Chandra Bose and W.M.P van der Aalst, "Context Aware Trace Clustering: Towards Improving Process Mining Results," The SIAM International Conference on Data Mining (SDM 2009), April 2009, pp. 401-412.

[10] L. Kaufman and P. J. Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis". by Leonard Kaufman, Peter J. Rousseeuw, March 1990, ISBN: 0-471-87876-6.

[11] D. Luengo and M. Sepúlveda, "Applying Clustering in Process Mining to Find Different Versions of a Business Process That Changes over Time," The Business Process Management Workshops (BPM 2011) Aug. 2011, pp. 153-158 doi:10.1007/978-3-642-28108-2_15.

[12] J. Munoz-Gama, J. Carmona, and W.M.P. van der Aalst, "Conformance Checking in the Large: Partitioning and Topology," The 11th International Conference on Business Process Management (BPM 13), Aug. 2013, pp. 130–145, doi:10.1007/978-3-642-40176-3_11.

[13] M. Pechenizkiy, N. Trčka, E. Vasilyeva, W. van der Aalst and P. De Bra, "Process Mining Online Assessment Data," The 2nd International Conference on Educational Data Mining (EDM 2009), July 2009, pp. 279–288.

[14] M. Reichert, "Visualizing Large Business Process Models: Challenges, Techniques, Applications," The 1st Int'l Workshop on Theory and Applications of Process Visualization (BPM 2012) Sep. 2012, LNCS Vol. 132, pp. 725-736, doi:10.1007/978-3-642-36285-9_73.

[15] M., Romero and C., Ventura, "Data mining in education," The Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol 3, Feb. 2013, pp. 12–27.

[16] G. Seber, "Multivariate Observations," Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[17] M. Song, C. W. Günther, and W.M.P. van der Aalst, "Trace Clustering in Process Mining," Business Process Management Workshops (BPM 2008) Sep. 2008, LNCS Vol. 17, pp 109-120, doi:10.1007/978-3-642-00328-8_11.

[18] V. Southavilay, K. Yacef, and R. A. Calvo, "Process mining to support students' collaborative writing," The 3rd International Conference on Educatinal Data Mining (EDM 2010) June 2010, pp. 257-266.

[19] H. Spath, "Cluster Dissection and Analysis: Theory," FORTRAN Programs, Examples, New York: Halsted Press, 1985.

[20] N. Trčka and M. Pechenizkiy "From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining," The 9th Conference on Intelligent Systems Design and Applications (ISDA 2009), Dec. 2009, pp. 1114–1119, doi:10.1109/ISDA.2009.159.

[21] N. Trčka, M. Pechenizkiy, and W. van der Aalst, "Process Mining from Educational Data (Chapter 9)," Handbook of Educational Data Mining.. CRC Press, 2010, pp. 123–142, doi: 10.1201/b10274-11.

[22] G. M. Veiga and D. R. Ferreira,"Understanding Spaghetti Models with Sequence Clustering for ProM," Business Process Management Workshops (BPM 2009) Sep. 2009, LNCS vol. 43, pp. 92–103, doi:10.1007/978-3-642-12186-9_10.

 ISBN: 978-1-61208-364-3

A Distribution for Service Model

Silvia Maria Prado*, Francisco Louzada†, José Gilberto Rinaldi ‡ and Benedito Galvão Benze §
*Federal University of Mato Grosso
Cuiabá, Brazil
Email: silviamprado@gmail.com
†University of São Paulo
São Carlos, Brazil
Email: louzada@icmc.usp.br
‡ UNESP University
Presidente Prudente, Brazil
Email: gilberto@fct.unesp.br
§ Federal University of São Carlos
São Carlos, Brazil
Email:benze@ufscar.br

***Abstract*—In this paper, we developed a flexible service model for the minimum service time called Minimum-Conway-Maxwell-Poisson-exponential distribution, denoted by MINCOMPE distribution, with the service rate dependent on the state of the system including the idle period. This distribution is a new approach where it is possible to look only at the service and capture variations of the system. In addition, this distribution is to model the dependency between the interarrival and service times. The MINCOMPE distribution contains submodels, such as Minimum-geometric-exponential, Minimum-Poisson-exponential and Minimum-Bernoulli-exponential, which express variations of the system. The properties of the proposed distribution are discussed, including formal proof of its probability density function and explicit algebraic formulas for their reliability and moments. The parameter estimation is based on the usual maximum likelihood method. Simulated and real data are shown to illustrate the applicability of the model.**

Keywords:**Conway-Maxwell-Poisson distribution; MINCOMPE distribution; minimum service time.**

I. INTRODUCTION

In this paper, we studied a specific system where the interarrival times are the same as the service times. In this system, there is a dependency between the interarrival and service times where the service is attached to the arrival. Hence, when the service finishes another customer arrives in the system and enters into the service directly. When the number of customers increases, the service becomes faster and the interarrival time decreases. Therefore, it is necessary to have an adjustement mechanism in order to restablish the balance of the system. The possible adjustments are to change the service rate and/or the opening of new service channels.

We proposed a distribution that describe this system which we called Minimum-Conway-Maxwell-Poisson-exponential distribution, denoted by MINCOMPE distribution, with service rate dependening on the state of the system. The MINCOMPE distribution contains various submodels, which can be obtained by varying the pressure parameter, such as, Minimum-geometric-exponential, Minimum-Poisson-exponential and Minimum-Bernoulli-exponential. This submodels capture the oscillations of the system due to the increase of the number of curtomers.

The MINCOMPE distribution was obtained using a compound of two distributions, the Conway-Maxwell-Poisson for the number of customers, denoted by COM-Poisson, and the exponential distribution for the interarrival time. The main goal was to observe the minimum interarrival times when the number of customers in the system is unknown.

It is necessary to consider the following system for the compound; a single server where the service time is exponential distributed and the mean depends on the system state and is given by $\mu_m = m^{\phi}\mu$, where the number of customers is indicated by m. The degree to which the service rate is affected by the system state is indicated by ϕ and it is called pressure parameter; the arrivals in the system occur at random; the interarrival times are exponentially distributed with mean λ; the customers are served on a First-Come-First-Served (FCFS).

It is generally believed that, the usual queue model has the service rate independent of the system state however, it is a special case when $\phi = 0$ so that $\mu_m = \mu$ for all m. Moreover, when $\phi = 1$ the service rate is directly proportional to the system state and the opening of new service channels. When ϕ values are greater than one, the service rate is more proportional to an increase in work.

In other words, the pressure parameter is a defense mechanism when there is a backlog of work. An increase in effort on the part of the server is an obvious source of increase in service rate.

Moreover, "the Poisson arrivals see Time Averages property, denoted by PASTA, was used, meaning when arrivals are Poisson, the fraction of arrivals who find a process in some state (busy or idle) is equal to the fraction of time the process is in that state", this property is described in [1].

In the literature, there are few references to be considered compound and state dependent service rate. We would like to mention the modeling studies: Jongbloed and Koole [2] studied a call center as a queueing model with Poisson arrivals having an unknown varying arrival rate. Srikanth and Manjunath [3] analyze queueing models where the joint density of the

 ISBN: 978-1-61208-364-3

interarrival time and the service time were described by a mixture of joint densities.

This paper has been organized as follows. In Section A, we presented the MINCOMPE distribution and some of its properties for minimum interarrival time or minimum service time.

In Section B, we derived the expressions for the probability density function, and r-th raw moments of the MINCOMPE distribution.

In Section C, we described the maximum likelihood estimation of the parameters of the model and demonstrated some numerical results with simulation and real data. Finally, Section II contains final remarks.

A. The Distribution for Minimum Service Time

The process can be described as follows. Let be M a random variable denoting the number of customers in the system, $m = 0,1,2\ldots$, with COM-Poisson distribution described in [4] and [5], with probability mass function (pmf) expressed as

$$P_m(M=m;\rho,\phi) = \frac{1}{[Z(\rho,\phi)]}\frac{\rho^m}{(m!)^\phi}, m = 0,\ldots \quad (1)$$

where $Z(\rho,\phi) = \sum_{j=0}^{\infty} = \frac{\rho^j}{(j!)^\phi}$ is normalizing constant, ρ is traffic intensity with $\rho < 1$ and $\phi \in (-\infty,\infty)$. The stability condition $\rho = \lambda/\mu < 1$ means that the arrival rate λ must be less than the service rate μ.

"Note that, the COM-Poisson distribution is undefined when $\rho \geq 1$, $\phi = 0$. Extending ϕ to its two extremities, the COM-Poisson distribution in (1) can be seen as a continuous bridge between the geometric ($\phi = 0$, *with* $0 < \rho < 1$), the Poisson ($\phi = 1$) and Bernoulli ($\phi \to \infty$) distributions. This distribution is overdispersed when $\phi \in [0,1)$ and underdispersion when $\phi \in (1,\infty)$" described in [4].

We assumed that the interarrival times and the service times follow the exponential distribution. Let Y_i, $i = 1,2,\ldots$ be random variables denoting interarrival times exponentially distributed with mean λ and given by

$$f(y_i;\lambda) = \lambda e^{-\lambda y_i}. \quad (2)$$

Most queueing models assume that interarrival times are statistically independent of the service times. However, such an assumption is not always valid. It is also important to take in consideration the possibility of the arrivals of customers being attached to the service as a control of the flow of customers or queue control models. If that is the case, customers will arrive in the system when a services finish. In other words there is no difference between the interarrival time and the service time.

In this paper, we are interested in observing only the minimum interarrival times or minimum service time as this represents how fast the system works and it is given by Y

$$Y = min[Y_1,\ldots Y_m], \quad (3)$$

considering that this is a crucial fact in order to stablish customer loyaty.

Considering the dependence between interarrival times and service times, we derived the distribution of the minimum interarrival or minimum service time given by a compound COM-Poisson distribution for the number of customer and exponential distribution for service time. Therefore, if Y_i and M are densities given by (2) and (1) respectively, the minimum service time distribution is given by

$$f_Y(y,\theta) = \frac{\lambda}{Z(\rho,\phi)}\sum_{m=1}^{\infty} m\frac{\rho^m e^{-m\lambda y}}{(m!)^\phi}, y > 0, \quad (4)$$

where $\theta = (\rho,\lambda,\phi)^T$. In addition, (4) can be rewritten in the form

$$f_Y(y,\theta) = \frac{\lambda Z_1(\rho e^{-\lambda y},\phi)}{Z(\rho,\phi)}E(M_1), \quad (5)$$

where $M_1 \sim COM-Poisson(\rho e^{-\lambda y},\phi)$.

In addition $Z_1(\rho,\phi) = \sum_{j=0}^{\infty} = \frac{(\rho e^{\lambda y})^j}{(j!)^\phi}$ is normalizing constant and $E(M_1) = \rho e^{\lambda y} d\log Z_1(\rho e^{-\lambda y},\phi)/d\rho$.

Therefore, the random variable Y has an MINCOMPE distribution if the cumulative distribution function takes the form

$$F_Y(y;\theta) = 1 - \frac{Z_1(\rho e^{-\lambda y},\phi)}{Z(\rho,\phi)}. \quad (6)$$

We rewritten (4) using the misture of exponential distribution and it is given by

$$f_Y(y,\theta) = \sum_{m=0}^{\infty} v_m f_E(y,m\lambda), \quad (7)$$

where $f_{E_Y}(y,\theta)$ denotes the exponential distribution function with parameter λ and the coefficient v_m was represented by COM-Poisson probabilities given by

$$\begin{aligned} v_m &= v_m(\rho,\phi) = P_m(M=m;\rho;\phi) \\ &= \frac{1}{Z(\rho,\phi)}\frac{m\rho^m}{(m!)^\phi}. \end{aligned} \quad (8)$$

where $\sum_{m=0}^{\infty} v_m = 1$. **Therefore, (6) can be rewritten and takes the form**

$$F_Y(y;\theta) = 1 - \sum_{m=0}^{\infty} v_m e^{-m\lambda y}. \quad (9)$$

The moments of the MINCOMPE distribution can be immediately obtained as linear functions of the exponential moments as

$$E(Y^r) = \lambda^{-r}\Gamma(r+1)\sum_{m=1}^{\infty} v_m m^{-r}. \quad (10)$$

In (11), the reliability function is shown

$$W_Y(y,\theta) = \frac{Z_1(\rho e^{-\lambda y},\phi)}{Z(\rho,\phi)}. \quad (11)$$

Thus, reliability function is the probability of no failures in the interval $[0,y]$ or equivalently, the probability to observe the service time after y time.

When the number of the arrival of customers in the system increases consequentely the interarrival times decrease. Due to this fact, there is a continiously pressure on the server to attend the high demand of work. Therefore, the system has an adjustement mechanism in order to restablish the balance of the system. In this case, the possible adjustements are to

 ISBN: 978-1-61208-364-3

change the service rate and/or open new service channels. These adjustements are captured according to the variations in the pressure parameter and described by corollaries below.

Corollary 1: When $\phi = 0$, the MINCOMPE distribution becames the Minimum-geometric-exponential distribution, denoted by MINGE distribution, for the minimum service time. The COM-Poisson is reduced to a geometric distribution and the service rate is independent of the system state. Therefore, the server is not accelerated and it is not stressed with the arrival of the customers. It is not necessary to do an adjustement in the system.

Therefore, the MINGE distribution is given by

$$f_Y(y,\theta) = \frac{\lambda e^{-\lambda y}(1-\rho)}{(1-\rho e^{-\lambda y})^2}. \quad (12)$$

The reliability function is obtained by

$$W_Y(y,\theta) = \frac{(1-\rho)e^{-\lambda y}}{(1-e^{-\lambda y}\rho)}. \quad (13)$$

When $\phi = 0$ in (10) the raw moments of Y is obtained and it is given by

$$E(Y^r) = \lambda^{-r}\Gamma(r+1)(1-\rho)\sum_{m=0}^{\infty}\sum_{k=0}^{m-1}\rho^m(k)^{-r}. \quad (14)$$

Corollary 2: When the pressure parameter assumed $\phi = 1$, the MINCOMPE distribution is reduced to the Minimum-Poisson-exponential distribution, denoted by MINPE distribution, for the minimum service time. The COM-Poisson is reduced to a Poisson distribution and the service rate is directly proportional to the system state and the server is accelerated. The adjustements mechanisms in order to restablish the balance of the system are opening of new service channel proportional to the number of customers and increase the service rate.

When $\phi = 1$ in (4) we obtained the MINPE distribution and it is given by

$$f_Y(y,\theta) = \frac{\lambda\rho e^{-\rho-\lambda y+\rho e^{-\lambda y}}}{(1-e^{-\rho})}\ y > 0. \quad (15)$$

The reliability function is given by

$$W_Y(y,\theta) = \frac{e^{\rho e^{-\lambda y}}}{(e^{\rho})}. \quad (16)$$

When $\phi = 1$ in (10) the raw moments of Y is obtained and it is given by

$$E(Y^r) = \lambda^{-r}\Gamma(r+1)e^{-\rho}\sum_{m=0}^{\infty}\rho^m(m!)^{-r}. \quad (17)$$

Corollary 3: When $\phi \to \infty$, the MINCOMPE was converted to Minimum-Bernoulli-exponential distribution, denoted by MINBE, for the minimum service time. The COM-Poisson is reduced to a Bernoulli distribution and the service rate is dependent of the system state. The server is accelerated, consequently the service rate increased.

If the random variable Y was defined as (3), replacing $\phi \to \infty$ in (4) and it is given by

$$f_Y(y,\theta) = \frac{\rho\lambda e^{-\lambda y}}{(1+\rho)}. \quad (18)$$

Therefore, the reliability function was presented by

$$W_Y(y,\theta) = \frac{1+\rho e^{-\lambda y}}{(1+\rho)}. \quad (19)$$

The raw moment of the exponential distribution was given by

$$E(Y^r) = \lambda^{-r}\Gamma(r+1)\frac{\rho}{(1+\rho)}. \quad (20)$$

B. Maximum Likelihood Estimation

The maximum likelihood estimation is considered the log-likelihood MINCOMPE distribution in (5) can be written as

$$\begin{aligned}\ell(\theta,y) &= -n\log Z(\rho,\phi) + \sum_{i=1}^{n}\log(Z_1(\rho e^{-\lambda y_i},\phi)) \\ &+ \sum_{i=0}^{n}\log E[M_1] \end{aligned} \quad (21)$$

where $\theta = (\rho,\lambda,\phi)^T$.

Denoted by Z^{θ_i} and $Z_1^{\theta_i}$ first derivatives of Z and Z_1 with aspect to any parameter θ_i of the MINCOMPE distribution. The components of the unit score function $U = (U_\rho, U_\lambda, U_\phi)^T$ is given by

$$U_\rho = -n\frac{Z^\rho}{Z} + \sum_{i=0}^{n}\frac{Z_1^\rho}{Z_1} + \sum_{i=0}^{n}\frac{E[M_1]^\phi}{E[M_1]}, \quad (22)$$

and

$$U_\lambda = \sum_{i=0}^{n}\frac{Z_1^\lambda}{Z_1} + \sum_{i=0}^{n}\frac{E[M_1]^\lambda}{E[M_1]}, \quad (23)$$

and

$$U_\phi = -n\frac{Z^\phi}{Z} + \sum_{i=0}^{n}\frac{Z_1^\phi}{Z_1} + \sum_{i=0}^{n}\frac{E[M_1]^\rho}{E[M_1]}. \quad (24)$$

The numerical computation of the above moments can be easily performed in software packages such as R and Matlab. Numerical maximization of the log-likelihood function is performed with the RS method [6] in the gamlss package. These methods were discussed in detail in [7], and [4].

We show the log-likelihood functions for other models in corollaries below.

Corollary 4:

Where $\phi = 0$, the minimum service time had MINGE distribution and the log-likelihood function is accorded by

$$\begin{aligned}\ell(\rho,\lambda) &= n\lambda - \sum_{i=0}^{n}\lambda y + n\log\lambda(1-\rho) \\ &- 2\sum_{i=0}^{n}\log(1-\rho e^{-\lambda y_i}). \end{aligned} \quad (25)$$

 ISBN: 978-1-61208-364-3

Corollary 5: Where $\phi = 1$, the minimum service time has MINPE distribution and the log-likelihood function is given by

$$\ell(\rho,\lambda) = n\log(\rho\lambda) - n\rho - \sum_{i=0}^{n}\lambda y_i + \rho\sum_{i=0}^{n}e^{\lambda y_i} - n\log(1-e^{\rho}). \quad (26)$$

Corollary 6: Where $\phi \to \infty$, the minimum service time has MINBE distribution and the log-likelihood function is given by

$$\ell(\rho,\lambda) = n\log\lambda\rho - \sum_{i=1}^{n}\lambda y_i - n\log(1+\rho). \quad (27)$$

C. Numerical Results

The numerical results are important to describe the behavior of the model and its applicability in different situations. We presented three pieces of data: the simulated data and two real data; data from a Brazilian supermarket checkout and data from the access to a website.

1) Simulation: For the simulated data, we have chosen $M/M/1$ model [8]. The aim was to look for the data where few customers remained in the queue. This particular set of data was then used to test the new distribution when the pressure parameter took the value $\phi = 0$. In this case, the system was not accelerated and the service rate was independent of the state of the system. Therefore, the $M/M/1$ model was simulated with the intensive traffic $\rho = 0.9$ with the arrival rate 0.9. We have stablished 1,000 arrivals as the ending point of the simulation. The proposal was to adjust the empirical models and it was based on the comparison between the observed and predicted values. The simplest way to make this comparison is graphically, which consists of comparing the reliability function to the Kaplan-Meier estimator. Thus, Figure 1 shows the behavior of the MINGE distribution it is compared with the Kaplan Meier estimates [9] for the simulated data with $\rho = 0.9$. Clearly, the MINGE distribution yields a close concordance with the Kaplan-Meier estimates.

2) Real Data:

- To begin with,the express checkout in the supermarket real data was analyzed. It is often felt that, the minimum service time is one of the key points in order to establish customer loyalty. Therefore, supermarkets use a variety of methods to reduce the service time at checkouts. The most traditional method for example, is the express lines. In the express lines the amount of items which customers can bring to an express checkout counter is limited. When analysing the supermarket checkouts, a particular express checkout presented a similar behavior of the system studied; the service was very fast and many customers entered directely into the service. The remaining number of customers in the queue was insignificant. We have used this data to test the MINGE distribution. The Kolmogorov-Smirnov test was used to prove that the interarrival times and the service times presented an exponential distributions. A total sample of 85 customers were observed with intensity traffic $\rho =$ 0.816 where the mean service was 1.12 minutes. The MINGE distribution was used when the service was independent of the state of the system. The server was able to absorve all the works and it was not necessary to adjust the system. Figure 2 shows the MINGE distribution and the Kaplan-Meier estimates. Indeed, the MINGE distribution has a close concordance with the Kaplan-Meier estimates. The maximum likelihood estimates are given by $\hat{\rho} = 0.876$, $\hat{\lambda} = 0.90$ minutes and the mean $E(Y) = 1.908$ minutes.
- Finally, we analysed accesses to the website "Tendencias Profissionais" [10]. A survey with 26 questions was alocated on this website. Data collection

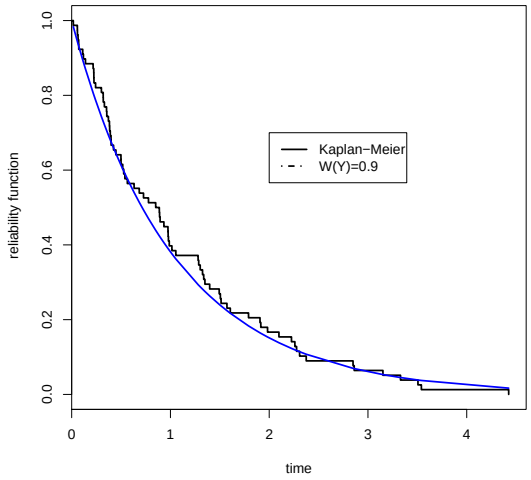


Figure 1: Kaplan-Meier estimates and reliability function $W(y)$ for $\rho = 0.9$.

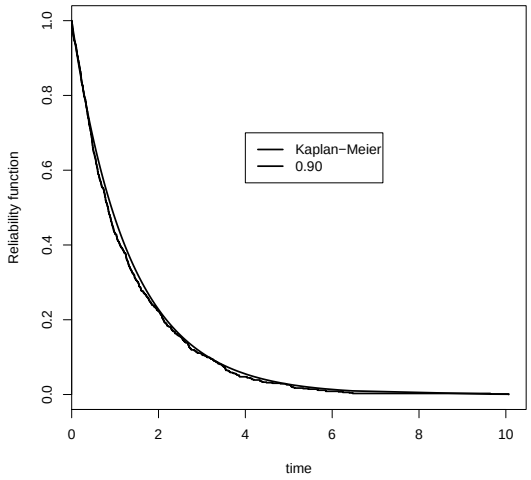


Figure 2: Kaplan-Meier estimates and reliability function $W_Y(y)$ for $\hat{\rho} = 0.876$.

 ISBN: 978-1-61208-364-3

started on 20 October, 2010 and it was available for 20 days. As a general rule, the internet allows the rapid dissemination of information. The survey was distributed through social networks and 1,000 emails were sent to the main communication agencies in Brazil. On the website "Tendencias Profissionais", the arrival time of customers was registered. A sequence of the arrival time was observed. The suitability of the exponential distribution was tested for interarrival times (Y). The Kolmogorov-Smirnov test was used, therefore $D_{max} < D_n^{\alpha}$ was obtained and $D_{max} = 0.114$ and $D_n^{\alpha} = 0.122$ was the critical value. Moreover, the suitability of the Poisson model for the number of customers (M) was tested. A new test for the Poisson distribution [8]. The new test takes into account the non-homogeneity of the process as well as the under-dispertion of data. Therefore, the new test in which $T_{new} = 4\sum_{n=20}^{i=1}(\lambda_i - \bar{\lambda})^2 = 0.011$, where λ_i is the arrival rate per day and $\bar{\lambda}$ is the average arrival rate of 20 days of observation. If $T_{new} > \chi^2_{n-1,1-\alpha}$, the hypothesis $H_0 M \sim Poisson(\lambda)$ is rejected. In this case, it was obtained that $T_{new} = 0.10 < \chi^2_{20-1,0.05}$, the hypothesis H_0 was rejected. Thus, the use of MINPE distribution was justified. Figure 3 shows the MINPE distribution and the Kaplan-Meier estimates. The MINPE distribution yields a close concordance with the Kaplan-Meier estimates. Moreover, the mean rate for answering the survey was 6.5 minutes and the minimum time was 1.2 minutes. In addition, the maximum likelihood estimates were given by $\hat{\rho} = 0.98$, $\hat{\lambda} = 0.449$ minutes and $E(Y) = 1.08$ minutes

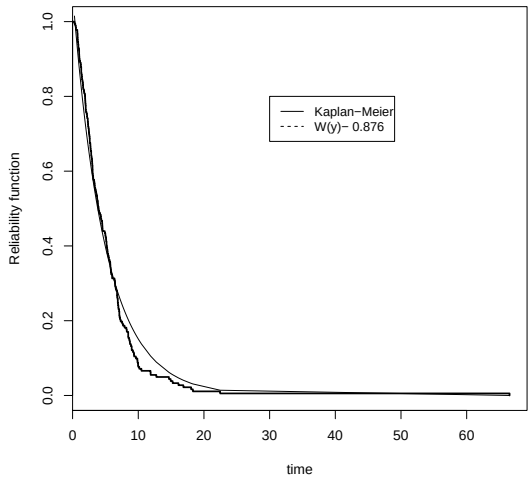


Figure 3: Kaplan-Meier estimates and reliability function $W(y)$ for $\hat{\rho} = 0.98$.

II. Final Remarks

In this paper, we proposed a distribution for the minimum service model with service rate dependent on the state of the system called the Minimum-Conway-Maxwell-Poisson-exponential distribution and denoted by MINCOMPE distribution. This distribution describes the service, not considering the number of customers. In addition, there is a dependence between interarrival times and service times. In other words, the service is attached to the arrival and the interarrival time is the same as the service time. Hence, when the customer arrives in the system, he enters into the service directly. Therefore, it is necessary to have an adjustement mechanism in order to restablish the balance of the system. As a result, the service rate increases and/or new channels of the service can be opened. We studied three situations for the server. Firstly, the pressure parameter took on the zero value, $\phi = 0$ and in this case, the server dit not accelerate and the service was independent from the state of the system. Afterwards, the pressure parameter took on the value of one, $\phi = 1$, accelerating the server and opening new service channels. Finally, the server increased even more the service rate and the pressure parameter assumed the value infinity. The MINCOMPE distribution generalizes other usual distributions for each variation of the pressure parameter, such as the Minimum-geometric-exponential, Minimum-Poisson-exponential and Minimum-Bernoulli-exponential. The properties of the proposed distribution were discussed, including a formal proof of its pdf and moments. An estimation of the parameters was obtained by the maximum likelihood method. In order to illustrate the model. Real and simulated data were set as illustrations of how to fit the MINCOMPE distribution. To conclude, we believe that the MINCOMPE distribution has a practical approach within a service model with the state dependent service rate. In addition, this can be applied to various practical situations.

References

[1] R. W. Wolff, *Stochastic modeling and the theory of queues*. Prentice Hall Englewood Cliffs, NJ, 1989, vol. 14.

[2] G. Jongbloed and G. Koole, "Managing uncertainty in call centres using poisson mixtures," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 307–318, 2001.

[3] S. K. Iyer and D. Manjunath, "Queues with dependency between interarrival and service times using mixtures of bivariates," *Stochastic models*, vol. 22, no. 1, pp. 3–20, 2006.

[4] T. P. Minka, G. Shmueli, J. B. Kadane, S. Borle, and P. Boatwright, "Computing with the com-poisson distribution," *Pittsburgh, PA: Department of Statistics, Carnegie Mellon University*, 2003.

[5] R. W. Conway and W. Maxwell, "A queueing model with state dependence services rates," *The Journal of Industrial Engineering*, vol. XII, no. 2, pp. 132–136, 1961.

[6] R. Rigby and D. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 507–554, 2005.

[7] G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright, "A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 1, pp. 127–142, 2004.

[8] D. Gross and C. Harris, *Fundamentals of queueing theory (Series in probability & statistics)*. Boston: Wiley, 1998.

[9] J. P. Klein and M.-J. Zhang, *Survival analysis, software*. Wiley Online Library, 2005.

[10] F. A. F., P. S.M., F. G. C., K. R.M.I, and M. V.N, "Tendencias profissionais," *IPEA*, vol. 4, no. 1, pp. 273–314, 2012.

 ISBN: 978-1-61208-364-3

Compressed SIFT Feature-based Matching

Shmuel Tomi Klein

Computer Science Department
Bar Ilan University, Israel
Email: tomi@cs.biu.ac.il

Dana Shapira

Computer Science Department
Ashkelon Academic College, Israel
Email: shapird@ash-college.ac.il

***Abstract*—The problem of compressing a large collection of feature vectors so that object identification can further be processed on the compressed form of the features is investigated. The idea is to perform matching against a query image in the compressed form of the descriptor vectors retaining the metric. Specifically, we concentrate on the Scale Invariant Feature Transform (SIFT), a known object detection method. Given two SIFT feature vectors, we suggest achieving our goal by compressing them using a lossless encoding for which the pairwise matching can be done directly on the compressed files, by means of a Fibonacci code. Experiments show that this approach incurs only a small loss in compression efficiency relative to standard compressors requiring a decoding phase.**

Keywords–Data Compression; Feature vectors; SIFT; Fibonacci code.

I. INTRODUCTION

The tremendous storage requirements and ever increasing resolutions of digital images, necessitate automated analysis and compression tools for information processing and extraction. A main challenge is detecting patterns even if they were rotated or scaled, working directly on the compressed form of the image. In a more general setting, a collection of images could be given, and the subset of those including at least one object, which is a rotated or scaled copy of the original object, is sought. An example for the former could be an aerial photograph of a city in which a certain building is to be located, an example for the more general case could be a set of pictures of faces of potential suspects, which have to be matched against some known identifying feature, like a nose or an eyebrow.

Invariance is obtained by using certain transforms, e.g., the one called Scale Invariant Feature Transform (SIFT) by Lowe [1], a high probability object detection and identification method, which is done by matching the query image against a large database of local image features. Lowe's object recognition method transforms an image into a large set of feature vectors, each of which is invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. Feature descriptor vectors are computed for the extracted key points of objects from a set of reference images, which are then stored in a database. An object in a new image is identified after matching its features against this database using the Euclidean L_2 distance.

Query feature compression can contribute to faster retrieval, for example, when the query data is transmitted over a network, as in the case when mobile visual applications are used for identifying products in comparison shopping. Moreover, since the memory space on the mobile device is restricted, working directly on the compressed form of the data is sometimes required.

The rest of the paper is organized as follows. Section 2 reviews some of the related work; Section 3 gives a brief description of SIFT; Section 4 presents our lossless encoding for SIFT feature vectors, especially suited for CFBM; Section 5 presents the algorithm used for compressed pairwise matching the feature vectors without decompression; finally, Section 6 presents results on the compression performance and the last section suggests how to extend this work.

II. RELATED WORK

Wagner et al. [2] developed object recognition algorithms especially designed for a restricted amount of available RAM, such as mobile phones. Wagner uses a fast corner detector for feature detection, and off-line preprocesses the features in different scales, while using only a fixed scale level, matching then on-line the phone's camera scale. Tackling this problem from another angle is by using good known methods for a non restricted Random Access Memory environment, but making them work in a compressed domain.

A feature descriptor encoder is presented in Chandrasekhar et al. [3]. They transfer the compressed features over the network and *decompress* them once data is received for further pairwise image matching. Chen et al. [4] perform tree-based retrieval, using a scalable vocabulary tree. Since the tree histogram suffices for accurate classification, the histogram is transmitted instead of individual feature descriptors. Also, Chandrasekhar et al. [5] encode a set of feature descriptors jointly and use tree-based retrieval when the order in which data is transmitted does not matter, as in our case. Several other SIFT feature vector compressors were proposed, and we refer the reader to [6] for a comprehensive survey. We propose a special encoding, which is not only compact in its representation, but can also be processed directly *without* any decompression.

Figure 1 visually represents our approach as opposed to the traditional one of feature based object detectors and previous research regarding feature descriptors compression. The client uses any feature detector for extracting key points from the

 ISBN: 978-1-61208-364-3

image, and computes the relevant vectors. These features are then sent along a network to the server, where pairwise pattern matching is applied against the stored database, as shown in Figure 1(a). Figure 1(b) depicts the scenario assumed in previous research that deals with compressed feature descriptors: compression is applied to the vectors before transmission, and decompression is performed once the descriptors are received on the server's side. Unlike traditional work, the current suggestion omits the decompression stage, and performs pairwise matching directly on the compressed data, as shown in Figure 1(c). Similar work, using quantization, has been suggested by Chandrasekhar et al. [7]. We do not apply quantization, and rather use a lossless encoding.

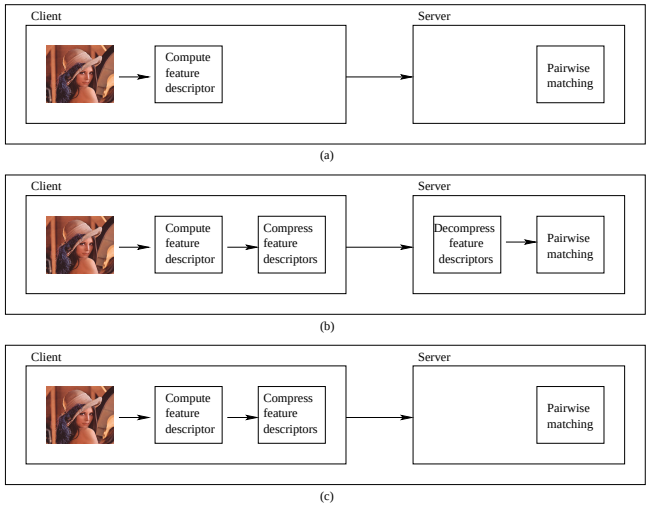


Figure 1. Block diagram showing (a) the traditional image retrieval system, (b) the scenario assumed by previous research, as opposed to (c) the compressed feature based matching problem.

Working on a shorter representation and saving the decompression process may save processing time, as well as memory storage, making sure not to hurt the true positives and false negatives probabilities. Moreover, representing the same set of feature descriptors in less space can allow us keep a larger set of representatives, which can result in a higher probability for object identification by reducing the number of mismatches.

The main idea is to perform the matching against the query image in the compressed form of the feature descriptor vectors so that the metric is retained, i.e., vectors are close in the original distance (e.g., Euclidean distance based on nearest neighbors according to the Best-Bin-First-Search algorithm in SIFT) if and only if they are close in their compressed counterparts. This can be done either by using the same metric but requiring that the compression does not affect the metric, or by changing the distance so that the number of false matches and true mismatches does not increase under this new distance. In the present work, we stick to the first alternative and do not change the L_2 metric used in SIFT.

For the formal description of the general case, let $\{\vec{f_1}, \vec{f_2}, \ldots, \vec{f_n}\}$ be a set of feature descriptor vectors generated using some feature based object detector, and let $\| \quad \|_M$ be a metric associated with the pairwise matching of this object detector. The *Compressed Feature Based Matching Problem* (CFBM) is to find a compression encoding of the vectors, denoted $\mathcal{E}(\vec{f_i})$, and an equivalent metric m so that for every $\epsilon > 0$ there exists a $\delta > 0$ in which $\forall i, j \in \{1, \ldots, n\}$

$$\|\vec{f_i} - \vec{f_j}\|_M < \epsilon \iff \|\mathcal{E}(\vec{f_i}) - \mathcal{E}(\vec{f_j})\|_m < \delta. \qquad (1)$$

This is an extension of the Compressed Pattern Matching paradigm introduced by Amir and Benson [8]. Given a pattern P, a text T and complementing encoding and decoding functions $\mathcal{E}$ and $\mathcal{D}$, the Compressed Matching problem is to locate P in the compressed text $\mathcal{E}(T)$. While the traditional approach searches for the pattern in the decompressed text, i.e., searching for P in $\mathcal{D}(\mathcal{E}(T))$, compressed matching calls for rather compressing the pattern too, and looking for $\mathcal{E}(P)$ in $\mathcal{E}(T)$, with the necessary adaptations. In our case, previous work on feature compression would use complementary encoding and decoding functions $\mathcal{E}$ and $\mathcal{D}$, and apply decompression on the vectors, so that $\mathcal{D}(\mathcal{E}(\vec{f_i})) = \vec{f_i}$.

III. Brief Description of SIFT

Matching features across different images appearing in different scales and rotations is a common problem in computer vision, and SIFT is one of the famous tools dealing with it. The SIFT algorithm first preprocesses the original image in order to construct a *scale space* to ensure scale invariance. SIFT repeatedly generates progressively blurred out images of the original image and resizes it to half the size. The blurred images are used to generate another set of images. The Laplacian of Gaussian (LoG) operation calculates second order derivatives on the blurred images. The blur smoothes out the noise and makes the second order derivative more stable. The LoG operation locates edges and corners in the image, which are used for finding keypoints. However, since calculating the LoG is computationally intensive, it is approximated by the Difference of Gaussians (DoG), calculating the difference between two consecutive scales, resulting in scale invariant keypoints.

Each pixel of the DoG scales is compared to all 26 of its neighbors, 8 neighbors in the current scale image and 18 more in the images of the scales one above and below it. Maxima and minima pixels are chosen as keypoints, which cannot be detected in the lowest or highest scales. Edges and low contrast pixels are eliminated from the set of keypoints. An orientation is calculated for each keypoint, choosing the most dominant one(s) around the keypoint. Any further calculations are done relative to this orientation. This effectively cancels out the effect of orientation, making it rotation invariant.

Highly distinctive vectors are then created for each keypoint as follows. A 16×16 window of pixels around the keypoint is taken. The window is split into sixteen 4×4 windows, each of which used to generate a histogram of 8 bins. Each bin corresponds to a different orientation (first bin for 0-44 degrees, second for 45-89 degrees, etc.), and the gradient orientations are put into these bins. To achieve rotation independence, the keypoint's rotation is subtracted from each orientation, so that each gradient orientation is relative to that of the keypoint. Finally, the 128 values which are attained are normalized.

The object detection and identification is done by pairwise matching the feature vectors of the query image against a large database of local image features using the L_2 norm.

 ISBN: 978-1-61208-364-3

IV. LOSSLESS ENCODING FOR SIFT FEATURE VECTORS

Given two SIFT feature vectors, we suggest achieving our goal to compress them using a lossless encoding so that the pairwise matching can be done directly on the compressed form of the file, by means of a *Fibonacci code* [9]. Note that while the encoding will be different, the metric used in SIFT does not change, or in terms of the above notation, M and m refer to the same Euclidean metric generally denoted as L_2.

A. The Fibonacci Code

The Fibonacci code is a universal variable length encoding of the integers based on the Fibonacci sequence rather than on powers of 2. A subset of these encodings can be used as a fixed alternative to Huffman codes, giving obviously less compression, but adding simplicity (there is no need to generate a new code every time), robustness and speed [10], [9]. The particular property of the binary Fibonacci encoding is that there are no adjacent 1's, so that the string 11 can act like a *comma* between codewords. More precisely, the codeword set consists of all the binary strings for which the substring 11 appears exactly once, at the left end of the string.

The connection to the Fibonacci sequence can be seen as follows: just as any integer k has a standard binary representation, that is, it can be uniquely represented as a sum of powers of 2, $k = \sum_{i \geq 0} b_i 2^i$, with $b_i \in \{0, 1\}$, there is another possible binary representation based on Fibonacci numbers, $k = \sum_{i \geq 0} f_i F(i)$, with $f_i \in \{0, 1\}$, where it is convenient to define the Fibonacci sequence here by $F(0) = 1, F(1) = 2$ and $F(i) = F(i-1) + F(i-2)$, for $i \geq 2$. This Fibonacci representation will be unique if, when encoding an integer, one repeatedly tries to fit in the largest possible Fibonacci number.

For example, the largest Fibonacci number fitting into 19 is 13, for the remainder 6 one can use the Fibonacci number 5, and the remainder 1 is a Fibonacci number itself. So, one would represent 19 as $19 = 13 + 5 + 1$, yielding the binary string 101001. Note that the bit positions correspond to $F(i)$ for increasing values of i from right to left, just as for the standard binary representation, in which $19 = 16+2+1$ would be represented by 10011. Each such Fibonacci representation starts with a 1; so, by preceding it with an additional 1, one gets a sequence of uniquely decipherable codewords.

Decoding, however, would not be instantaneous, because the set lacks the prefix property. For example, a first attempt to start the parsing of the encoded string 1101111111110 by 110 11 11 11 11 would fail, because the remaining suffix 10 is not the prefix of any codeword. So, only after having read 5 codewords in this case (and the example can obviously be extended) would one know that the correct parsing is 1101 11 11 11 110. To overcome this problem, the Fibonacci code defined in [10] simply reverses each of the codewords. The adjacent 1s are then at the right instead of at the left end of each codeword, thus yielding the prefix code {11, 011, 0011, 1011, 00011, 10011, 01011, 000011, 100011, 010011, 001011, 101011, 0000011,...}.

A disadvantage of this reversing process is that the order preserving of the previous representation is lost, e.g., the codewords corresponding to 17 and 19 are 1010011 and 1001011, but if we compare them as if they were standard binary representations of integers, the first, with value 83, is larger than the second, with value 75. At first sight, this seems to be critical, because we want to compare numbers in order to subtract the smaller from the larger. But, in fact, since we calculate the L_2 norm, the *square* of the differences of the coordinates is needed. It therefore does not matter if we calculate $x - y$ or $y - x$, and there is no problem dealing with negative numbers. The reversed representation can therefore be kept.

B. Using a Fibonacci Code for SIFT Vectors

We wish to encode SIFT feature vectors, each consisting of exactly 128 coordinates. Thus, in addition to the ability of parsing an encoded feature vector into its constituting coordinates, separating adjacent vectors could simply be done by counting the number of codewords, which is easily done with a prefix code.

Empirically, SIFT vectors are characterized by having smaller integers appear with higher probability. To illustrate this, we considered the Lenna image (an almost standard compression benchmark) and applied Matlab's SIFT application on it, generating 738 feature vectors. The number of occurrences of 0 was 28,182, and that of the following numbers 1 to 25 is plotted in Figure 2.

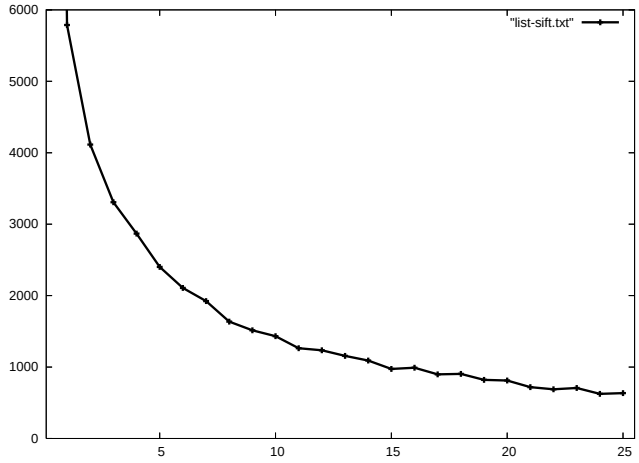


Figure 2. Value distribution in a feature vector.

Feature vectors also contain repeated zero-runs, as could be expected by the high number of zeros. We therefore chose representing a pair of adjacent 0s by a single codeword. That is, the pair 00 is assigned the first Fibonacci codeword 11, a single 0 is encoded by the second codeword 011, and generally, the integer k is represented by the Fibonacci codeword corresponding to the integer $k+2$, for $k \geq 0$. The usual approach for using an universal code, such as the Fibonacci code, is first sorting the probabilities of the source alphabet symbols in decreasing order and then assigning the universal codewords by increasing codeword lengths, so that high probability alphabet symbols are given the shorter codewords. In our case, in order to be able to perform compressed pairwise matching, we omit sorting the probabilities, as already suggested in [11] for Huffman coding. Figure 2 shows that the order is not strictly monotonic, but that the fluctuations are very small. Indeed, experimental results show that encoding the numbers themselves instead of their indices has hardly any influence (0.1% on our test images).

 ISBN: 978-1-61208-364-3

As an example, consider a feature vector of 128 coordinates, the first 20 of which are

$$0, 0, 0, 0, 0, 0, 0, 0, 10, 3, 6, 4, 0, 0, 2, 4, 10, 83, 69, 0, \ldots$$

corresponding to a point of interest of Lenna's Image. The Fibonacci encoding of this feature vector is

11 11 11 11 101011 00011 000011 10011 11

1011 10011 101011 1000101011 0010010011 011 . . .

using 70 bits, rather than 160 bits for the first 20 elements of the original SIFT vector. Note that since all numbers are simply shifted by 2, the difference between two Fibonacci encodings is preserved, which is an essential property for computing their distance in the compressed form.

V. Compressed Pairwise Matching

Given two compressed feature vectors one needs to compute their L_2 norm. Each component is first subtracted from the corresponding component, then the squares of these differences are summed. The algorithm for computing the subtraction of two corresponding Fibonacci encoded coordinates A and B is given in Figure 3. We start by stripping the trailing 1s from both, and pad, if necessary, the shorter codeword with zeros at its right end so that both representations are of equal length. Note that the term first, second and next refer to the order from right to left.

```
Sub(A, B)
scan the bits of A and B from right to left
a1 ⟵ first bit of A
a2 ⟵ second bit of A
while bits of A not empty
{   a3 ⟵ next bit of A
    b1 ⟵ next bit of B
    a1 ⟵ a1 − b1
    a2 ⟵ a1 + a2
    a3 ⟵ a1 + a3
    a1 ⟵ a2
    a2 ⟵ a3
}
b ⟵ value of last 2 bits of B
if b ≠ 0 then b ⟵ 2 − b
return 2 ∗ a1 + a2 − b
```

Figure 3. Subtraction of Fibonacci Codewords.

At the end of the while loop, there are 2 unread bits left in B, which can be 00, 10 or 01, with values 0, 1 or 2 in the Fibonacci representation, but when read as standard binary numbers, the values are 0, 2 and 1. This is corrected in the commands after the while loop of the algorithm. The evaluation relies on the fact that a 1 in position i of the Fibonacci representation is equivalent to, and can thus be replaced by, 1s in positions $i + 1$ and $i + 2$. This allows us to iteratively process the subtraction, independently of the Fibonacci number corresponding to the leading bits of the given numbers. Processing is, therefore, done in time proportional to the size of the compressed file, without any decoding.

As an example, consider the numbers $A = 130$ and $B = 65$, encoded by the strings representing 132 and 67, which are 10001001011 and 1010100011, respectively. Figure 4 shows the results of applying the subtraction algorithm on A and B, which appear, in their reduced form (without trailing 1, but with B padded by 0 to get to the same length) in the boxed first line and last column. At the end, b_1 is assigned the value 1, and the result is indeed $130-65 = 65 = 2*25+16-1$. Note that had we subtracted A from B, the values in columns a_1 and a_2 would be negative or 0 (except in the first row), but the algorithm would still work correctly. In that case, the values in the last line would be -14 and -25, and indeed $65 - 130 = -65 = 2 * (-25) - 14 - 1$.

							a_3	a_2	a_1	b_1
1	0	0	0	1	0	0	1	0	1	0
	1	0	0	0	1	0	0	2	1	1
		1	0	0	0	1	0	0	2	0
			1	0	0	0	1	2	2	0
				1	0	0	0	3	4	0
					1	0	0	4	7	1
						1	0	6	10	0
							1	10	16	1
								16	25	

Figure 4. Example of direct differencing.

```
L2Norm(V1, V2)
while V1 and V2 are not empty
{   remove first codeword from V1
                and assign it to A
    remove first codeword from V2
                and assign it to B
    if A ≠ B then
        if A = 11 then
            S ⟵ Sub(B, 011)
            V1 ⟵ 011 ‖ V1
        else if B = 11 then
            S ⟵ Sub(A, 011)
            V2 ⟵ 011 ‖ V2
        else S ⟵ Sub(A, B)
        SSQ ⟵ SSQ + S^2
}
return √SSQ
```

Figure 5. Compressed differencing of the coordinates.

To calculate the L_2 norm, the two Fibonacci encoded input vectors have to be scanned in parallel from left to right. In each iteration, the first codeword (identified as the shortest prefix ending in 11) is removed from each of the two input vectors, and each pair of coordinates is processed according to the procedure Sub(A, B) above. The codeword 11, representing two consecutive zeros, needs a special treatment only if the other codeword, say B, is not 11. In this case, 11 should be replaced by two codewords 011, each representing a single zero. We thus perform Sub$(B, 011)$, and then concatenate the second 011 in front of the remaining input vector, to be processed in the following iteration. The details appear in the algorithm of Figure 5, where $\|$ denotes concatenation and SSQ is initialized to 0.

 ISBN: 978-1-61208-364-3

VI. COMPRESSION PERFORMANCE

We considered three images for our experiments: *Lenna*, *Peppers* and *House*, which were taken from the Signal and Image Processing Institute Image Data Base [12]. We first applied SIFT on all images receiving 737, 872, and 991 interest points, respectively. Table 1 presents the compression performance of our Fibonacci encoding suitable for compressed matching as compared to other compressors. The second column shows the original sizes of the SIFT feature vectors in bytes. The third column, headed Fib, presents the compression performance, as a percentage of the original size, in which each number is represented by its Fibonacci encoding, which is useful for compressed pairwise matching. To evaluate the compression loss due to omitting the sorting of the frequencies, we considered the compression where each symbol is encoded using the Fibonacci codeword assigned according to its position in the list of decreasing order of frequencies. These values appear in the 4th column headed Ordered Fib.

For comparison, the compression achieved by a Huffman code is also included in the fifth column as a lower bound. As can be seen, encoding the numbers themselves instead of their indices induces a negligible compression loss. The high probability for small integers also reduces the gap between the performances of Fibonacci and Huffman codes.

TABLE I. COMPRESSION EFFICIENCY OF THE PROPOSED ENCODINGS (IN PERCENT OF ORIGINAL SIZE).

Image	Original Size	Fib	Ordered Fib	Huffman	gzip	bzip
Lenna	236,382	27.82	27.78	26.2	34.7	30.7
Peppers	279,422	27.3	27.2	25.7	34.3	30.4
House	325,778	29.5	29.4	27.3	35.6	31.7

The last two columns give the compression performances of gzip and bzip2. These are adaptive compression schemes, and as such no real competitors to Huffman or Fibonacci coding: while their performance on text files is often superior, taking advantage also of the order in which the characters appear, and not just of their frequencies, they cannot be used when direct access to a part of the compressed file is required, as in our case of SIFT feature vectors, and they require a sequential scan from their beginning for the decoding. In this particular case, their compression is also worse than that of Huffman or Fibonacci. This can be explained by the fact that they need to encode also the separating blanks or newlines between the elements of the feature vectors, which constitute a substantial part of the files, whereas Huffman and Fibonacci encode the elements themselves, and not their representations, so the original file can be reconstructed without having to encode the separators explicitly.

VII. CONCLUSION AND FUTURE WORK

We have dealt with the problem of compressing a set of feature vectors known as SIFT, under the constraint of allowing processing the data directly in its compressed form. Such an approach is advantageous not only to save storage space, but also to the manipulation speed, and in fact improves the whole data handling from transmission to processing.

Our solution is based on encoding the vector elements by means of a Fibonacci code, which is generally inferior to Huffman coding from the compression point of view, but has several advantages, turning it into the preferred choice in our case: (a) simplicity – the code is fixed and need not be generated anew for different distributions; (b) the possibility to identify each individual codeword – avoiding the necessity of adding separators, and not requiring a sequential scan; (c) allowing to perform subtractions using the compressed form – and thereby calculating the L_2 norm, whereas a Huffman code would have to use some translation table.

The experiments suggest that there is only a small loss, of 6–8%, in compression efficiency relative to the optimal Huffman codes, which might be worth a price to pay for the improved processing. Relative to other standard compressors, like gzip or bzip, there is even an improvement in compression, contrarily to what one might expect on text files, for example. This is due to the fact that the separators between the vector elements need not be encoded in the Fibonacci approach.

The basic techniques of the present work can be extended to a different, yet related problem: the *Compressed Approximate Pattern Matching* paradigm. When searching for a pattern in a given text one may also be interested in locating strings that are not completely identical to the original pattern, but are quite similar. In the literature, this problem is referred to as *Approximate Pattern Matching*, which is to find all occurrences of substrings in a given text T that are at a given "distance" k or less from a pattern P under some metric. A common choice is the edit distance metric, in which the distance between two strings is defined as the minimum number of insertions, deletions or substitutions of single characters performed on one of the strings in order to convert it to the other. The case where $k = 0$ corresponds to the classical pattern matching problem.

The *Compressed Approximate Matching Problem* (CAMP) is locating *similar* patterns to the searched one working directly on the compressed form of the text. Defining *similarity* formally necessities the existence of a metric so that if the distance between two patterns under this metric is small, searching for one of them in the compressed form of the file will be able to locate both patterns. Approximate compressed pattern matching was first introduced by Amir and Benson [8] as an open problem. It has been solved for many cases, e.g., for byte Huffman coding of words [13], for run length encoded strings [14], for Lempel-Ziv compressed text in [15][16], and Straight Line Programs [17][18].

More formally, given a pattern P, a compressed text $\mathcal{E}(T)$, and a metric $\| \quad \|_M$, the CAMP is to locate all patterns Q in $\mathcal{E}(T)$ so that $\|P - Q\|_M \leq \epsilon$ for some $\epsilon \geq 0$. This is a generalization of the compressed pattern matching problem in which $\epsilon = 0$.

A tempting definition is dealing with two metrics, $\| \quad \|_M$ and a corresponding metric $\| \quad \|_m$ so that if $\|P - Q\|_M \leq \epsilon$ for some $\epsilon \geq 0$ in T, then there exist a corresponding metric $\| \quad \|_m$ and $\delta \geq 0$ so that $\|\mathcal{E}(P) - \mathcal{E}(Q)\|_m \leq \delta$ in the compressed file $\mathcal{E}(T)$. However, this raises some difficulties, as an occurrence of $\mathcal{E}(Q)$ in the encoded file does not necessarily correspond to an occurrence of an approximated pattern. For example, if the encoded file uses Huffman coding, an occurrence of the compressed pattern $\mathcal{E}(P)$ might appear in the encoded file, without implying that there is a corresponding occurrence of the original pattern in the original file, since

 ISBN: 978-1-61208-364-3

$\mathcal{E}(P)$ is not necessarily aligned on codeword boundaries. We intend to deal with these extensions in future work.

REFERENCES

[1] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60 (2), 2004, pp. 91–110.

[2] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg, "Pose tracking from natural features on mobile phones," in Proceedings of the International Symposium on Mixed and Augmented Reality, 2008, pp. 125–134.

[3] V. Chandrasekhar et al., "Transform Coding of Image Feature Descriptors," in Visual Communications and Image Processing, vol. 7257 (1), 2009, pp. 725 710–725 710–9.

[4] D. M. Chen et al., "Tree Histogram Coding for Mobile Image Matching," in Data Compression Conference, DCC–09, 2009, pp. 143–152.

[5] V. Chandrasekhar et al., "Compressing Feature Sets with Digital Search Trees," in ICCV Workshops, 2011, pp. 32–39.

[6] V. Chandrasekhar et al., "Survey of SIFT compression schemes," in Int. Workshop on Mobile Multimedia Processing (WMMP), 2010.

[7] V. Chandrasekhar et al., "Compressed Histogram of Gradients: A Low-Bitrate Descriptor," International Journal of Computer Vision, vol. 96(3), 2012, pp. 384–399.

[8] A. Amir and G. Benson, "Efficient two-dimensional compressed matching," in Data Compression Conference DCC–92, Snowbird, Utah, 1992, pp. 279–288.

[9] S. T. Klein and M. Kopel Ben-Nissan, "On the Usefulness of Fibonacci Compression Codes," The Computer Journal, vol. 53, 2010, pp. 701–716.

[10] A. S. Fraenkel and S. T. Klein, "Robust universal complete codes for transmission and compression," Discrete Applied Mathematics, vol. 64, 1996, pp. 31–55.

[11] S. T. Klein and D. Shapira, "Huffman Coding with Non-Sorted Frequencies," Mathematics in Computer Science, vol. 5(2), 2011, pp. 171–178.

[12] [Online]. Available: http://sipi.usc.edu/database/

[13] E. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates, "Fast and flexible word searching on compressed text," ACM Trans. Inform. Syst. (TOIS), vol. 18 (2), 2000, pp. 113–139.

[14] V. Mäkinen, G. Navarro, and E. Ukkonen, "Approximate Matching of Run-Length Compressed Strings," Algorithmica, vol. 35 (4), 2003, pp. 347–369.

[15] G. Navarro and M. Raffinot, "A general practical approach to pattern matching over Ziv-Lempel compressed text," in Proceedings of Combinatorial Pattern Matching (CPM), 1999, pp. 14–36.

[16] J. Kärkkäinen, G. Navarro, and E. Ukkonen, "Approximate string matching on Ziv-Lempel compressed text," Discrete Algorithms, vol. 1 (3-4), 2003, pp. 313–338.

[17] P. Bille et al., "Random access to grammar-compressed strings," in Symposium on Discrete Algorithms (SODA), 2011, pp. 373–389.

[18] T. Gagie, P. Gawrychowski, C. Hoobin, and S. J. Puglisi, "Faster Approximate Pattern Matching in Compressed Repetitive Texts," ISAAC, 2011, pp. 653–662.

 ISBN: 978-1-61208-364-3

Privacy Friendly Mobile Intelligent Advertising Framework

Preethi Satishchandra, Sanjay Addicam, Kalpana Algotar
Retail Services Division
Intel Corporation
Chandler, USA
emails: {preethi.satishchandra, addicam.v.sanjay, kalpana.a.algotar}@intel.com

***Abstract*— This paper explains a framework known as "Mobile Intelligent Advertising Framework" which provides personalized recommendations to the user in a privacy friendly manner. Mobile devices, such as smartphones, and tablets are everywhere now. So, retailers and advertisers want to rely more on mobile data to recommend products to their consumers and most importantly they want to understand consumer interests to recommend meaningful ones. Our framework includes an Android app, which tries to understand the user passively using the available mobile data that the user gives access to, such as browsing history, accelerometer data, call log history, etc., and recommends a product after data analysis. The app can also communicate with any digital sign nearby which triggers the sign to play targeted ads to the user viewing the sign instead of random ones. This paper explains the important components of the framework briefly and presents an overview of current state of the art in capturing user's interests through mobile data for providing relevant recommendations.**

Keywords-Mobile; Intelligent Advertising Framework; Digital Signs; Data Mining; Lifestyle Analysis; Privacy.

I. INTRODUCTION

In recent years, mobile devices such as smartphones and tablets have grown significantly and they present new opportunities and challenges. This applies to the field of retail too where retailers and advertisers now not only have a new medium to showcase their products and ads to their consumers but also should give importance to new and important fields such as mobile data to show relevant stuffs to the consumers. Mobile data is a good source of data for providing personalized recommendations. Unlike random ones, these personalized targeted recommendations such as ads can provide more value to the consumers thereby benefitting both retailers and advertisers too. There are few applications out there with context-aware platforms, such as Qualcomm's Gimbal [7] and Google Now [6]. But, these do not address privacy as our framework attempts to do by handling all the analysis locally within the device.

Our paper is organized in 3 sections. The first section is Introduction; the second section explains the entire framework, which is further divided into many subsections. The last section is the conclusion and future work.

Section 2 begins by explaining the entire framework briefly. This is followed by describing the data needed for analysis, various data sources and its collection. This is followed by the description of the lifestyle analysis and the lifestyle model. Using this model, lifestyle analysis is done based upon which relevant recommendations are provided to the user. This is followed by the subsection, describing the various tools that were used to build the framework - RapidMiner, Lingpipe. The following sub section explains the communication between the digital sign and the mobile phone, to play targeted ads on the sign instead of random ones which is followed by the explanation of the privacy importance emphasized by our framework which distinguishes it from many other existing context-aware platforms and finally the subsection presenting experiments and results.

II. MOBILE INTELLIGENT ADVERTISING FRAMEWORK (MIAF)

Big data in terabytes and petabytes is floating around from various sources, which can provide valuable insights on current and future trends of an industry. This applies to many industries, such as healthcare, transportation, retail, finance, and many more. Data have proved to be important in mobile too, which can be utilized to provide many recommendations, such as ads, fitness updates, etc. These mobile data can come from different sources and sensors. Some of them are call log data, browsing history, accelerometer data, battery usage, ambient light, app usage, location data and much more. Developing an intelligent framework which fetches the right data, integrates data from various sources, does analysis and provides relevant recommendations in a privacy friendly manner is not trivial. Our framework approaches to solve this problem.

As essential part of MIAF, an Android app was created to facilitate data analysis within the mobile. When a user installs this app, he or she has to provide some minimum information to get recommendations which is age and gender bracket, e.g., Adult Male. The app can access only the part of mobile data to which the user gives permission to.

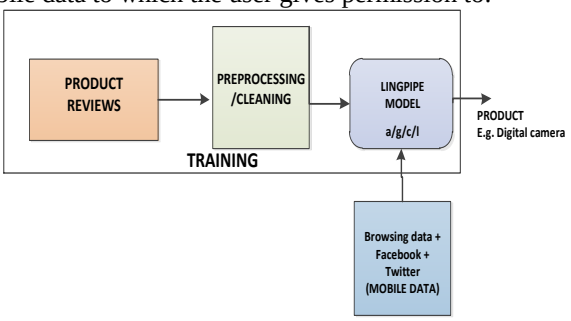


Figure 1. MIAF Architecture

Fig. 1 explains the overall MIAF architecture briefly. We are collecting data, such as product reviews, cleaning the data and training a classification engine based on Lingpipe [3], a text processing tool kit suitable for mobiles. The products selected are grouped into different age, gender, city, lifestyle groups (a/g/c/l, respectively). Lingpipe classification models for different a/g/c/l groups are built by training the model with respective products. Currently, we have around 80 a/g/c/l profiles. So, there are 4 age groups (Child, Young Adult, Adult, and Senior), 2 genders, 2 cities and 5 lifestyles. For example, adult male, from New York with social lifestyle. So, when a user installs the app, information such as age, gender, location is easily captured and further lifestyle analysis of the user is done. Based on the user's profile, the respective trained Lingpipe model is imported to the mobile. The new real-time mobile data, such as browsing history, Facebook and Twitter data are processed and passed into this model to get relevant product recommendations. So, the entire mining and analysis happens locally without the data leaving the device. Besides providing product recommendations, the app can also communicate to any digital sign nearby to play relevant ads and also this app can be integrated behind any retailer's app who is interested in providing personalized recommendations to their loyal consumers. This framework also includes many other features, such as targeted advertising based on facial detection [1][9]; this paper mainly concentrates on the mobile part of the framework.

A. Data

The Lingpipe classification engine requires training data to provide product and lifestyle recommendations. Product reviews fetched from various websites served as a good source of training data using which valuable product insights were extracted for each of the product category. Reviews were preferred instead of product description as they can be correlated with the user's data more and also reviews have become more important as consumers now trust reviews more than description before purchasing a product. A web crawler developed using PHP [14] was used to fetch the reviews for a set of preselected products. Further data based on different demographics (a/g/c) were obtained from Neilsen Company [5]. These data comprised of different kinds of mobile data, such as app usage data, browsing history, battery usage, call log data, etc.

B. Lifestyle Analysis

Knowing the lifestyle of a user helps to recommend better products as consumers usually buy products based on their lifestyles, e.g., a fitness cautious person may prefer product, such as running shoes, sports good, etc. Determining the lifestyle of a user after doing analysis of mobile data is not a trivial problem. MIAF architecture can recommend different lifestyles to a user which are social, talks a lot, travels a lot, fitness cautious and busy. The varied data sources used to determine lifestyle are call log (duration over some period), battery usage (percentage over some period), accelerometer (jogging, running or walking) and number of locations (e.g., number of distinct locations visited in a day) data. Table 1 depicts how data for recommending lifestyle looks for particular demographics after collecting and doing some analysis in a day.

TABLE I. TABLE FOR LIFESTYLE ANALYSIS

Data sources	**Young Adult/Male/New York**		
Call log (min)	23	100	130
Battery usage (percentage)	50	40	80
Accelerometer	Jogging	Walking	Walking
App Usage (Social)(percent)	3	10	4
App Usage (Talks a lot)	4	10	20
App Usage (Productivity)	3	10	2
App Usage (Fitness)	5	12	4
App Usage (Travels)	4	15	10

C. Lifestyle Model

A lifestyle model was built using the historical data obtained from Neilsen [5] company. The varied data sources used to compute lifestyle were obtained based on different demographics, as shown in Table 1. To build this model, a weighted table was constructed by clustering each of the data sources into 5 groups (since 5 lifestyles) for every a/g/c group. K means clustering algorithm [10] was used and the 5 centroids of the 5 clusters were assigned as weights for each of the data sources. Table 2 shows the weights computed for some of the data sources after clustering the historical data for a particular a/g/c group. Such weighted tables are created for all the data sources for every a/g/c profile.

TABLE II. SAMPLE TABLE FOR LIFESTYLE MODEL

Data sources	**Social**	**Talks a Lot**	**Travels**	**Fitness**	**Busy**
Call log (min)	0.3	0.2	0.4	0.1	0.5
Battery usage (percentage)	0.4	0.8	0.3	0.5	0.5

These weights are used as multipliers; so, when the new user mobile data comes (all in numerical values), they are multiplied with these multipliers and finally summed up to determine one lifestyle. The lifestyle with the greatest weight will be selected. This is explained in Table 3; if Sum1 is the greatest, that particular user is tagged with social lifestyle. The weights are updated as and when, new data is collected and added to the historical data collection.

TABLE III. SAMPLE LIFESTYLE TABLE WITH WEIGHTS

Data sources	**Social**	**Talks a Lot**	**Travels**	**Fitness**	**Busy**
Call log – **50min**	50*w1	50*w2	50*w3	50*w4	50*w5
Battery usage**40%**	40*w1	40*w2	40*w3	40*w4	40*w5
App Usage (Social) **30%**	30*w1	30*w2	30*w3	30*w4	30*w5

 ISBN: 978-1-61208-364-3

App Usage (Talks a lot) **20%**	20*w1	20*w2	20*w3	20*w4	20*w5
App Usage (Productivity)**10%**	10*w1	10*w2	10* w3	10*w4	10*w5
App Usage (Fitness) **20%**	20*w1	20*w2	20*w3	20*w4	20*w5
App Usage (Travels) **20%**	20*w1	20*w2	20*w3	20*w4	20*w5
Sum	Sum1	Sum2	Sum3	Sum4	Sum5

D. Rapidminer

Rapidminer [2], an open source data mining tool, served as a scratchpad to understand and build some data mining models such as classification and clustering. This understanding eased the development process of MIAF app. Product reviews fetched from various websites were processed and a classification model based on Support Vector Machine (SVM) [11] was developed in Rapidminer. The measure of accuracy of the model gave some confidence to build the same on the Android Mobile platform. SVM outperformed other classification engines, such as K- Nearest Neighbor (KNN) [13] and Naïve Bayes [12].

Rapidminer was also used for implementing clustering algorithm for detection of lifestyle as discussed above. K – Means clustering algorithm served well as we wanted to divide the data into 5 (k = 5) lifestyles to find 5 centroids.

E. Lingpipe

Lingpipe [3] is a Java-based framework text processing tool kit which served well for Android data mining. Its Software Development Kit (SDK) is easy to integrate with the mobile platform. Lifestyle models were built based on different demographics as discussed above. Further, for various a/g/c/l profiles, Lingpipe product classification models are built for product recommendation and these are built offline in the cloud. Depending upon the user profile respective trained model is imported to the mobile through MIAF. Cloud is used to build all the pre-trained data models and storage of the same. None of the user information is pushed into the cloud as it is handled within the device itself. By having all the training handled in the cloud there is no overload on the mobile Central Processing Unit (CPU) and the battery, as training of the data model requires more CPU overhead. User privacy is also preserved by utilizing the cloud only for the required and by not pushing everything to it.

TF-IDF (Term Frequency – Inverse Document Frequency) classifier [15] was used as it had better classification accuracy then other classifiers, such as Naïve Bayes [12]. SVM was not supported by Lingpipe. Tokenization, transformation of tokens, stop word removal, stemming and filtering of tokens was used for cleaning of text. Depending on the user profile (a/g/c/l), the respective model is imported from the cloud to the mobile to provide product recommendations. This model keeps updating, too, periodically, if needed. For example, if a user previously was in New York and recently moved to San Francisco, a new product model will be imported to the user mobile to provide relevant product recommendations.

MIAF app receives mobile data which are made up of browsing history keywords, twitter keywords and Facebook keywords. Browsing pages are parsed and top n words based on TF-IDF score are extracted, same with Facebook page and Twitter page. All these top TF-IDF words extracted are concatenated and passed into the model imported into the mobile to recommend products. All the word vectors are stored in SQLite database [16] within the device. So, no user mobile data leaves the mobile. The data are cleaned from SQLite every week, so that there is no overload on the database and not cleaned very frequently too for the creation of meaningful user profiles. Currently, in a device, products can be recommended from 100 product lists.

F. Talking of Mobile and Digital Sign

Digital signs [4] are part of every industry where they are present in airports, shopping malls and retailer shops. MIAF app can make the ads played on the digital sign more valuable. To bring more value for ads, we wanted our app to talk with the digital sign. Extensible Messaging and Presence Protocol (XMPP) [17], a messaging protocol was used to establish this communication channel. So, when a user with MIAF app walks near any digital sign communication channel between them is built and the ad played on the digital sign will be related to the ad recommended on the user's mobile. This is explained in Fig. 2, where the communication between the mobile and the digital sign is established using XMPP, iPhone ad is played on the sign after the word "iPhone" is transferred from mobile to the sign.

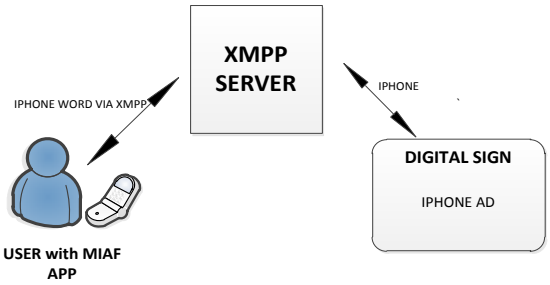


Figure 2. Communication between Phone and Digital Sign

G. Privacy

Most of the applications that exist now do not handle user's data securely, as most of the mobile data analysis are pushed into the cloud due to limited mobile computing resources. Our app handles all the required analysis locally in the mobile with minimal effect on the battery usage. Only a single word vector such as iPad, sun glass, etc., goes out of the device. No user data are withheld. The user can opt in and opt out of the service anytime. The products or ads are targeted in a more generalized way, where the targeting is done based on demographics, but not for any particular individual. We are looking into adding more levels of security and privacy by encrypting the data and also investigating on integrating with McAfee [8] software on Android.

 ISBN: 978-1-61208-364-3

H. Experiments and Results

MIAF app was shared among a few employees within the company for feedback and also among few customers. The results have been positive. Due to time constraints and other issues we have not documented the results for publishing purposes. But, we have conducted an experiment to prove targeted advertising is more relevant then non – targeted advertising based on AVA [1] technology, which is an Intel based technology used to target ads, based on demographics. The experiment was conducted in a supermarket over a period of 9 months. There was monthly sale increase of 16% and 3% monthly viewership time increase with targeted ads compared to non- targeted ones and the results are published [9]. By adding the mobile part, we believe sales and viewership can increase more bringing in more value to the consumers and the retailers too. Further, we compared MIAF to other similar applications out there. Google Now [6] has the proximity beacon missing and with Gimbal [7], the analytics apps need to be created by the customer and most importantly both the applications do not handle the analysis locally within the device like MIAF.

III. CONCLUSION AND FUTURE WORK

In this paper, we have attempted to explain a software framework for mobiles and digital signage using which relevant and personalized recommendations can be reached to the right consumer at the right time in a privacy friendly manner. With the limited data available, user's interest is captured passively by analyzing the lifestyle of the user and all this analysis happens locally within the device to protect the privacy. Further, based on lifestyle relevant products are recommended. In the future, we want to make this application more efficient and scalable by adding more features such as Graph databases and also by expanding the product lists from 100's to 1000's. We are also extending the framework to not only recommend ads or products, but also predict next user's activity, e.g., a user usually calls at noon every day; so, it is noon now and he/she may want to call. In the future, we want to extend MIAF to other platforms than Android.

REFERENCES

[1] P. Tian, S.V. Addicam, K. Chiranjeevi and S. Malik, "Intelligent Advertising Framework for Digital signage," in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2012, pp. 1532–1535.

[2] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks," in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), August 2006, pp. 935–940.

[3] Alias-i. 2008, LingPipe 4.1.0 - http://alias-i.com/lingpipe [retrieved: May, 2014]

[4] Digital Signage - http://en.wikipedia.org/wiki/Digital_signage [retrieved: May, 2014]

[5] Neilsen - http://www.nielsen.com/us/en.html [retrieved: June, 2014]

[6] Google Now - http://www.google.com/landing/now/# [retrieved: May, 2014]

[7] Qualcomm Gimbal - https://www.gimbal.com/ [retrieved: May, 2014]

[8] McAfee - http://www.mcafee.com/us/ [retrieved: May, 2014]

[9] K. Algotar, S.V. Addicam and P. Satishchandra, "Case Study: Using Video Analytic Data to Target Advertisements," in proceedings of Journal of Emerging Trends in Computing and Information Technology, vol. 5, no. 3, March 2014, pp. 206–209.

[10] K-means Clustering - http://en.wikipedia.org/wiki/K-means_clustering [retrieved: June, 2014]

[11] Support Vector Machine - http://en.wikipedia.org/wiki/Support_vector_machine [retrieved: June, 2014]

[12] Naïve Bayes Classifier - http://en.wikipedia.org/wiki/Naive_Bayes_classifier [retrieved: June, 2014]

[13] KNN - http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [retrieved: June, 2014]

[14] PHP - http://www.php.net/ [retrieved:May, 2014]

[15] TFIDF Classifier - http://alias-i.com/lingpipe/docs/api/com/aliasi/classify/TfIdfClassifierTrainer.html [retrieved: May, 2014]

[16] SQLite database - http://www.sqlite.org/ [retrieved: June, 2014]

[17] XMPP - http://xmpp.org/ [retrieved: May, 2014]

 ISBN: 978-1-61208-364-3

Data Leakage Detection Using Information Retrieval Methods

Adrienn Skrop
Department of Computer Science and Systems Technology
University of Pannonia
Veszprém, Hungary
skrop@dcs.uni-pannon.hu

Abstract— **Data leakage is an uncontrolled or unauthorized transmission of classified information to the outside. It poses a serious problem to companies as the cost of incidents continues to increase. Different solutions have been developed to prevent data loss; however none of them can provide absolute protection due to insider negligence. It is essential to discover data leakage as soon as possible, thus the purpose of this research is to design and implement a data leakage detection system based on different, semantic driven information retrieval models and methods. After describing briefly the idea, the architecture of the system and potential methods are discussed.**

Keywords—data leakage; interaction information retrtieval; hyperbolic information-retrieval; cryptography.

I. INTRODUCTION

Data or information leakage can be defined as an uncontrolled, unauthorized transmission of classified information to the outside or simply an unauthorized dissemination of information. Data leakage can be described by many closely related terms, as the following definitions show. Information leak can be an uncontrolled, unauthorized transmission of classified information from a data center or computer system to the outside. Such leakage can be accomplished by physical removal of data storage devices or even plain old human memory [5]. Data breach is also an unauthorized dissemination of information. It may be due to an attack on the network or outright theft of paper documents, portable disks, USB drives or laptops [12]. An information exposure is the intentional or unintentional disclosure of information to an actor that is not explicitly authorized to have access to that information [13]. Data exfiltration, also called data extrusion, is the unauthorized transfer of data from a computer. Such a transfer may be manual and carried out by someone with physical access to a computer or it may be automated and carried out through malicious programming over a network [11]. Industrial espionage is the theft of trade secrets by the removal, copying or recording of confidential or valuable information in a company for use by a competitor [24]. As the definitions indicate, data leakage can occur in many forms and in any place. Thus, a number of solutions have been developed to prevent data loss. On one hand, encryption may prevent lost or stolen data from being viewed or used by non-authorized individuals. On the other hand, different data leakage products help monitor, manage, and protect data to minimize the risks of data loss and ensure compliance with security policies [26]. However, none of these solutions can provide absolute protection. According to a Symantec study, more than 40 per cent of data breaches were estimated to be due to insider negligence [25]. The purpose of this research is to design and implement a data leakage detection system based on different information retrieval models and methods.

In Section 2, the problem of data leakage is presented. Section 3 shows a potential architecture of a data leakage detection system. Section 4 presents the methods that are planned to be used in the system. Section 5 concludes the paper and presents ideas for future work.

II. DATA LEAKAGE PREVENTION

Data leakage is an incident in which sensitive, protected or confidential data has potentially been viewed, stolen or used by an individual unauthorized to do so. Information and data leakage is on the rise for multiple reasons, e.g., the poorly performing economy, frequent job changes, market advantage achieved by acquisition of trade secrets. This situation poses a serious problem to companies and organizations. The number of leakage incidents and the cost they inflict continues to increase. In most cases, the end product is not as valuable as obtaining the means of production, the research and development, or the know-how. Data loss can be caused by malicious intent or by unintentional mistake. Both cases can diminish a company's brand, reduce shareholder value, and damage the company's goodwill and reputation [15]. Data leakage prevention has been studied both in academic research areas and in practical application domains e.g., [1][16]. A number of methods and systems have been developed to prevent data leakage. For example, in [17], IRILD, an information retrieval based cyclical hashing approach for information leak detection is presented. Cyclical hashing is employed to split the document into multiple parts and generate fingerprints for these parts. This series of fingerprints are checked against the series of fingerprints of outgoing documents. In [18] the problem of giving sensitive data to a set of supposedly trusted third parties is discussed. Data allocation strategies were proposed, that improve the probability of identifying leakages caused third party agents. In [19] a framework is presented for detecting sensitive data exfiltration by an insider attack. However, data leakage detection systems cannot provide absolute protection. Thus, if we cannot

 ISBN: 978-1-61208-364-3

prevent data loss it is essential to discover data leakage as soon as possible.

III. DETECTING DATA LEAKAGE

The purpose of this research is to design and implement a cloud technology based data leakage detection system using different information retrieval models and methods. Cloud-implemented systems and services are available from anywhere and from any device. The only condition is that the device should have Internet access. Further benefits of clouds computing are: reduced IT costs, scalability, flexibility, etc. [23]. The research is expected to result in a system that is suitable for detecting sensitive information on the Web. The implemented data leakage detection system goes beyond the currently available services for comparing the contents of the documents. For example, plagiarism checking services examine and compare the documents word by word to find copy-paste content. These methods have the disadvantage that it can only take into account the words in the documents. In these methods, information about the semantics is not included.

The goal of the data leakage system is to monitor the Web and collect information according to users' preferences. Figure 1 shows the model of the system. The Web data sources are compared with user's confidential documents. Semantically meaningful similarity of data sets might indicate data leakage. Usually, the similarity of documents is determined using a repetition-based hard similarity metric S_H of any two words w_i and w_j:

$$S_H(w_i, w_j) = \begin{cases} 1, if\ w_i = w_j \\ 0, otherwise. \end{cases} \quad (1)$$

This approach ignores all potential semantic correlations between different words. In our system not the pure content, but the meaning of Web documents and user documents are compared. The system may consist of a number of modules. In this section, these modules are introduced briefly. The modules are represented in Figure 2.

The document collection contains sensitive, protected or confidential information. In order to protect these documents, encryption is required. Thus, the Cryptographic module resides on the clients' server. All the other services reside in the cloud. The Cryptographic module is responsible for preparing an encrypted version of the documents. In order to do it, an adequate mathematical model is necessary.

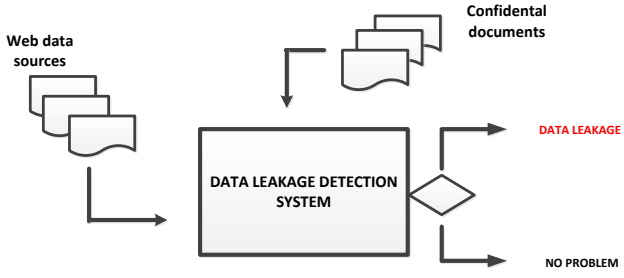


Figure 1. Data leakage detection system.

In information-retrieval the vector space model [20] is the basis of many systems. It is a simple and intuitively appealing framework for implementing term weighting and ranking. In this model, documents are assumed to be a part of an *n* dimensional vector space, where *n* is the number of index terms. In this model, every document is represented by a vector of index terms [6]. Applying the vector space model as the basis of the Cryptographic module the question arises: how to choose index terms. A previous idea is to let the user define index terms.

Text mining module will use the output of the Cryptographic module to define search Queries. Search queries are determined using the predefined vector space. Ontology may be used to add more semantics to search queries. Queries are submitted to the Search module. The Search module is responsible for discovering Web pages and collecting relevant data. It can be implemented as a conventional keyword-based metasearch engine.

Metasearch engines are search engines that search other engines. They submit the search query to several other search engines and return a summary of the results. This strategy gives the search a broader scope than searching with a single search engine [14]. The Search module incorporates a Crawler module that investigates the structure of Web sites, determines those pages of Web sites that contain relevant data, and indexes these pages using keywords. Text mining module converts Web documents into their vector space representations.

The scoring module matches the mathematical vector space representations of Web documents and user documents. Similarity is defined as a kind of distance. Based on this mathematical distance, the system can determine whether Web documents and user documents are close enough. If Web documents are close enough to confidential user documents data leakage warning signs appear. A number of mathematical models can be used to calculate similarity. In Section 4, two different approaches are proposed.

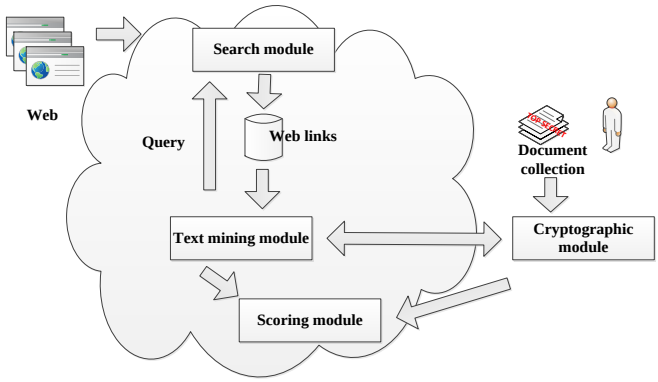


Figure 2. Data leakage detection modules.

 ISBN: 978-1-61208-364-3

IV. METHODS

In this section, methods that can be used in the scoring module are presented. Given a search query, the retrieved Web documents and confidential user documents, the scoring module computes a relevance score that measures the similarity between the Web documents and user document. Different methods will be implemented to satisfy different information needs or user preferences.

Web documents and user documents are represented using the vector space model. In this model, the vector space has to be defined first. On one hand, text mining methods may be used to identify index terms in confidential user documents. On the other hand, users may be asked to define index terms themselves. Considering the sensitive nature of user documents, the last solution may be better. The vector space representation of documents is as follows.

Given a finite set T of index terms $T = \{t_1, ..., t_i, ..., t_n\}$ defined by the user, any Web document W_j is assigned a vector $\mathbf{v}_j$ of finite real numbers, as follows:

$$\mathbf{v}_j = (w_{ij})_{i=1,...,n} = (w_{1j}, ..., w_{ij}, ..., w_{nj}) \quad (2)$$

The weight w_{ij} is interpreted as an extent to which the index term t_i characterizes a Web document. Confidential user documents also have to be represented as a vector. An appropriate safe approach can be to create an artificial document, i.e., the weights of index terms are determined by the user. As a result, a confidential user document is assigned a vector $\mathbf{v}_k$ of finite real numbers, as follows:

$$\mathbf{v}_k = (w_{ik})_{i=1,...,n} = (w_{1k}, \ldots, w_{ik}, \ldots, w_{nk}) \quad (3)$$

A Web document W_j is represented to a user having confidential document C_k if they are similar enough, i.e., a similarity measure S_{jk} between the Web document vector $\boldsymbol{v}_j$ and the confidential user document vector $\boldsymbol{v}_k$ is over some threshold K, i.e.,

$$S_{jk} = s(\mathbf{v}_j, \mathbf{v}_k) > K \quad (4)$$

In the classical vector space model (VSM) [3], different weighting schemes and similarity measures can be used, e.g., Cosine measure, Dice's coefficient [21]. However, the categoricity of the system can be varied by both changing the weighting scheme and similarity measure at the expense of a costly computation. To avoid costly computation, we propose the use of a VSM over the Cayley-Klein Hyperbolic Geometry [22].

In the classical VSM, the feature space is mathematically modelled by the orthonormal Euclidean space. In the hyperbolic information-retrieval model, the vector space is defined over the Cayley-Klein Hyperbolic Geometry. In hyperbolic IR (HIR), the similarity measure is derived from the hyperbolic distance. The hyperbolic similarity measure $S_{j,k}$ is defined as follows [4]:

$$S_{j,k} = \sigma_{j,k} = \left(ln\left(e \cdot \frac{r+\sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}}{r-\sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}} \right) \right)^{-1} \quad (5)$$

where

$$r > max_{\boldsymbol{v}_j} d_E(\boldsymbol{v_j v}_k) \quad (6)$$

and

$$d_E(\boldsymbol{v_j v}_k) = \sqrt{\sum_{i=1}^{n}(w_{ij}-w_{ik})^2}, \quad (7)$$

represents the Euclidean distance of the vectors. Given a VSM based on the Cosine measure using the term-frequency-normalized weighting, this scheme can be replaced with this hyperbolic IR model producing exactly the same answers and ranking. For technical disciplines, the usage of the term-frequency-normalized weighting scheme is recommended as yielding good results [2]. In the hyperbolic model, the categoricity of the system can be varied by only modifying the radius of the hyperbolic space and without using a different weighting scheme and similarity measure. Experiments demonstrated that categoricity in HIR can be varied more than O(n) faster, where n is the number of index terms, than in the VSM [4]. Thanks to the variable categoricity of the measure, the degree of similarity is easily variable in the system. This property can be utilized to vary similarity measure to satisfy different information needs or user preferences.

Besides the VSM, other techniques are considered to be used to determine the similarity of documents. Interaction information-retrieval (I^2R) model [7][8] may prove to be applicable too. Clustering is a well-known technique applied in IR. It is typically used to group documents to be searched. A special case of clustering is adaptive clustering, i.e., a clustering in which the cluster structure is being developed under or is being influenced by an interaction with the user. One way of conceiving adaptive clustering is to adopt a connectionist-based view.

In the data leakage detection system, adaptive clustering can be implemented as follows. Any Web document is represented by an object. An object o_i, i = 1, 2, ..., M, is assigned a set of identifiers. Identifiers are predefined index terms t_{ik}, k = 1, 2, ..., n_i. There are weighted and directed links between any pair (o_i, o_j), $i \neq j$, of objects.

One weight is the relative frequency [7][8] – denoted by w_{ijp} – of a term given a Web document, i.e., the ratio between the relevance r_{ijp} of index term t_{jp} in Web document object o_i, and the length n_i of o_i, i.e., the total number of index terms assigned to o_i:

$$w_{ijp} = \frac{r_{ijp}}{n_i}, p = 1, \ldots, n_j \quad (8)$$

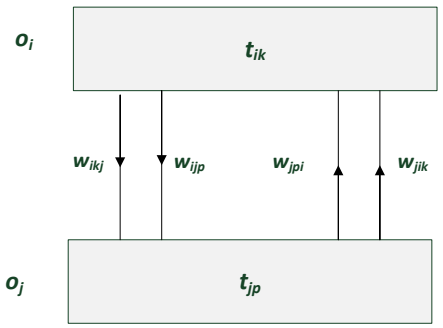


Figure 3. Connections between object pairs.

The other weight is the extent to which a given index term reflects the characteristic of a Web document, i.e., the inverse document frequency [7][8]. If r_{ikj} denotes the relevance of index term t_{ik} in o_j, and df_{ik} is the number of Web documents that can be indexed by t_{ik}, then w_{ikj} is given by the inverse document frequency formula, and thus, represents the extent to which t_{ik} reflects the characteristic of o_j:

$$w_{ikj} = r_{ijp} log \frac{2M}{df_{ik}}. \quad (9)$$

The other two connections - in the opposite direction - have the same meaning as above: w_{jik} corresponds to w_{ijp}, while w_{jpi} corresponds to w_{ikj}. Figure 3 shows the connections between the object pairs.

The Web documents are represented as an interconnected network; every document is linked to every other document. The confidential user document is conceived as being an object, too. It is interconnected with the already interconnected Web documents causing some of the already existing connections to change because of the change of M and df_{ik}. The objects are conceived as being a network of artificial neurons in which a spreading of activation takes place according to a winner takes all strategy. The activation is initiated at the confidential document, and spreads over along the strongest connection thus, passing on to another neuron, and so on. The strength of the connection between any pair (o_i, o_j), $i \neq$ i, of objects, and thus, between the confidential document and another Web document object o_i is defined as follows [7][8]:

$$K_{ij} = \sum_{p=1}^{n_j} w_{jpi} + \sum_{k=1}^{n_i} w_{jik} \quad (10)$$

After a number of steps, the spreading of activation reaches an object that was already affected. This is analogous to a local memory recalled by the confidential document. Those Web documents are retrieved by the system which belongs to this circle. These retrieved documents may indicate data leakage. The corresponding Web documents are ranked in the order of maximal activation. The advantages of the interaction model are as follows [9][10]. On one hand, this method also avoids costly computation. The complexity of weights computation is polynomial. The retrieval process takes polynomial time. On the other hand, the interaction retrieval method allows for a relatively high precision within 50%−70%. Standard test collections based evaluation showed that this method is useful when high precision is favored at low to middle recall values [9].

V. CONCLUSION

Data leakage prevention and protection might ensure that sensitive or confidential information remains safe and secure. Many software solutions were developed to provide data protection. However, malicious attacks and insiders' negligence cannot be completely eliminated. Thus, it is essential to discover data leakage as soon as possible.

In this paper we introduced a semantic information-retrieval based approach to address the problem of data leakage detection. The idea of the system is to monitor the Web, collect information according to users' preferences and indicate data leakage based on semantic similarity of documents. A modular system architecture that composed of Web searching, text mining and scoring was proposed. A connectionist and a hyperbolic IR model were suggested to be implemented in the scoring module, because these models avoid costly computation. The system is under implementation. After implementing the system, experiments have to be carried out to evaluate its effectiveness and precision.

Our future work includes the investigation of automatic summarization methods that can be implemented in the Cryptographic module to extract keywords from sensitive documents. Another open problem is the extension of the Cryptographic module with query expansion techniques so that module can handle semantically related forms of keywords.

ACKNOWLEDGMENT

This research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004 - National Research Center for Development and Market Introduction of Advanced Information and Communication Technologies.

REFERENCES

[1] A. Shabtai, Y. E. Asaf, and R. Lior, A survey of data leakage detection and prevention solutions. Springer, 2012, ISBN: 978-1-4614-2052-1.

[2] C. T. Meadow, Text Information Retrieval Systems. Academic Press, 2000, ISBN: 0124874053.

[3] G. Salton, "Automatic phrase matching," In: Hayes, DG, Ed., Readings in Automatic Language Processing. American Elsevier Publishing Company, Inc., New York, 1966, pp. 169-188.

[4] J. Góth, and A. Skrop, "Varying retrieval categoricity using hyperbolic geometry," Information Retrieval, vol. 8(2), 2005, pp. 265-283.

[5] M. E. Kabay. *Glossary of Computer Crime Terms*. [Online]. Available from: http://www.mekabay.com/overviews/glossary [retrieved: May, 2014]

[6] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval: The Concepts and Technology behind Search (2nd Edition). ACM Press Books, Addison-Wesley Professional, 2011, ISBN: 0321416910.

[7] S. Dominich, "Connectionist interaction information retrieval," Information processing & management, vol. 39.2, 2003, pp. 167-193, doi: 10.1016/S0306-4573(02)00046-8.

[8] S. Dominich, "Interaction information retrieval," Journal of Documentation, vol. 50.3, 1994, pp. 197-212, doi: 10.1108/eb026930.

[9] S. Dominich, "Connectionist interaction information retrieval," Information Processing & Management, vol. 39, 2003, pp.167-193, doi.: 10.1016/S0306-4573(02)00046-8.

[10] S. Dominich, A. Skrop, and Zs. Tuza, "Formal Theory of Connectionist Web Retrieval," Soft Computing in Web Information Retrieval, Studies in Fuzziness and Soft Computing, vol. 197, 2006, pp. 163-194.

[11] TechTarget. *WhatIs.com. Definition: data exfiltration.* [Online]. Available from: http://whatis.techtarget.com/ [retrieved: May, 2014]

[12] The Computer Language Company Inc. *Encyclopedia.Definition of: data breach.* [Online]. Available from: http://www.pcmag.com/encyclopedia/term/61571/data-breach [retrieved: May, 2014]

[13] The MITRE Corporation. *Common Weakness Enumeration (CWE) is a list of software weaknesses. CWE-200: Information Exposure.* [Online]. Available from: http://cwe.mitre.org/data/definitions/200.html [retrieved: May, 2014].

[14] W. B. Croft, D.Metzler, and T. Strohman, Search engines: Information retrieval in practice (p. 283). Reading: Addison-Wesley, 2010.

[15] CISCO. *Cisco Data Loss Prevention.* [Online]. Available ffrom: http://www.cisco.com/c/en/us/products/security/email-security-appliance/dlp_overview [retrieved: May, 2014]

[16] Symantec. *Phishing – The latest tactics and potential business impacts.* White paper. Oct 11, 2012, [Online]. Available from: http://whitepapers.itnews.com.au/content22479

[17] E. Gessiou, Q. H. Vu, and S. Ioannidis, "IRILD: an Information Retrieval based method for Information Leak Detection," In Proceedings of European Conference on Computer Network Defense, 2011, pp. 33–40, IEEE.

[18] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," Knowledge and Data Engineering, IEEE Transactions on, vol. 23(1), 2011, pp. 51-63.

[19] Y. Liu, C. Corbett, K. Chiang, R.Archibald, B.,Mukherjee, and D. Ghosal, "SIDD: A framework for detecting sensitive data exfiltration by an insider attack," In System Sciences, 2009, HICSS'09, pp. 1-10, IEEE.

[20] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18(11), 1975, pp. 613-620.

[21] W. B. Frakes, and R. Baeza-Yates, Information retrieval, Data Structures and Algorithms. Prentice Hall, 1992, ISBN: 0-13-463837-9.

[22] J. Bolyai, APPENDIX: The Theory of Space. Akadémiai Kiadó, Hungary, Budapest, 1987, ISBN: 9630515121.

[23] T. Velte, A. Velte, and R. Elsenpeter, Cloud Computing, A Practical approach. McGraw Hill Professional, 2009, ISBN: 0071626956.

[24] InvestopediaUS. *Industrial Espionage.* [Online]. Available from: http://www.investopedia.com/terms/i/industrial-espionage.asp [retrieved: May, 2014].

[25] D. S. Wall, "Organizational security and the insider threat: Malicious, negligent and well-meaning insiders," Technical report, Symantec, 2011.

[26] CDW. *Data loss prevention.* Online]. Available from: http://www.cdw.com/content/solutions/data-loss-prevention.aspx [retrieved: May, 2014].

 ISBN: 978-1-61208-364-3

Document Retrieval in Big Data

Feifei Pan
Computer Science Department
New York Institute of Technology
New York, USA
Email: fpan@nyit.edu

***Abstract*—Nearest Neighbor Search for similar document retrieval suffers from an efficiency problem when scaled to a large dataset. In this paper, we introduce an unsupervised approach based on Locality Sensitive Hashing to alleviate its search complexity problem. The advantage of our proposed approach is that it does not need to scan all the documents for retrieving top-K Nearest Neighbors, instead, a number of hash table lookup operations are conducted to retrieve the top-K candidates. Experiments on two massive news and tweets datasets demonstrate that our approach is able to achieve over an order of speedup compared with the traditional Information Retrieval method and maintain reasonable precision.**

Keywords–Document Retrieval; Locality Sensitive Hashing; Big Data.

I. INTRODUCTION

The Nearest Neighbor Search (NNS) task [1] aims at automatically finding the top K objects (e.g., documents) which are most semantically similar to a given query object. NNS is essential to motivate the progress in many search related tasks and is fundamental to a broad range of Natural Language Processing (NLP) down-stream tasks, including name spelling correction [2], document translation pair acquisition [3], large-scale similar noun list generation [4], unsupervised mining of lexical variants from noisy texts [5], and large-scale first story detection from news and tweets [6].

Nowadays, data are being collected at unprecedented speed and scale everywhere around us: various news agencies produce thousands of news articles while Twitter generates over 500 million Tweets everyday. The traditional Information Retrieval (IR) method to tackle NNS is to represent documents in the vector space and find the candidate documents that share the highest probabilities with the query document [7]. However, it is very time consuming or even infeasiable to brute force compute the similarity score of the query document and all other documents in a large dataset. Thus, it is critical to find other solutions to deal with the search efficiency problem. In order to make it scale to big data, some researchers attempted to reduce the dimentionality of the representative vectors or add in time constraints to narrow down the search range [8]. Both of these works are able to save some computational costs, but can not solve the problem fundamentally.

Hashing has been successfully applied to several non-NLP problems including object recognition [9][10], image retrieval [11][12] and image matching [13][14]; however, it received limited attention in NLP fields. The general idea of hashing is to represent each document as a binary code (1-bit of a binary code is one digit of "0" or "1"). Its advantage is two-fold: (1). The capability to store large amount of documents in memory. For example, we can store 250 million documents with 16G memory using 64 bit for each document, while English Gigaword fifth edition [15] stores 10 million documents with 26G. (2). The time efficiency to process binary codes. For example, calculation of hamming distance between a pair of binary codes is far faster than cosine similarity over a pair of document vectors.

The paper is structured as follows: in Section I-A and II, we briefly introduce the terminologies and previous related work. In Section III, we compare our Locality Sensitive Hashing (LSH) based approach with the traditional IR method in tackling NNS task. We talk about the experiment results in Section IV and make the conclusion in Section V.

A. Background and Terminology

Here are some background information and terminologies: Bit is the basic unit of a binary code; one bit is either a digit of "0" or "1". A binary code is a bit sequence assigned to represent an object. For example, we can represent a document as "00101100". A hash table is a table containing all the binary codes for a set of documents while the documents with the same binary codes are located within the same bucket. Hamming Distance between two binary codes is the number of positions at which the corresponding bits are different. Hash Lookup is to find candidate neighbors in the hash table buckets within a given hamming distances from h_q, given a query q with a binary code h_q. In practice, the given hamming distance is usually set to 2. Hash Lookup Success rate is the probability to find any candidate neighbors in the buckets within a given hamming distances from h_q.

II. RELATED WORK

Locality Sensitive Hashing (LSH) [16] is one of the notable schemes for data-independent hashing. It uses random projections to construct randomized hash functions, therefore similar data points have a higher probability to be mapped into the same bucket. To address the problem of learning similarity-preserving binary code for efficient retrieval from a large scale collection, several data dependent hash schemes were developed. Weiss et al. [17] designed Spectral Hashing (SH) which was motivated by spectral graph partitioning and it used a spectral relaxation to obtain an eigenvector solution. Liu et al. [18] utilized anchor graphs to discover the neighborhood structure inherent in the data, and Gong and Lazebnik [19] proposed an Iterative Quantization (ITQ) approach by minimizing the quantization errors. Generally speaking, data dependent hash schemes are able to learn better quality binary codes than randomized algorithms, so we adopt ITQ to hash documents to binary codes. However, in most of the data dependent hash schemes, the generated binary codes suffer from poor hash table lookup success rate problem which

 ISBN: 978-1-61208-364-3

makes the learnt binary codes inefficient for practical use. In this paper, we aim to alleviate the search efficiency problem by taking the advantages of LSH.

III. Nearest Neighbor Search

In this section, we first show the motivation of adopting hashing techniques to NNS problem in Section III-A. Then, we introduce the details of LSH in Section III-B.

A. Traditional NNS

The most traditional way of finding Nearest Neighbors is to represent each document as a term vector, e.g., each element of the vector is the tf-idf weight of a term:

$$tf(t,d) = log(1 + f(t,d)) \quad (1)$$

$$idf(t,D) = \frac{log(|D|)}{log(|\{d \in D : t \in d\}|)} \quad (2)$$

$$tf\text{-}idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

where t is a term, d is a document, D is the whole corpus and f(t,d) is the frequency of term t in document d. Given a query document q, we used the cosine similarity metric [20] to judge the similarity of q with a document candidate c:

$$distance(q,c) = cos(\theta) = \frac{q \cdot c}{||q||||c||} = \frac{\sum_{i=1}^{n} q_i \times c_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2}\sqrt{\sum_{i=1}^{n}(c_i)^2}} \quad (4)$$

The higher similarity score between q and c, the closer they are. To find out the top-K nearest neighbors for a query document q, one needs to first compute the cosine similarity scores between q and each document candidate, then pickup the K documents with the highest similarity scores. However, brute force search does not scale to big data since the computational complexity for each query is $O(n)$ and other computational costs such as similarity calculation and ranking similarity scores are not immaterial.

Hashing schemes aim to remove the curse of dimensionality and largely save computation cost. We will introduce LSH in the following section.

B. Locality Sensitive Hashing

The underlying intuition of LSH is that if two objects are close, then after a "projection" operation they will remain close together. In other words, similar data points are more likely to be mapped into the same bucket with a high collision probability.

Binary Code Learning

Given a LSH setting of M bits and L hash tables, a query data point q and a candidate data point c will collide if and only if:

$$h_{ij}(q) = h_{ij}(c), i \in [1...L], j \in [1...k] \quad (5)$$

and the hash function $h_{ij}(x)$ is defined as:

$$h_{ij}(x) = sgn(u_{ij}^T \cdot x), sgn(u) = \begin{cases} 1 & \text{u} \geq 0 \\ 0 & \text{Otherwise} \end{cases} \quad (6)$$

where u_{ij} are randomly generated vectors with components randomly selected from a Gaussian Distribution, e.g., $N(0,1)$. In this case, the probability of two points x_1 and x_2 colliding under LSH can be calculated as:

$$P_{coll} = 1 - \frac{\theta(x_1,x_2)}{\pi} \quad (7)$$

where $\theta(x_1,x_2)$ is the angle between x_1 and x_2. Given the desired probability of missing a nearest neighbor δ, we can approximate the required number of hash table L:

$$L = log_{1-P_{coll}^M}\delta \quad (8)$$

In this paper, given the term vector as input, we learn the binary code for each document and we try different settings for M and L to see how they influence the results.

Document Retrieval

After learning the binary codes for all documents, inverse table lookup operations are conducted to find nearest neighbors of a query document given the binary code of the query document h_q:

1) lookup the bucket that has the same binary code as h_q and retreive all the documents within the bucket. 1 hash table lookup operation is needed.
2) lookup the buckets that have Hamming Distance 1 with h_q and retreive all the documents within the buckets. $length(h_q)$ hash table lookup operation is needed.
3) lookup the buckets that have Hamming Distance 2 with h_q and retreive all the documents within the buckets. $length(h_q) \times (length(h_q) - 1)$ hash table lookup operation is needed.
4) Document candidates in the same buckets will be randomly pickup to form the top-K nearest neighbors.

We only conduct hash table lookup for the buckets which have Hamming Distance smaller than or equals to 2 with h_q. Otherwise, it requires too many hash table operations and it is not efficient for a large dataset.

IV. Experiments

A. Data

For the experiments, we use a news dataset and a tweets dataset, in order to see how the proposed methods perform for different genres. The news dataset is the English portion of the standard TDT-5 [21] dataset. It consists of $278,109$ documents from the 6-month time period since April 2003. 126 topics with an average of 51 documents per topic are annotated, and other unlabeled documents are irrelevant to them. 400 randomly selected labeled documents are used for testing. The tweets dataset is gathered through Twitter Application Programming Interface (API) from a time period between October 25th, 2012 and November 4th, 2012, filtered by hashtags '#hurricane' and '#sandy'. As a result, we collected 1.49 million tweets before, during and after Hurricane Sandy hit the northeast of the US.

 ISBN: 978-1-61208-364-3

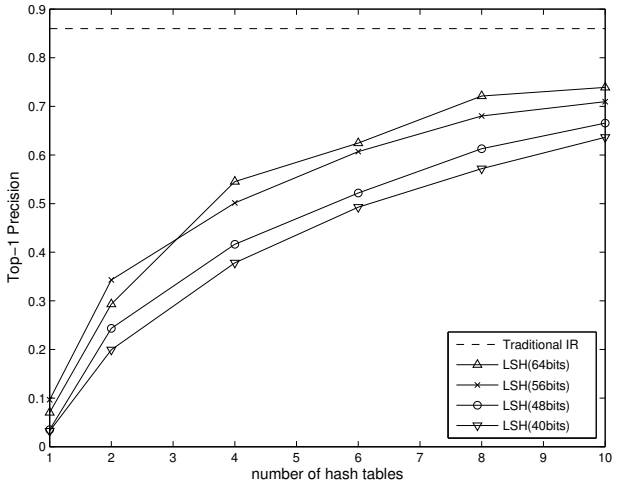


Figure 1: LSH Top-1 NNS Precision.

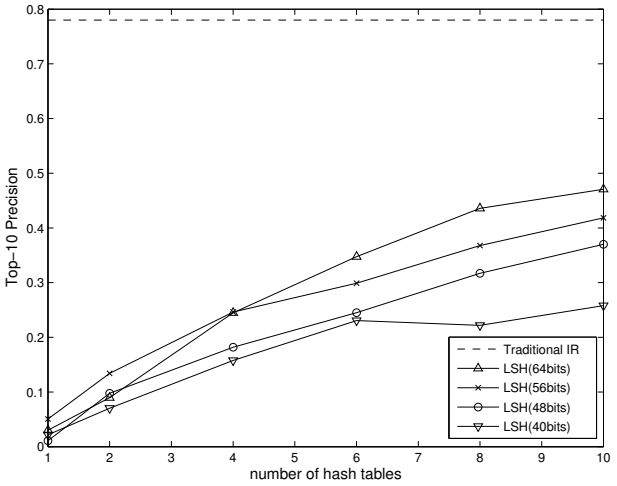


Figure 2: LSH Top-10 NNS Precision.

TABLE I: SAMPLE QUERY TWEETS AND THE CORRESPONDING TOP RETRIEVED SIMILAR TWEETS.

Query 1: @danburyweather: Mayor Bloomberg Tells NYC Residents: 'Be Prepared To Evacuate' Read more: http://t.co/oe2IUB2E
Mayor Bloomberg Tells NYC Residents: 'Be Prepared To Evacuate' http://t.co/eIXJBzmO
#Hurricane Sandy @businessinsider: Mayor Bloomberg Tells NYC Residents: 'Be Prepared To Evacuate' by @DinaSpector http://t.co/l76jFAAH
Watch out NYC
I can't wait to hear Mayor Bloomberg say Zone Ah again in Spanish when talking about hurricane evacuation procedures. #Frankenstorm
For NYC residents - hurricane evacuation zone finder tool - http://t.co/mB1WSDPf
@DanSkeldonNBC40 Which storm was Bering hyped more Sandy or Irene? Do you think Cape May County residents need to evacuate for Sandy?
RT @EasternSurfMag: STAY ALERT and BE PREPARED this weekend
RT @Chels_sahagian3: Just watching the news about this hurricane is making me more and more scared:(
NYC Hurricane evacuation map
Watch there be a hurricane on Monday and Bloomberg stills makes us go to school. Jew bastard
Query 2: People return home from shelters after hurricane Sandy: More than 1800 people who were housed in shelters prior to the hurricane
People return home from shelters after hurricane Sandy http://t.co/36cSHU2r
Jamaica: More than 1800 people return home from sutlers after hurricane Sandy
At the height of #Sandy more than 1800 people were housed in Red Cross shelters
258 shelters in 16 states safely housed 11
@JeffreyYoung_HC are hospitals exempted from evacuation NYU endangered 300 lives for not evacuating prior to the storm. http://t.co/QvqQmwCG
@Tek_Roo FEMA organized with states prior to Sandys landing to get people evacuated and resources in place for rescue.
RT @nycgov: RI @NYCMayorsOffice: Mayor: we will keep shelters open until New Yorkers can safely return to their homes. #Sandy
West Deptford shelter housed 40 people during Hurricane Sandy
Thanks toRedCross258 shelters in 16 states safely hosed 11000 people. #Sandy Recovery begins today Every dollar helps http://t.co/ghhORZvz
Thousands in New York remain homeless and in shelters nearly a week after Hurricane Sandy. It seems like things are returning to normal.

We select 20 informative tweets as testing queries. For each tweet, we remove hashtags, URLs and @ information. For each news article and tweet, we apply the Stanford Tokenizer [22] for tokenization, remove stopwords based on the stop list from InQuery [23], and apply Porter Stemmer [24] for stemming.

To evaluate the system performance for new articles, we use the topic labels of documents as ground truth: if one retrieved document shares the same topic label with the query document, they are true neighbors. We evaluate the precision of the top-K candidate documents returned by each method and calculate the average precision across all queries. Since we do not have groundtruth for tweets, we will not report system performance for tweets dataset in the paper. Instead, we show system outputs given some tweet queries to demonstrate the effectiveness of our proposed method.

B. Results

Our LSH-based approach aims to alleviate the search efficiency problem of NNS. We compare the average search time of the traditional IR method and our LSH-based method. In the news dataset, LSH-based method only needs about one twentieth of the search time as the traditional IR while in tweets dataset the differences becomes even larger: LSH-based method only needs one twentieth of the time approximately. It clearly proves the superiority of our approach.

Furthermore, we compare the top-K NNS precision of the traditional IR method and our LSH-based method. Fig. 1 and Fig. 2 are the top-K NNS results for the news dataset. Generally speaking, with longer binary code length or more number of hash tables, the top-K NNS precision keeps increasing and it reaches convergence approximately in the setting of 64 bits and 10 hash tables. In Fig. 1, it is clear that when retrieving top-1 Nearest Neighbors, although LSH can not well approximate the performance of traditional IR, it is still able to produce an acceptable performance. However, in Fig. 2, LSH is surpassed by traditional IR by a large margin when retrieving top-10 Nearest Neighbors. It suggests that our approach is more reliable when K is small. Table I shows two sample query

 ISBN: 978-1-61208-364-3

tweets and the correspoding similar tweets returned by the LSH-based approach with 64 bits and 10 hash tables setting. The returned tweets are the most tropically related results to the query tweets and it demonstrates that our approach can be adapted to tweets as well.

V. Conclusion and Future Work

In this paper, we propose an efficient LSH based solution for document retrieval in big data. Although our approach is unable to achieve similar performance as the traditional IR method, it boosts the search time over an order of magnitude. Experiments on two genres show that our approach is flexible and feasible in practical use, especially for retrieving a small number of documents in a large dataset. In the future, we plan to conduct manual annotation on the tweets dataset in order to carry out quantitative evaluations. After that, we will focus on improving the document representation to further boost the precision of the LSH based approach.

References

[1] A. Andoni, "Nearest neighbor search: the old, the new, and the impossible," in PhD Dissertation in MIT, 2009.

[2] R. Udupa and S. Kumar, "Hashing-based approaches to spelling correction of personal names," in EMNLP, 2010, pp. 1256–1265.

[3] K. Krstovski and D. A. Smith, "A minimally supervised approach for detecting and ranking document translation pairs," in The sixth ACL Workshop on Statistical Machine Translation, 2011, pp. 207–216.

[4] D. Ravichandran, P. Pantel, and E. H. Hovy, "Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering," in ACL, 2005, pp. 622–629.

[5] S. Gouws, D. Hovy, and D. Metzler, "Unsupervised mining of lexical variants from noisy text," in EMNLP, 2011, pp. 82–90.

[6] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in HLT-NAACL, 2010, pp. 181–189.

[7] J. Allan, V. Lavrenko, D. Malin, and R. Swan, "Detections, bounds, and timelines: Umass and tdt-3," in In Proceedings of Topic Detection and Tracking Workshop, 2000, pp. 167–174.

[8] S. Petrovic, "Real-time event detection in massive streams," in PhD Theis at University of Edinburgh, 2012.

[9] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," in IEEE Trans. Pattern Anal. Mach. Intell., vol. 30(11), 2008, pp. 1958–1970.

[10] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in CVPR, 2008, pp. 1–8.

[11] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, 2009, pp. 2143–2157.

[12] H. Xu, J. Wang, Z. Li, G. Zeng, S. Li, and N. Yu, "Complementary hashing for approximate nearest neighbor search," in ICCV, 2011, pp. 1631–1638.

[13] S. Korman and S. Avidan, "Coherency sensitive hashing," in ICCV, 2011, pp. 1607–1614.

[14] C. Strecha, A. A. Bronstein, M. M. Bronstein, and P. Fua, "Ldahash: Improved matching with smaller descriptors." IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, 2012, pp. 66–78.

[15] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda, "English gigaword fifth edition (http://catalog.ldc.upenn.edu/ldc2011t07)," 2011.

[16] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in STOC, 1998, pp. 604–613.

[17] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in NIPS, 2008, pp. 1753–1760.

[18] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in ICML, 2011, pp. 1–8.

[19] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in CVPR, 2011, pp. 817–824.

[20] A. Singhal, "Modern information retrieval: A brief overview," in Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24(4), 2001.

[21] K. M. S. S. David Graff, Junbo Kong, "Tdt5 multilingual text (https://catalog.ldc.upenn.edu/ldc2006t18)," 2006.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[23] J. Callan, W. Croft, and S. Harding, "The inquery retrieval system," in Proceedings of the Third International Conference on Database and Expert Systems Applications, 1992, pp. 78–83.

[24] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14(3), 1980, pp. 130–137.

 ISBN: 978-1-61208-364-3

Bag-of-Features Tagging Approach for a Better Recommendation with Social Big Data

Ming Cheung

HKUST-NIE Social Media Lab,
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology,
Hong Kong
Email: cpming@ust.hk

James She

HKUST-NIE Social Media Lab,
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology,
Hong Kong
Email: eejames@ust.hk

***Abstract*—The interests of users are always important for personalized content recommendations on friendships, events and media content from the social big data. However, those interests may not be specified, which makes the recommendations challenging. One of the possible solutions is to analyze the user's interests from the shared content, especially images with manually annotated tags. They are shared on online social networks such as Flickr and Instagram. However, the accuracy of the recommendation is greatly affected by the accuracy of the tag, which is not always reliable. This paper demonstrates how a bag-of-features (BoF)-based tagging approach can help to improve the accuracy of recommendations using an unsupervised algorithm. A set of auxiliary tags is used to represent user interests and, hence, the recommendation. The approach is evaluated with over 500 user and 200k images from Flickr. It is proven that by BoF tagging (BoFT), friendship recommendation is possible without friendship/tag information and the recall and the precision rate are improved by about 50% over using user tags.**

Keywords–Image tagging; recommendation; online social network; bag-of-features; annotation; big data.

I. INTRODUCTION

Nowadays, sharing social content has become part of our lives, in which billions pieces of content are shared. Recommending content that matches the user's interests from the social big data is important for any social networks. However, the user's interests may be hidden, that is the interests are not specified in the user profile. With an incomplete set of data, the content recommendation may be inaccurate. On the other hand, the user's interests are reflected in the abundance of social content, especially image, shared on the networks. A good recommendation is possible through analyzing the users' interests reflected among shared images. One of the most important applications is friendship recommendation, the inference of the connection between two users [1]. One of the possible ways to analyze a user's interests from shared images is through tagging [2]. Tagging, the act of using text to annotate a social content, is one of the most basic and essential features in any social networks that helps content recommendation [3]. The tags describe the image and reflect the users' interests since users with similar interests are more likely to upload images with similar tags. Connections can be discovered with the tendency to make friends with someone who shares similar interests reflected in the tags. Fig. 1(a) to (d) is a set of images and their tags by different users in Fotolog, Flickr, Twitter and Instagram. The tags include the name of the object, location, time or even the feeling felt at the time by the user. They are the major types of tags [4].

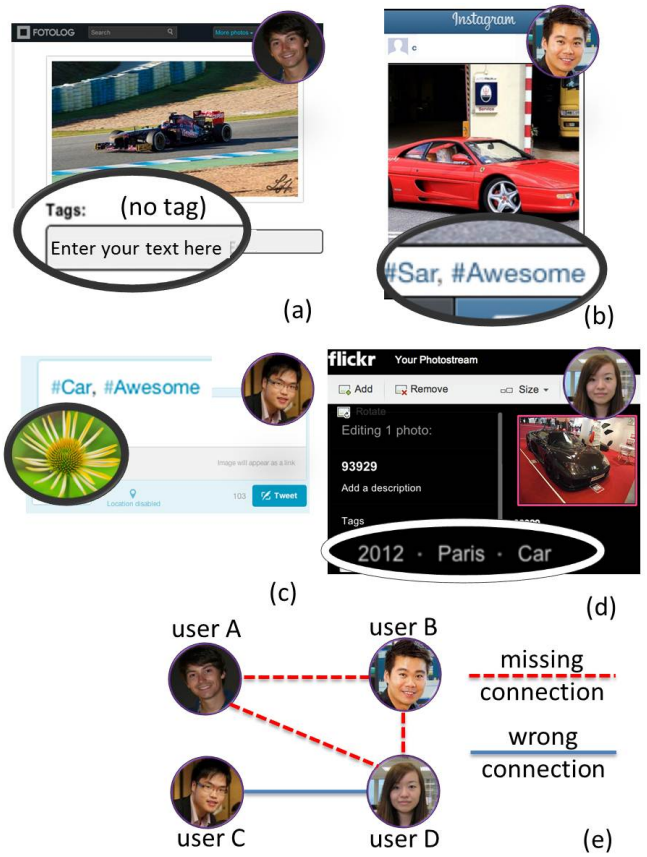


Figure 1: Examples of tagging on: (a) Fotolog, (b) Instagram, (c) Twitter, (d) Flickr, (e) and the corresponding social graph.

A tie between 2 people indicates that there is a connection between them such as friendship. The strength of connections among users can be measured by tie strength. People with higher tie strength such as best friends have a higher influence on the user. One of the important elements in measuring the tie strength is the common interests they share [5]. As tags reflect the interests of users, the tie strength can be estimated by the similarity of tags and hence, calculate possible connections. For example, if two users both upload images with the tag, "Car", the users are similar in terms of their interest and can be connected. Reliable tags that can reflect the nature of an image

 ISBN: 978-1-61208-364-3

are essential for this calculation. However, in most of the social networks, the tags are added manually and are not always reliable [6]. They may not be accurate, suitable for analysis, or, sometimes even available. Some users are not interested in annotating the images they upload because of the considerably longer time required than simply uploading the images, as the user shown in Fig. 1(a). For those tagged images, the tags may not be a good description of the content. Users may type the tag wrongly such as the tag, "Sar" instead of "Car", in Fig. 1(b). Some irrelevant tags are added intentionally to increase the popularity. An example is the tag "Car" for the picture of flowers of Fig. 1(c). For a good annotated image, analyzing the tags is still not an easy task. The tags may have different levels of details [8] or diverse details. For example, in Fig. 1(d), the user annotates the object, time and place. As a result, the social graph by the tags, as shown in Fig. 1(e) has wrongly connected user C with user D while leaving users A and B without any connection. These are some common examples of how a user annotates an image on social networks and how that annotation affects the discovery of connections.

This paper proposes a novel approach using BoF-based tagging that makes better recommendations through users' interests discovered from images uploaded. Instead of using a supervised approach in [9], this paper proposes an unsupervised approach in which images are grouped visually by BoF. The approach is also evaluated using a dataset of 542 users and 201006 images and the actual relationship among users. The results prove that the proposed approach can help to make a better recommendation. The main contributions are the following: 1) propose a novel way to represent user interest with auxiliary tags in an unsupervised manner; 2) introduce a friendship recommendation approach based on the auxiliary tags; and 3) verify the recommendation with the actual relationship from the scraped data. Section II in this paper discusses previous works. Section III is the general context of a BoF-based tagging, followed by how to connect people with similarity in Section IV. Section V shows the details of the experimental result and Section VI concludes the paper.

II. Previous Works

Recommending personalized content from billions of shared content is always a challenging task. Information overload may occur so that users have difficulty processing the huge amount of available content. A possible solution is a recommendation system based on a trust-based approach [10], where the user's social connection is considered for filtering the content. The interest shared is a way to measure tie, the strength of the social connection [7]. A hybrid approach can combine interests and the social connections. However, obtaining the user's interest is not always available. Although users can enter their interests for better recommendations, they may not want to spend time on the annotation. One of the possible solutions is to analyze the interests reflected in the content they have shared, especially images. This analysis can be based on the user annotated tags that describe the images. However, those tags may be inappropriate, wrong or have a different degree of details. One of the most well-known solutions is Collaborative Filtering (CF) [11][12], in which the same social content will be tagged by many users. The final tag quality can be improved [13] by analyzing the tags from different users. Although there is a promising result by applying CF, it is not suitable for systems with a large amount of images. The reason behind is that only small portion of images are popular and receive many tags. While it gives some of the images appropriate tags, most of them are left without proper tags for analysis. In image sharing platforms, such as Flickr, a user can upload hundreds of images at a time which makes CF inappropriate.

Another possible way is a content-based approach, in which the visual features are considered in order to annotate an image [14][15]. However, determining the relationship between the features and the tags is not a trivial task. The same object can be visually different among images. In this paper, BoF-based tagging [16] is applied. The proposed approach makes use of computer vision techniques in object recognition tasks to infer interests for friendship recommendations. BoF is an image-based approach that detects low-level features, and encodes an image into a feature vector. An unsupervised method is used for the learning. Images are grouped based on the similarity of their features vectors and hence the similarity of 2 users can be calculated. With this approach, it is possible to obtain the tag given to an image and, therefore, recommendation.

Among different types of content recommendation, friendships, or connections among people, is one the most important and fundamental functions. This problem has long been studied. One of the possible ways to make the recommendation is by the existing connections among people [17][18]. However, this may limit the recommendations from millions of users as the connections among users may not available. Friendship recommendation is also possible with user interests [18] inferred from user input [19] or user generated content [20] and other personal information [1]. Interests are combined with the existing connections with a machine learning algorithm in [21] for recommendation. In [22], the authors focus on how to make use of the group information on Flickr for friendship recommendation. The co-occurrence in images can also be a cue on friendship recommendation [23].

III. BoF-based Tagging

BoF has been a popular approach to many computer vision tasks because of its simplicity [16]. BoF is a method to represent images into feature vectors of local image descriptors. Fig. 2 is the process of the proposed approach in which Fig. 2(a) is the use of BoF in this work. The different parts of the BoF tagging are introduced in this section.

A. Feature Extraction

Feature extraction is a process to obtain the local features in step 1 of Fig. 2. These features can be detected by Harris Affine detector, or Maximally Stable Extremal Regions detector [16]. The extracted features are relatively consistent with viewing angles and lighting conditions. They are represented in a way that is independent of the size and orientation, such as scale-invariant feature transform (SIFT) [24].

B. Codebook Generation

Codebook generation (step 2 of Fig. 2) is a process to obtain the visual words that can represent the features obtained in the feature extraction in step 1 of Fig. 2. It is a clustering process that groups similar features. The mean vectors of each group are defined as the visual word, which can be used to represent

 ISBN: 978-1-61208-364-3

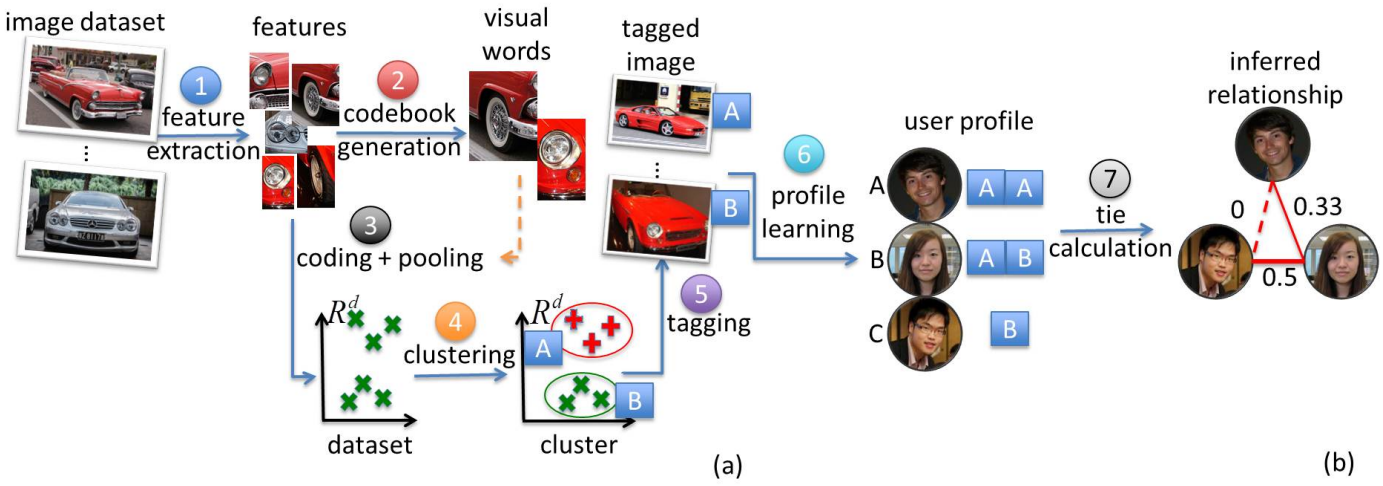


Figure 2: Flow chart of the proposed approach: (a) BoF-based tagging, (b) relationship recommendation.

the features in these images. One of the possible techniques to obtain those clusters is by k-mean clustering [25], which groups the visual features based on their visual similarity. The codebook generation is an offline process that does not need to be updated in real time.

C. Feature Coding and Pooling

Feature coding is to encode features with the visual words. Each feature in every image is represented by a visual word in feature coding. The images are then represented by a feature vector in the feature pooling. This process is carried out in encoding the images in the dataset (step 3 of Fig. 2). One of the most common approaches is using the histogram that counts the number of occurrences of each visual word in the image. The feature vector obtained is used in the clustering to group images that are visually similar.

D. Clustering and Tagging

The goal of clustering is to group images with similar feature vectors, that is, group images that are visually similar. Each cluster obtained in this operation corresponds to similar objects to which an auxiliary tag is assigned. After obtaining the cluster in step 4 of Fig. 2, the images in any cluster are assigned with the same auxiliary tag to reflect that they are visually similar and belong to the same group. It is an unsupervised operation no assumption is made or information on the image is known.

IV. Proposed BoF-Based Recommendation

This section introduces how to find similarities among people from the result of the BoF tagging (BoFT). The first part introduces how to learn the user profile from the result of BoF tagging, while the second part discusses how to make recommendations based on the user profile.

A. BoF Tagging and User Profile

The user profile, which reflects the interests of the users is the key in the content recommendation. A user profile can be obtained based on user manual input, in which the user manually inputs what kind of content is their favor. In the proposed approach, it is assumed that no user input is needed and the user profile is obtained from the image uploaded as in step 5 of Fig. 2. The histogram of the tags is used as the user profile in the proposed approach.

B. User Profile and Recommendation

When the user profile is obtained, the next step is to make a recommendation to the user. Recommendations are based on the tie, the strength of the relationship between two people. An item favored by a user may also be liked by friends of the user with strong ties. For example, user A likes Ferrari, while user B likes BMW and user A is user C's best friend. As a result, it is more likely that user C likes Ferrari. Content recommendation is then possible with the value of the tie calculated by user profile with the following formula:

$$S_{cosine}(A,B) = \frac{T_A \cdot T_B}{||T_A|| \cdot ||T_B||} \quad (1)$$

where T_A is the set of tags in the images uploaded by user A and T_B is the set of tags in the images uploaded by user B. The pairwise similarity is calculated based on the user profile. It is possible to obtain the tie between two people and find that people with similar interests have a higher similarity. The tie is assumed to be undirected, which means that the tie is the same from user A to B and user B to A. Different types of recommendation is possible with user ties, in particular, the focus of this paper is on friendship recommendation.

V. Experimental Results

In this section, the dataset, the experiments and the results are discussed. In this paper, the discussion focuses on discovering the connections among users by the tendency of people to make friends with people who share similar interests. The results show that it is possible to infer connections by using the BoF tagging.

A. Dataset and Experimental Setup

The setting of the experiment is shown in Fig. 3. A set of 201006 images uploaded by 542 users is scraped from Flickr,

 ISBN: 978-1-61208-364-3

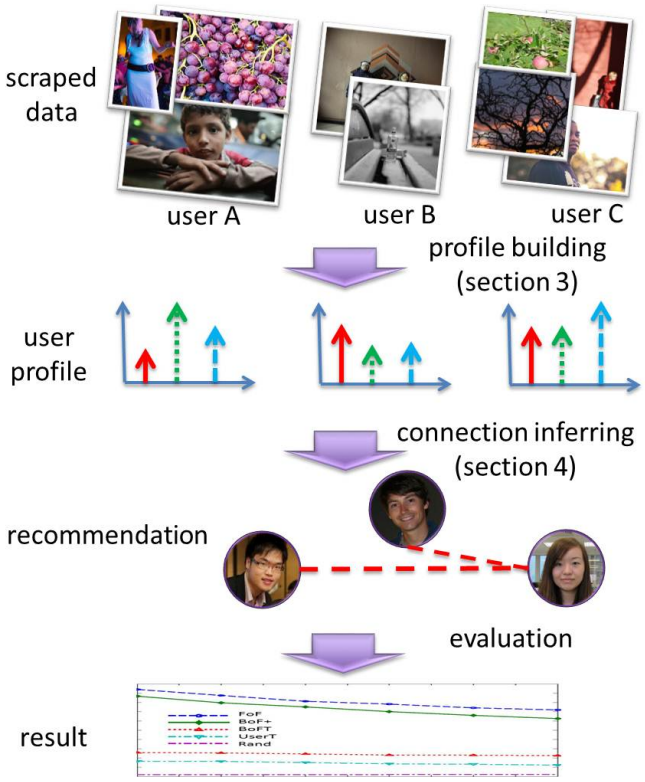


Figure 3: Experiment setting

an online social network for image sharing with millions of images uploaded and tagged, and the BoF tagging approach is used to tag those images. The 542 users are selected randomly from images under the same tag query page to provide diversity. The user profiles from the uploaded images are built with the tags obtained. Then connections among users are inferred with the tie calculation and evaluated with the actual connections scraped with the images. Tables I and II show the attributes scraped for the users and the images.

TABLE I: MAJOR ATTRIBUTES FOR IMAGES

Attribute	Description
ImageID	the unique ID for the image
Tag	the set of user annotated tags

TABLE II: MAJOR ATTRIBUTES FOR USERS

Attribute	Description
UserID	the unique ID for the user
ImageUploaded	the set of images uploaded by the user
FriendList	the user ID of the user friends

In the dataset, there are a total of 2827409 tags among the images, on average there are 14 tags per image and 5422 tags per user. There are 152938 unique tags. The average number of friends of a user is 170 for which there are 902 connections among the 542 users. The goal of the experiment is to infer those connections using the set of images uploaded by the users, even without using the friendship information.

The features of all the images are detected by the Harris-Affine key point detector. They are described by the

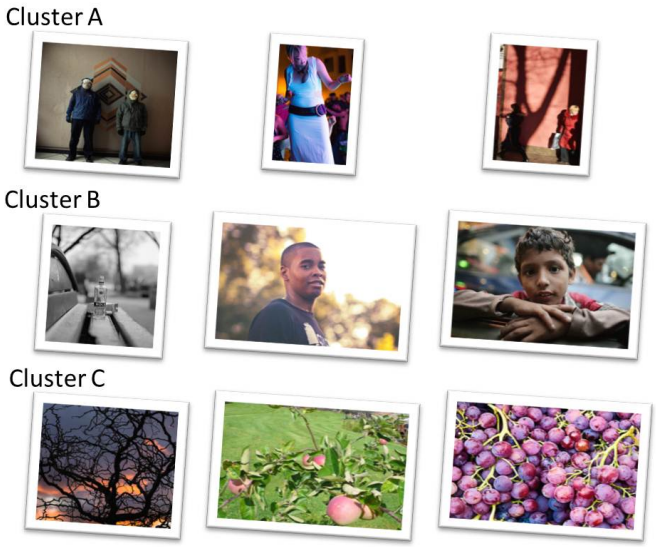


Figure 4: Examples of images in clusters

128-dimension SIFT descriptor. The visual words are obtained by k-mean and then used to represent all the images as feature vectors. A clustering operation is then used to group images with similar feature vectors. The images in the same cluster are assigned with the same auxiliary tag, in which each cluster has a unique auxiliary tag as in Fig. 4. The auxiliary tags are then used to build the user profile. the friendship recommendation is based on the similarity of the user profile of the users.

Different approaches are implemented for comparisons. The first approach is a random approach (Rand), in which users are recommended randomly. This is the baseline for the comparison to simulate the condition that user information such as friendship is not available. Two other approaches are also implemented to compare the proposed one. The first one is the Friend-of-Friend (FoF) [26], in which the similarity is based on the Jaccard similarity of the friend list. The similarity between two users with more common friends is higher. The FoF approach serves as the upper bound of the proposed approach to simulate the condition that information are available. The second approach is similar to the proposed approach, but instead of using auxiliary tag, the user annotated tags are used (UserT). The similarity is based on the tag they used for their images. It is also interesting to check if the additional information from BoFT can improve the preference of the upper bound, FoF. BoFT+ is defined as the following:

$$S_{BoFT+} = \beta * S_{BoFT} + (1 - \beta) * S_{FoF} \quad (2)$$

where β is a constant, S_{BoFT}, S_{FoF} and S_{BoFT+} are the similarities of the BoFT, FoF and BoFT+. A study on the performance with different values of β is carried. It is measured by the area under curve (AUC) on the recall rate with 5 to 10 recommendation and shown in Fig. 5. A higher value in AUC implies a better performance. It is observed that AUC is maximum when β is 0.024. It implies that the majority of information is from FoF. In all approaches, no recommendation is made if the similarity between that user to others are all 0. For example, in the FoF approach, no recommendation is made if a user has no friend. It is a valid

 ISBN: 978-1-61208-364-3

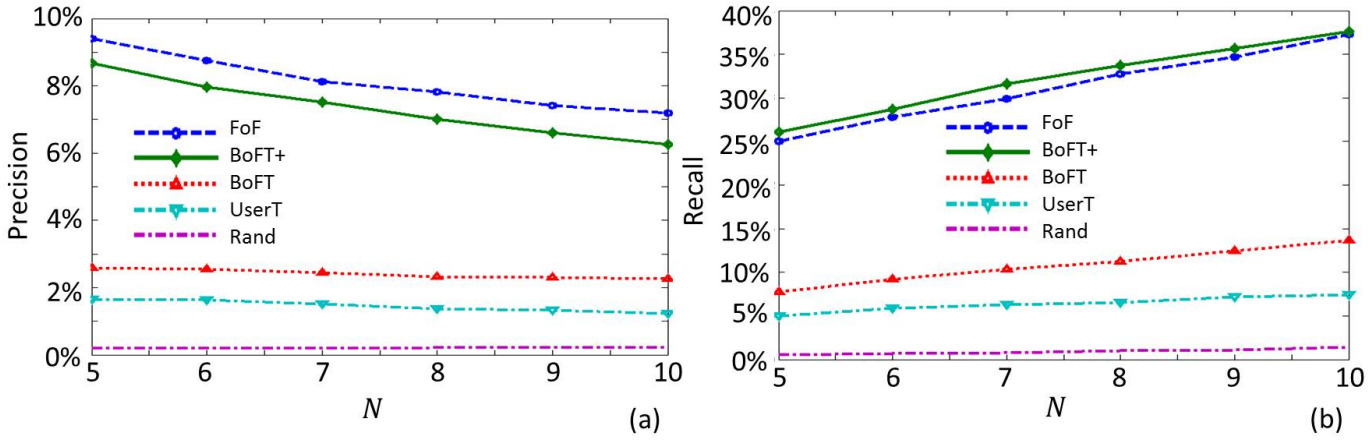


Figure 6: Result rate of different approaches: (a) precision, (b) recall.

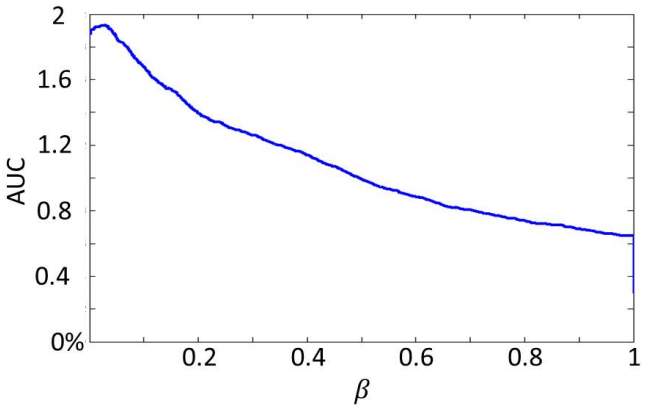


Figure 5: β vs. AUC for $N \subset [5, 10]$

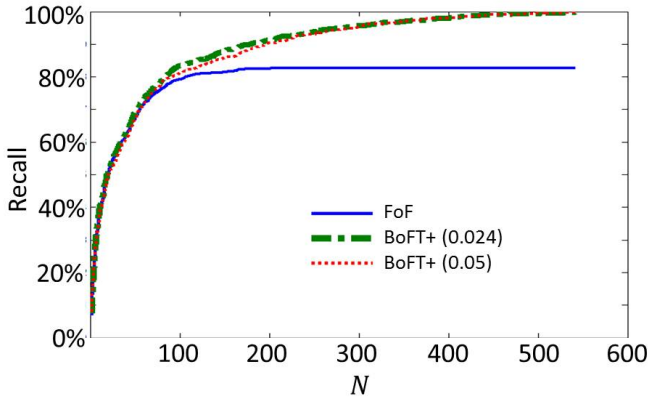


Figure 7: Recall rate for BoFT+ and FoF for $N \subset$[1,542].

assumption when information is not available.

The results are evaluated by two popular rates, top N recall rate and top N precision rate as the following:

$$Precision = \frac{T_p}{(T_p + F_p)} \tag{3}$$

$$Recall = \frac{T_p}{(T_p + F_n)} \tag{4}$$

where T_p is the true positive (the recommended connection is an actual connection), while F_p and F_n are false positive and false negative respectively. F_p is the case that the recommended connection is not a connection, while F_n are the connections that are not recommended. The physical meaning for precision rate is the percentage of recommended items is actual connections. The recall is the percentage of actual connections that is recommended. The higher the values are, the better the approach is. When more items are recommended, recall rate is increased, but precision rate decreases. A list of recommendations is generated for each user with the proposed approaches. The approaches are evaluated with the top N per user recall rate and the top N precision rate, in which the top N users with the highest similarity, are recommended.

B. Results

Fig. 6 shows the top-N precision and recall rate of different algorithms for 5 to 10 recommendations per users. It is observed that FoF, BoFT, UserT and BoFT+ are all better than using user tags in terms of the two rates. It is clear that BoFT approach outperforms UserT and random approach and is much closest to the upper bound, FoF. BoFT+ can only improve the performance slightly. A detailed discussion can be found in the next subsection.

C. Discussion

In the experiment, it is observed that all approaches are better than the random one. The use of BoFT provides a significant improvement on the performance of the recommendation. By BoFT, the recall and the precision rate are improved by about 50% over UserT. Although the recall rate for BoFT+ is slightly higher than FoF, the precision rate of FoF is higher than BoFT+ as shown in Fig. 6. As discussed in previous section, the additional BoFT information can only slightly improve the FoF approach. It is also interesting to check the performance when N is large. The top N recall rate for BoFT+ with 2 values of β and FoF are shown in Fig. 7. When N is smaller than 100, the BoFT+ approach are only slightly better than FoF. However, when N is large, the improvement is more significant and the recall is increased by 13.3% with AUC by BoFT+. As mentioned in the previous section, the higher in

 ISBN: 978-1-61208-364-3

FoF approach in terms of the precision rate is that some users have no common friends with others. If two people have no common friend, no recommendation is possible. On one hand, users with no common friend are never recommended to each other in FoF approach and results in a higher precision rate. The number of common friends between 2 users can be small, or even zero. As a result, FoF approach may only capture relations with high confidence (with common friends) but those without any common friends. A high precision rate but a lower recall rate are obtained. On the other hand, BoFT can connect people through the images they have uploaded and therefore most users can be reached. The BoFT approach provides a new way to connect people. The next research challenge is how to improve the prediction performance such that it is close to the upper bound with the shared images.

VI. CONCLUSION

This paper proposes a novel BoF based Tagging approach to make better recommendations matching users interests discovered from images uploaded It demonstrated how a BoF-based approach can help discover hidden users' connections from the shared images on a social network for a better recommendation. As user friendship/interests are not always available, this paper proposes an unsupervised approach to classify image according to their visual features. Those visual features represent the user interest, and hence recommendation. The proposed approach is evaluated by friendship recommendation with a scraped dataset from Flickr with over 500 users and 200k images. It is proven that the proposed approach can recommend friendship to user based on the image shared by the users. With our proposed approach, BoFT, the recall and the precision rate are improved by about 50%.

ACKNOWLEDGMENT

This work is supported by HKUST-NIE Social Media Lab., HKUST

REFERENCES

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2009, pp. 201-210.

[2] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in Proceedings of the 17th International Conference on World Wide Web, April 2008, pp. 675-684.

[3] T. C. Zhou, H. Ma, M. R. Lyu, and I. King, "UserRec: A user recommendation framework in social tagging systems." In Proceedings of AAAI, July 2010, pp. 1486-1491.

[4] T. Hammond, T. Hannay, B. Lund, and J. Scott, "Social bookmarking tools (I) a general review," D-Lib Magazine, vol. 2, 2005.

[5] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in Social Computing (SocialCom), 2010 IEEE Second International Conference on, August 2010, pp. 88-95.

[6] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: Context and content in community-contributed media collections," in Proceedings of the 15th International Conference on Multimedia, September 2007, pp. 631-640.

[7] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2009, pp. 211-220.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, November 2004, pp. 91-110.

[9] M. Cheung and James She "Predicting Peoples Interests and Connections Using Their Shared Photos," unpublished.

[10] F. E. Walter, S. Battiston, and F. Schweitzer, "A model of a trust-based recommendation system on a social network," Auton. Agents Multi-Agent Syst., vol. 16, February 2008, pp. 57-74.

[11] X. Zhang, Z. Li, and W. Chao, "Tagging image by merging multiple features in a integrated manner," Journal of Intelligent Information Systems, vol. 39, August 2012, pp. 87-107.

[12] B. Sigurbjrnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in Proceedings of the 17th International Conference on World Wide Web, April 2008, pp. 327-336.

[13] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," IEEE Transactions on Multimedia, vol. 14, February 2012, pp. 883-895.

[14] E. Moxley, J. Kleban, J. Xu, and B. Manjunath, "Not all tags are created equal: Learning flickr tag semantics for global annotation," in Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, June 2009, pp. 1452-1455.

[15] X. Zhang et al., "Social image tagging using graph-based reinforcement on multi-type interrelated objects," Signal Processing, vol. 93 no. 8, August 2013, pp. 2178-2189

[16] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu, "Joint Social and Content Recommendation for User-Generated Videos in Online Social Network," IEEE Transactions on Multimedia, April 2013, pp. 698-709

[17] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2010, pp. 194-201.

[18] I. Guy et al., "Personalized recommendation of social software items based on social relations," in Proceedings of the Third ACM Conference on Recommender Systems, October 2009, pp. 53-60.

[19] W. H. Hsu, A. L. King, M. S. Paradesi, T. Pydimarri, and T. Weninger, "Collaborative and structural recommendation of friends using weblog-based social network analysis," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, March 2006, pp. 55-60.

[20] X. Xie, "Potential friend recommendation in online social network," in Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom), December 2010, pp. 831-835.

[21] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2010, pp. 393-402.

[22] R. Parimi and D. Caragea, "Predicting friendship links in social networks using a topic modeling approach," in Advances in Knowledge Discovery and Data Mining Anonymous Springer, May 2011, pp. 75-86.

[23] H. Kim, J. Jung, and A. El Saddik, "Associative face co-occurrence networks for recommending friends in social networks," in Proceedings of Second ACM SIGMM Workshop on Social Media, October 2010, pp. 27-32.

[24] F. Tian, X. Shen, F. Shang, and K. Zhou, "Automatic image tagging by multiple feature tag relevance learning," in Pattern Recognition Anonymous Springer, 2012, pp. 505-513.

[25] D. A. Forsyth and J. Ponce, Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference, 2002.

[26] L. L and T. Zhou, "Link prediction in complex networks: A survey," in Physica A: Statistical Mechanics and its Applications, vol. 390, March 2011, pp. 1150-1170.

 ISBN: 978-1-61208-364-3

Trace Analysis Exploration Using Semantic Web Tools Use Case: You Tube Network Traffic

Oscar Alberto Santana-Alvarez
University of Guadalajara
Zapopan Jalisco México
oscar_santana@cucea.udg.mx

Liliana Ibeth Barbosa-Santillán
University of Guadalajara
Zapopan Jalisco México
ibarbosa@cucea.udg.mx

Gerardo Padilla-Zárate
Intel Corporation. Guadalajara Design Center.
Tlaquepaque, Jalisco, México
gerardo.padilla.zarate@intel.com

***Abstract*—Analysis and exploration of information gathered through local networks are tasks that must constantly be done. Such information is useful for the local network administrators because it gives them a good input about the most effective maintenance procedure for the local network. Maintenance may include activities such as watching over preferences among users, keeping track of the size of the files users are accessing, most viewed videos, etc. It is especially important to watch over Video On Demand (VoD) traffic, such as YouTube-like services providers because of the size of the files they handle and their popularity among users, especially students. One approach to address monitoring activities is network traffic traces, which are sequences of events that recorded specific aspects of a web site. Such traces are usually stored as plain text files (i.e., logs). This paper presents the trace analysis exploration using semantic Web tools, with focus on the You Tube Network Traffic approach, which is based on semantic methods in facilitating network trace analysis by populating two Network Traffic Trace Turtle Files (NTTTF) with network traces obtained by monitoring means. The queries used over the NTTTFs allow us to identify key information presented in the network traffic trace associated with different aspects of our case study. The results showed the feasibility of this approach, where NTTFs improved the way valuable information is being found. Analysis of network traces showed information such as the most viewed videos, the slowest YouTube servers, etc. With this information, more accurate maintenance procedures can be followed. The analysis and exploration of approximately 1,100,000 (one million one hundred thousand) YouTube network traffic traces was performed by means of semantic queries.**

Keywords*—Network Traffic Traces; Semantic Analysis; Local Area Networks.

I. INTRODUCTION

The analysis and exploration of network traces obtained by means of network monitoring is a time consuming task because of the enormous quantity of data and the structure of the files gathered. Besides, human intervention is necessary in order to interpret the information stored in traces. Therefore, it is necessary to implement a framework that support network administrators with the task of finding key information about the use of the local network. This paper presents the exploration and analysis of information stored in network traffic traces. Network traffic traces are data in text format that stores specific aspects of a network. It proposes an approach that facilitates the way network administrators analyze and extract valuable information from network traces by means of semantic web tools. We have the hypothesis that the use of semantic web tools in the analysis and exploration of network traces will allow us to find valuable information in a more manageable and friendly way in order to support network managers.

A. Contributions

The main contributions of the present paper are as follows:

- It presents the steps involved in our approach, which allows the analysis of YouTube Traffic Network traces. The steps include the conceptualization of flat flow of data into NTTTFs. It also includes the queries that were performed in order in order to extract key information about users' behavior about YouTube traffic at local network. The contribution of this paper falls into the category of querying for mining.
- It presents the schema of the NTTTFs, which allows us to perform semantic queries with sparl. NTTTFs help us identify key information and relate important information inside network traces.

B. Structure of the paper

The structure of the paper is organized as follows: In Section 2, we present the related work, while in Section 3, our approach is fully explained. In Section 4, we explain the experiments we performed with our approach. Section 5 explains the results we obtained and, finally, in Section 6, we conclude and mention future work on the subject.

II. RELATED WORK

1) Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications: Zink et al. [1] present a full investigation on how they found out that certain videos are far more popular than others by performing a statistical analysis on the networking traces. They took into account users local preferences instead of international preferences. Therefore, they demonstrated by means of simulations that by performing some actions in the network infrastructure such as the implementation of proxy catching, they could reduce network traffic significantly. We took the same traces Zink et al. used, in order to perform a semantic analysis.

 ISBN: 978-1-61208-364-3

2) Execution Trace Exploration and Analysis Using Ontologies: Al Haider et al. [2] obtained execution traces from a software of a basic calculator and they performed a simple semantic analysis by using ontologies. Although they utilized a reasoner and sparql instructions, they did it with barely 100,000 traces. One of the contributions of the present paper is the implementation of a semantic analysis on a large amount of network traces.

3) Understanding Execution Traces Using Massive Sequence and Circular Bundle Views: Cornelissen et al. [3] proposed a a visual approach to understand the system at hand in order to maintain it. They do so by using dynamic information, e.g., execution traces. They developed a tool that shows the relationships between the method calls through curves in a circular view. They propose a visual approach to gather information from traces while we propose the execution of sparql queries in order to achieve the same goal.

4) Execution Trace Visualization in a 3D Space: Just as Cornelissen et al. [3] proposed a visual approach to understand the system to be analyzed by means of execution traces, Dugerdil et al. [4] proposed a visual approach to analyze execution traces based on 3D space.

5) Exploiting text mining techniques in the analysis of execution traces: Pirzadeh et al. [5] proposed an approach to analyze execution traces by taking into account the traces execution phases. They applied their approach to large traces of two different systems. We did not take into account traces execution phases because such analysis could not be applied to the network traces we choose due to their schema and the content as well.

6) Evaluating distributed real-time and embedded system test correctness using system execution traces: Hill et al. [6] proposed the evaluation of distributed real-time and embedded system test correctness by means of execution traces. Although the authors evaluated a distributed real-time and embedded system by means of execution traces, they did not do it by using semantic tools, which is an important difference from our work.

7) A survey of trace exploration tools and techniques: A survey of trace exploration tools and techniques was done by A. Hamou-Lhadj et al [7]. A review of the different tools and techniques for trace exploration has been made. None of them makes use of TTL files to perform a semantic analysis.

8) Relational Database Approach for Execution Trace Analysis: S. Alouneh et al. [8] proposed and approach to implement relational databases with execution traces. They visualize their proposed architecture and give advantages of their approach even though they did not implement it.

9) A Distributed Architecture for Dynamic Analyses on User-Profile Data: Antoniol et al. [9] proposed a model to represent dynamic information (traces). They collected and comprised traces from a distributed system and and left the door open to manipulate the resultant traces.

10) A Systematic Survey of Program Comprehension through Dynamic Analysis: Cornelissen et al. [10] made a systematic review on program comprehension through dynamic analysis (analysis of execution traces). This study was categorized into four facets: activity, target, method and evaluation.

11) SEAT-A usable trace analysis tool: Hamou-Lhadj et al. [11] proposed a software for a exploration analysis of execution traces. The software supports several features such as filtering techniques or working on several traces while we developed a tool that takes all the traces included in a folder and puts them together in a turtle format.

III. Our Approach

In the present section, an overview of our approach is being given. Figure 1 clarifies our approach.

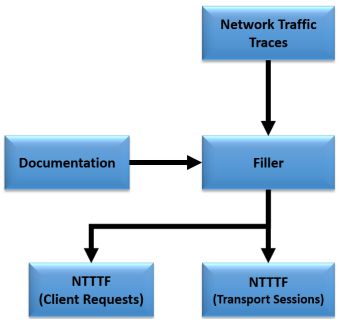


Fig. 1. Steps followed to obtain the NTTTFs.

First, documentation was obtained from UMass [12] traces repository. The downloaded documentation file has the schema of the network traces. Such a schema served us to code the algorithm of The Filler.

After that, both NTTTFs were populated by means of a software developed in .NET platform. This software is called The Filler. The Filler was developed in C# programming language. It generates two turtle files with .ttl extensions (NTTTF Client Requests and NTTTF Transport Session) and inserts into them the network traces also downloaded from the UMass [12] repository. Once we had the NTTTFs populated, semantic queries were performed in order to find key information.

Al Haider et al. [2] also proposed the use of semantic tools for traces. However, they used an ontology on execution traces they obtained from a software of a basic calculator. This leads to several differences; for example, the use of a reasoner to check consistency of the ontology. Also, they did its semantic analysis with barely 100,000 traces. One of the contributions of the present paper is the implementation of a semantic analysis on a large amount of network traces

Alouneh et al. [8] proposed an approach to implement relational databases with execution traces. They visualize their proposed architecture and give advantages of their approach even though the authors did not implement it. The main reason why we did not utilize relational databases is because we performed several tests before the experiments shown in the present paper. We had the serious restriction of using only 400,000 traces on the Jena Framework [13], so, we adjusted

 ISBN: 978-1-61208-364-3

the resultant schema of the NTTF files and we were able to perform sparql queries on 1,100,000 traces in Jena. That is to say, we wanted to put to the test Jena Framework on the quantity of traces it can handle. Doing it with relational databases has the major drawback that every corporation that has a software for relational databases has his own implementation. We wanted to make our research as portable as possible by using turtle text files. It is not the scope of this paper to compare the efficiency of tools that make use of text files against tools that make use of relational databases. The scope of the present paper is to demonstrate that with network traces, keeping traces as text files and the use of semantic tools, we can find information for the decision taking. Even though efficiency is an important variable to take into account, it is the advantages of using ontologies the point we want to show.

A. Documentation

As a first step, we obtain the documentation from the UMass [12] repository. The document identifies two types of network traces:

- Trace files that contain information about client requests for YouTube video clips.
- Trace files that contain information about the transport session for video clips requested by clients from the UMass campus network.

Table 1 shows in the left column the fields found in the transport session trace files and an example in the right column. This is for clarification purposes.

TABLE I
SCHEMA AND EXAMPLE OF A TRANSPORT SESSION NETWORK TRACE

Field	Example
id	# 5
source_ip (anonymized)	64.15.112.107
sport	80
dest_ip (anonymized)	148.85.44.11
dport	2365
_pro	6
dir	1
start_time	0.464492
finishtime	74.901594
duration (in seconds)	74.437101
datapkts	3182
size_in_bytes	4644692
rate	499.180
flags	16

All parameters are self-explanatory except, for the flags parameter. This attribute is being used for control purposes, in this special case, network administrators. One thing to note about this attribute is that its value was always 16 in all the traces. So, this attribute, nor by itself, neither in conjunction with another attribute, gave us any information at all about the trace. Table 2 shows the fields found in the documentation for the client request trace files on the left column, while an example network trace is being shown on the right column.

TABLE II
SCHEMA AND EXAMPLE OF A CLIENT REQUEST NETWORK TRACE

Field	Example
timestamp	1189828805.208862
YouTube Server IP (anonymized)	63.22.65.73
sport	80
Client IP (anonymized)	140.8.48.66
Request	GETVIDEO
Video Id	lML9dik8QNw
Content Server IP	158.102.125.12

B. Network Traffic Traces

Network traffic trace files contain the data collected by monitoring the local network. There are 21 files that contain two types of network traffic traces, as was said before, YouTube Client Requests and YouTube Transport Sessions. Figure 2 shows a fragment of the content of a network traffic trace file. Each file has .dat extension and the information stored in every one of them is plain text.

```
1189828805.208862 63.22.65.73 140.8.48.66 GETVIDEO lML9dik8QNw 158.102.125.12
1189828810.212831 63.22.65.73 35.139.191.73 GETVIDEO bBXyB7niEc0&origin 105.136.66.5
1189828829.197218 63.22.65.77 205.181.191.47 GETVIDEO qVEcloLm4I4 63.22.67.110
1189828830.228538 63.22.65.73 102.15.239.161 GETVIDEO FKOn8OMN7F8 63.22.64.40
1189828832.795029 63.22.65.77 102.15.228.156 GETVIDEO WPcL8swzt8k 254.212.22.126
1189828833.031500 63.22.65.77 140.8.55.234 GETVIDEO OBSwS-QAv68&origin 105.136.66.5
1189828833.901196 63.22.65.73 140.8.49.142 GETVIDEO B361KLb_PlI&origin 105.136.66.5
1189828838.399229 63.22.65.73 140.8.48.66 GETVIDEO I0AKol7FIAA 254.212.22.75
1189828841.713910 63.22.65.77 102.15.255.4 GETVIDEO dciqhrfFXFw&origin 105.136.66.5
1189828841.792919 63.22.65.77 102.15.226.143 GETVIDEO 0bpgvFTF37c 254.212.22.187
1189828842.720476 63.22.65.73 140.8.54.40 GETVIDEO jXq3hO82cTY&origin 105.136.66.5
1189828846.842593 63.22.65.77 205.181.182.53 GETVIDEO XXG1-hlMX1o 254.212.18.215
1189828852.098143 63.22.65.73 102.15.234.3 GETVIDEO Oeq3XblvufE 254.212.17.132
1189828852.779726 63.22.65.73 205.181.167.225 GETVIDEO nFkW-bReI0c&origin 105.136.66.5
1189828853.623512 63.22.65.73 102.15.239.161 GETVIDEO QA4aJnYOChE 63.22.64.65
1189828859.296750 63.22.65.73 140.8.48.66 GETVIDEO Ksxhx9Nh6lQ 254.212.18.195
1189828860.421157 63.22.65.73 205.181.190.224 GETVIDEO aNrY9OcS-y8 254.212.19.76
1189828863.792110 63.22.65.77 102.15.237.62 GETVIDEO tl1IgCfZ7PA 254.212.23.146
1189828870.055699 63.22.65.73 140.8.54.219 GETVIDEO _s1YngtN3y4 63.22.64.32
1189828880.704996 63.22.65.73 205.181.190.224 GETVIDEO M5lnwJL5Mew 254.212.29.90
1189828885.808688 63.22.65.77 102.15.229.19 GETVIDEO CH1VqsQPTrE 254.212.19.76
1189828887.811597 63.22.65.73 205.181.188.157 GETVIDEO 82w3FkY5kPE&origin 105.136.66.5
1189828895.114468 63.22.65.73 35.139.180.61 GETVIDEO 53d7ikNox3c 254.212.29.239
1189828896.696499 63.22.65.77 35.139.31.8 GETVIDEO _-sNIWi2fLs 254.212.23.155
```

Fig. 2. Network Traces located in the .dat files.

As can be seen, each line represents a network traffic trace and the order of the fields is the same as shown in Section 3.

C. The Filler

The Filler automatically creates two empty files, ClientRequests.ttl and TransportSession.ttl, the NTTTFs, and automatically populates them with the information stored inside the network traffic traces files. It was because we were dealing with a great number of elements in the trace files, approximately 1,100,000 (one million one hundred thousand) traces, and doing so manually would have been a very hard and time consuming task. The Filler was created in order to cope with this task. Figure 3 shows the main screen of the software with all the controls visible and enabled in order for the reader to look at all the controls that are present in the Filler.

Figure 4 shows the software running with the first step already being taken. It is a software that consists of three simple steps:

- Step 1 - Load Network Traces (Client Requests). By pressing this button, a window will appear in which the

 ISBN: 978-1-61208-364-3

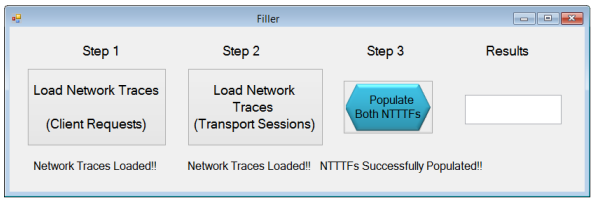


Fig. 3. Main screen of Filler. All controls enabled for clarification.

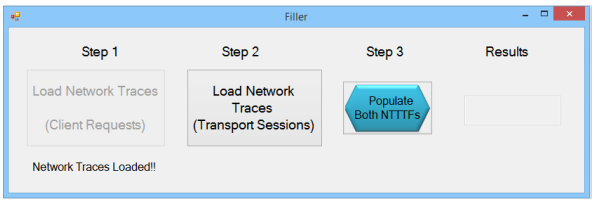


Fig. 4. The Filler running.

user can select the folder of the computer that contains the network traffic traces that correspond to the client requests. It is important to note that all the files that are inside the folder will be read in order to extract the traces from them. The latter was done in order to ensure minimum intervention from the user with the aim of making the software more easy to use. The folder needs to contain only the files that have traces within them. In our approach, the NTTTF Client Request was populated using two YouTube network traffic traces files. Their name start with the prefix flows.xxxx and have the .dat extension.

- Step 2 - Load Network Traces (Transport Sessions). By pressing this button a window will appear in which the user can select the folder of the computer that contains the network traffic traces that correspond to the transport session. The files that are inside the folder will be read in order to extract the traces from them. The folder needs to contain only the files that have traces within them. In our approach, the NTTTF Transport Session was populated using 19 YouTube network traffic traces files. Their name start with the prefix youtube.parsed.xxxxxx and have .dat extension.
- Step 3 - Populate Both NTTTFs. This button triggers the method that takes the selected folders and reads all the files included in them one by one. Each line of every file that is being read, is being parsed to the corresponding NTTTF with text marks that adjust the Turtle standard in order to obtain a Resource Description Format (RDF) [14] compliant file.

IV. Experimentation and Results

Once both NTTTF files (client request and transport session) were generated by The Filler, Jena Framework [13] was installed on the computer in order to perform the experiments. The Jena Framework has within the sparql [15] tool, which was used for performing semantic queries over the NTTTFs. Sparql is a RDF [14] query language that allows us to perform semantic queries, updating, copying and creation of data, etc. For our approach only the execution of semantic queries was needed.

The aim of the tests was to find valuable information that support the decision making of the local network managers. In order to perform the tests, the following steps were followed:

- Step 1 - Every sparql query was written down inside a text file with a .rq extension.
- Step 2 - The sparql application was executed with every query following the syntax:

sparql –query queryfile.rq –data NTTTFxx.ttl –time

where queryfile.rq was substituted by 1-Top100v1.rq, 2-TopDuration.rq and 3-TopSlowServers.rq

where NTTTFxx.ttl was substituted by NTTTFCR.ttl (Client Request) and NTTTFTS.ttl (Transport Session).

where the -time clause measures the time the query needed to execute. Finally, we interpret the results and demonstrate the valuable information that can be found by means of semantic tools.

A. First Test

The aim of the first test was to locate the top 30 most viewed videos. The purpose of this query is to provide local network managers with a list of the videos that can be stored in the cache local server in order to reduce the traffic outside of the local network, increasing the bandwidth left for other users. The first test made use of the NTTTF Client Request. Figure 5 shows the sparql query:

```
prefix
ntttfcr: <http://ntttfcr.com/ns/NTTTFCR#>

SELECT (COUNT(?ntttfcr) as ?Record_Quantity)
?VideoId

WHERE
{
  ?ntttfcr ntttfcr:VideoId ?VideoId.
}

GROUP BY ?VideoId
ORDER BY DESC(?Record_Quantity)
LIMIT 30
```

Fig. 5. First test sparql query.

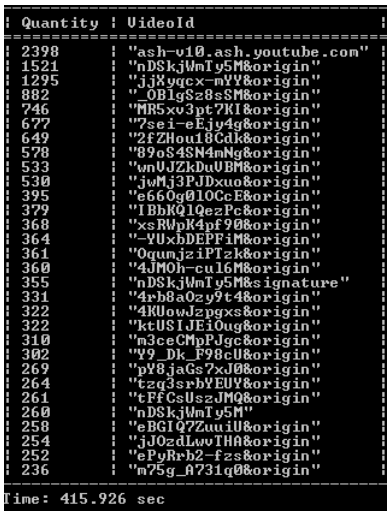
Here is the shell command that was necessary in order to execute the query:

sparql –query 1-Top100v1.rq –data NTTTFCR.ttl –time

Figure 6 shows the result thrown by the sparql tool.

The left column in Figure 6 shows the number of times a certain YouTube was visited. In the right column, there are the video identifiers. The VideoId that is on top is the video that has been accessed the most number of times (2398 times). This means that, if we backup this video in a local cache server, 2398 accesses outside of the local network will be saved. Now, by adding the top 100 hits of the same query, we get 28793 accesses. If we follow the same procedure, we would be saving

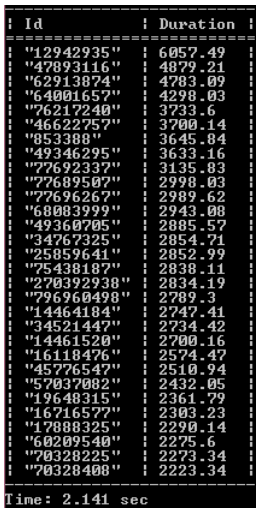
 ISBN: 978-1-61208-364-3



Quantity	VideoId
2398	"ash-v10.ash.youtube.com"
1521	"nDSkjWmIy5M&origin"
1295	"jjXyqcx-mYY&origin"
882	"_OBlgSz8sSM&origin"
746	"MR5xv3pt7KI&origin"
677	"7sei-eEjy4g&origin"
649	"2fZHou18Cdk&origin"
578	"89oS4SN4mNg&origin"
533	"wnUJZkDuVBM&origin"
530	"jwMj3PJDxuo&origin"
395	"e66Qg0lOCcE&origin"
379	"IBbKQlQezPc&origin"
368	"xsRWpK4pf90&origin"
364	"-YUxbDEPFiM&origin"
361	"OqumjziPIzk&origin"
360	"4JMOh-cul6M&origin"
355	"nDSkjWmIy5M&signature"
331	"4rb8aOzy9t4&origin"
322	"4KUowJzpgxs&origin"
322	"ktUSIJEiOug&origin"
310	"m3ceCMpPJgc&origin"
302	"Y9_Dk_P98cU&origin"
269	"pY8jaGs7xJ0&origin"
264	"tzq3srbYEUY&origin"
261	"tFfCsUszJMQ&origin"
260	"nDSkjWmIy5M"
258	"eBGIQ7ZuuiU&origin"
254	"jJOzdLwvTHA&origin"
252	"ePyRrb2-fzs&origin"
236	"m75g_A73iq0&origin"

Time: 415.926 sec

Fig. 6. Results of the first test.



Id	Duration
"12942935"	6057.49
"47893116"	4879.21
"62913874"	4783.09
"64001657"	4298.03
"76217240"	3733.6
"46622757"	3700.14
"853388"	3645.84
"49346295"	3633.16
"77692337"	3135.83
"77689507"	2998.03
"77696267"	2989.62
"68083999"	2943.08
"49360705"	2885.57
"34767325"	2854.71
"25859641"	2852.99
"75438187"	2838.11
"270392938"	2834.19
"796960498"	2789.3
"14464184"	2747.41
"34521447"	2734.42
"14461520"	2700.16
"16118476"	2574.47
"45776547"	2510.94
"57037082"	2432.05
"19648315"	2361.79
"16716577"	2303.23
"17888325"	2290.14
"60209540"	2275.6
"70328225"	2273.34
"70328408"	2223.34

Time: 2.141 sec

Fig. 8. Results of the second test.

28693 accesses outside of the local network. We think this would represent a huge saving in bandwidth if measures for saving would be applied. Another piece of information that can be noted from Figure 6 is the time it took for the sparql tool to perform the query, namely 415.926 seconds. This represents a small amount of time taking into account the quantity of traces analyzed.

B. Second Test

The second test we performed was to locate the id of the longest transport sessions. We suggest that a local backup of the longest transport sessions will increase overall bandwidth available for the users. This test made use of the NTTTF Transport Session. Figure 7 shows the sparql query:

```
prefix
ntttfts: <http://ntttfts.com/ns/NTTTFTS#>

SELECT ?Id ?Duration ?Size ?Rate
(((?Size/?Duration)/1024) AS ?Velocidad)
WHERE
{
  ?ntttfts ntttfts:Id ?Id.
  ?ntttfts ntttfts:Duration ?Duration.
  ?ntttfts ntttfts:Size ?Size.
  ?ntttfts ntttfts:Rate ?Rate.
}

ORDER BY DESC(?Percentage_Accuracy)
LIMIT 30
```

Fig. 7. Second test sparql query.

Here is the ms-dos command that was necessary in order to execute the query:

sparql –query 2-TopDuration.rq –data NTTTFTS.ttl

Figure 8 shows the result thrown by the sparql tool.

Figure 8 shows on the left column the identifiers of the transport sessions with the highest duration in seconds (right column). As can be seen, we have very long sessions (6057.49 seconds the longest of them). Presumably, the longer the transport sessions, the longer videos last. We are sure that by storing not only most viewed videos but also the longest videos in a local cache server will have a direct impact in the overall available bandwidth for the rest of the users of local network. Now, by summing the top 100 hits of the same query we get 202,608.46 seconds in total. This means that by backing up these videos, we will be unloading local network from this amount of time. Another piece of information that can be noted from Figure 8 is the time it took for the sparql tool to perform the query, namely 2.141 seconds.

C. Third Test

The third test we performed was to locate the slowest YouTube servers. We suggest that a local backup of the slowest servers' videos is necessary in order to increase overall bandwidth available for the users. This test made use of the NTTTF Transport Session. Here is the sparql query:

```
prefix
ntttfcr: <http://ntttfcr.com/ns/NTTTFCR#>

SELECT (COUNT(?ntttfcr) as ?Record_Quantity)
?VideoId

WHERE
{
  ?ntttfcr ntttfcr:VideoId ?VideoId.
}

GROUP BY ?VideoId
ORDER BY DESC(?Record_Quantity)
LIMIT 30
```

Fig. 9. Third test sparql query.

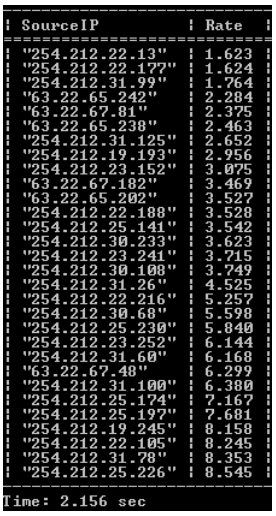
Here is the ms-dos command that was necessary in order to execute the query:

sparql –query 3-TopSlowServers.rq –data NTTTFTS.ttl

Figure 10 shows the result thrown by the sparql tool.

Figure 10 shows on the left column the IP addresses with the lowest rates (right column). This means that these are the slowest YouTube servers. The identification of slow servers

 ISBN: 978-1-61208-364-3



SourceIP	Rate
"254.212.22.13"	1.623
"254.212.22.177"	1.624
"254.212.31.99"	1.764
"63.22.65.242"	2.284
"63.22.67.81"	2.375
"63.22.65.238"	2.463
"254.212.31.125"	2.652
"254.212.19.193"	2.956
"254.212.23.152"	3.075
"63.22.67.182"	3.469
"63.22.65.202"	3.527
"254.212.22.188"	3.528
"254.212.25.141"	3.542
"254.212.30.233"	3.623
"254.212.23.241"	3.715
"254.212.30.108"	3.749
"254.212.31.26"	4.525
"254.212.22.216"	5.257
"254.212.30.68"	5.598
"254.212.25.230"	5.840
"254.212.23.252"	6.144
"254.212.31.60"	6.168
"63.22.67.48"	6.299
"254.212.31.100"	6.380
"254.212.25.174"	7.167
"254.212.25.197"	7.681
"254.212.19.245"	8.158
"254.212.22.105"	8.245
"254.212.31.78"	8.353
"254.212.25.226"	8.545

Time: 2.156 sec

Fig. 10. Results of the second test.

allows for actions to be taken to improve the situation, such as the search for alternate servers. These are actions taken by some network devices, such as routers, in order to enhance the overall performance of the network. Sometimes there is no solution, no alternate servers for example, to this problem. We are sure that by identifying slow servers and reducing access to them will not only diminish the network usage but also reduce the work load of network devices that perform actions to bypass them. Another piece of information that can be noted from Figure 10 is the time it took for the sparql tool to perform the query, namely 2.156 seconds, which is very little time for the information we gather.

D. Fourth Test

This test represents the last bastion we need to ensure we can have a direct impact on the enhancement of the local network by using semantic web tools. The fourth test we performed was to search for atypical values on the duration on the transpot sessions traces. An example of an atypical value may be a session that lasted too long for a small file with a good rate. This test made use of the NTTTF Transport Session. Figure 11 shows the sparql query.

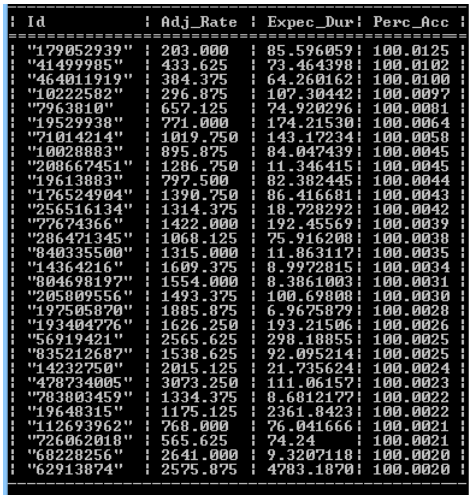
```
prefix
ntttfcr: <http://ntttfcr.com/ns/NTTTFCR#>

SELECT (COUNT(?ntttfcr) as ?Record_Quantity)
?VideoId

WHERE
{
   ?ntttfcr ntttfcr:VideoId ?VideoId.
}

GROUP BY ?VideoId
ORDER BY DESC(?Record_Quantity)
LIMIT 30
```

Fig. 11. Fourth test sparql query.

Here is the ms-dos command that was necessary in order to execute the query:

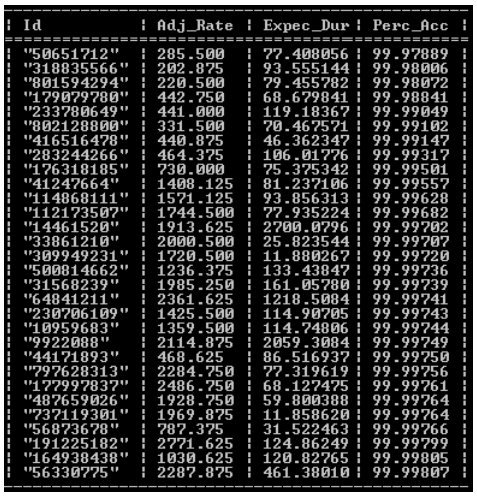
sparql –query 4-Anomalies.rq –data NTTTFTS.ttl

Figures 12 and 13 show the result given by the sparql tool.



Id	Adj_Rate	Expec_Dur	Perc_Acc
"179052939"	203.000	85.596059	100.0125
"41499985"	433.625	73.464398	100.0102
"464011919"	384.375	64.260162	100.0100
"10222582"	296.875	107.30442	100.0097
"7963810"	657.125	74.920296	100.0081
"19529938"	771.000	174.21530	100.0064
"71014214"	1019.750	143.17234	100.0058
"10028883"	895.875	84.047439	100.0045
"208667451"	1286.750	11.346415	100.0045
"19613883"	797.500	82.382445	100.0044
"176524904"	1390.750	86.416681	100.0043
"256516134"	1314.375	18.728292	100.0042
"77674366"	1422.000	192.45569	100.0039
"286471345"	1068.125	75.916208	100.0038
"840335500"	1315.000	11.863117	100.0035
"14364216"	1609.375	8.9972815	100.0034
"804698197"	1554.000	8.3861003	100.0031
"205809556"	1493.375	100.69808	100.0030
"197505870"	1885.875	6.9675879	100.0028
"193404776"	1626.250	193.21506	100.0026
"56919421"	2565.625	298.18855	100.0025
"835212687"	1538.625	92.095214	100.0025
"14232750"	2015.125	21.735624	100.0024
"478734005"	3073.250	111.06157	100.0023
"703803459"	1334.375	8.6812177	100.0022
"19648315"	1175.125	2361.8423	100.0022
"112693962"	768.000	76.041666	100.0021
"726062018"	565.625	74.24	100.0021
"68228256"	2641.000	9.3207118	100.0020
"62913874"	2575.875	4783.1870	100.0020

Fig. 12. Results of the fourth test with DESC clause.



Id	Adj_Rate	Expec_Dur	Perc_Acc
"50651712"	285.500	77.408056	99.97889
"318835566"	202.875	93.555144	99.98006
"801594294"	220.500	79.455782	99.98072
"179079780"	442.750	68.679841	99.98841
"233780649"	441.000	119.18367	99.99049
"802128800"	331.500	70.467571	99.99102
"416516478"	440.875	46.362347	99.99147
"283244266"	464.375	106.01776	99.99317
"176318105"	730.000	75.375342	99.99501
"41247664"	1408.125	81.237106	99.99557
"114868111"	1571.125	93.856313	99.99628
"112173507"	1744.500	77.935224	99.99682
"14461520"	1913.625	2700.0796	99.99702
"33861210"	2000.500	25.823544	99.99707
"309949231"	1720.500	11.880267	99.99720
"500814662"	1236.375	133.43847	99.99736
"31568239"	1985.250	161.05780	99.99739
"64841211"	2361.625	1218.5084	99.99741
"230706109"	1425.500	114.90705	99.99743
"10959683"	1359.500	114.74806	99.99744
"9922088"	2114.875	2059.3084	99.99749
"44171893"	468.625	86.516937	99.99750
"797628313"	2284.750	77.319619	99.99756
"177997837"	2486.750	68.127475	99.99761
"487659026"	1928.750	59.800388	99.99764
"737119301"	1969.875	11.858620	99.99764
"56873678"	787.375	31.522463	99.99766
"191225182"	2771.625	124.86249	99.99799
"164938438"	1030.625	120.82765	99.99805
"56330775"	2287.875	461.38010	99.99807

Fig. 13. Results of the fourth test with ASC clause.

Figures 12 and 13 show, from left to right, the Id of the transport session, adjusted rate, expected duration, percentage of accuracy. The main difference is that Figure 12 shows the top most different transport sessions Ids and Figure 13 shows the lowest different transport sessions. This query works as follows: first, we needed to adjust the rate to obtain 100% accuracy. That is the reason why we multiply the rate by 125. Later on, we divided the size and adjusted the rate. By doing this we obtained the expected duration of a transport session. Finally, we compared the duration of the transport session with the expected duration. This comparison gave us a percentage value of accuracy. Every value that has a bias of more than 1% would be considered as atypical. As it can be seen from both figures, there is no atypical value. This means that all values from this trace are perfectly explainable.

 ISBN: 978-1-61208-364-3

V. Conclusion and Future Work

Semantic web tools, such as sparql from Jena Framework, allow us to perform semantic queries in order to extract valuable information from raw data. As shown in Section 4, the information gathered by means of semantic queries could help people (local network administrators in this case) to better understand the system at hand in order to take precise decision over its maintenance. It is important to note that the total amount of time for the three queries to execute in a laptop with standard specifications (Core i5, 8Gb RAM, 1TB HDD) was of 420.223 seconds (7 minutes). It is a very small amount of time taking into account the network traces quantity, approximately 1100000 (1 million one hundred thousands). That is to say, approximately 10 minutes are necessary in order to follow our approach. It is our belief that more experimentation must be done in order to take advantage of more semantic tools that are at our disposal, for example, ontologies. Further investigation will be done with network traces but next time, we will take advantage of the characteristics of ontologies in order to perform more sophisticated and complex processes. Ontologies are formal representations of information and have characteristics such as datatype properties, object properties, inference of information, etc., that makes them attractive to be applied to traces. One way of applying ontologies to traces is by defining a model of an ontology that matches the schema of the trace at hand. After that, we need to look for the technology that best suits the expecting results, a software for example. Next steps are uncertain and will depend on what authors will seek to solve with the ontology. This will allow us to gather more accurate and concluding information for the decision making.

VI. Acknowledgments

Oscar Alberto Santana Alvarez will like to thank the Consejo Nacional de Ciencia y Tecnología (CONACyT) for their support for his PhD program. Also, we will like to thank to Michael Zink, Kyoungwon Suh and Jim Kurose for giving us access to the network traffic traces they gathered. These files were the principal input for the present paper.

References

[1] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," Comput. Netw., vol. 53, no. 4, pp. 501–514, Mar. 2009, [retrieved: 05, 2014]. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2008.09.022

[2] N. Al Haider, B. Gaudin, and J. Murphy, "Execution trace exploration and analysis using ontologies," in Proceedings of the Second International Conference on Runtime Verification, ser. RV'11. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 412–426, [retrieved: 05, 2014]. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29860-8_33

[3] B. Cornelissen et al., "Understanding execution traces using massive sequence and circular bundle views," in Proceedings of the 15th IEEE International Conference on Program Comprehension, ser. ICPC '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 49–58, [retrieved: 05, 2014]. [Online]. Available: http://dx.doi.org/10.1109/ICPC.2007.39

[4] P. Dugerdil and S. Alam, "Execution trace visualization in a 3d space," in Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on, April 2008, pp. 38–43.

[5] H. Pirzadeh, A. Hamou-Lhadj, and M. Shah, "Exploiting text mining techniques in the analysis of execution traces." in ICSM. IEEE, 2011, pp. 223–232, [retrieved: 05, 2014]. [Online]. Available: http://dblp.uni-trier.de/db/conf/icsm/icsm2011.html#PirzadehHS11

[6] J. H. Hill, P. Varshneya, and D. C. Schmidt, "Evaluating distributed real-time and embedded system test correctness using system execution traces." Central Europ. J. Computer Science, vol. 1, no. 2, pp. 167–184, 2011, [retrieved: 05, 2014]. [Online]. Available: http://dblp.uni-trier.de/db/journals/cejcs/cejcs1.html#HillVS11

[7] A. Hamou-Lhadj and T. C. Lethbridge, "A survey of trace exploration tools and techniques," in Proceedings of the 2004 Conference of the Centre for Advanced Studies on Collaborative Research, ser. CASCON '04. IBM Press, 2004, pp. 42–55, [retrieved: 05, 2014]. [Online]. Available: http://dl.acm.org/citation.cfm?id=1034914.1034918

[8] S. Alouneh, S. Abed, B. Mohd, and A. Al-Khasawneh, "Relational database approach for execution trace analysis," in Computer, Information and Telecommunication Systems (CITS), 2012 International Conference on, May 2012, pp. 1–4.

[9] G. Antoniol and M. D. Penta, "A distributed architecture for dynamic analyses on user-profile data." in CSMR. IEEE Computer Society, 2004, pp. 319–328, [retrieved: 05, 2014]. [Online]. Available: http://dblp.uni-trier.de/db/conf/csmr/csmr2004.html#AntoniolP04

[10] B. Cornelissen, A. Zaidman, A. van Deursen, L. Moonen, and R. Koschke, "A systematic survey of program comprehension through dynamic analysis." IEEE Trans. Software Eng., vol. 35, no. 5, pp. 684–702, 2009, [retrieved: 05, 2014]. [Online]. Available: http://dblp.uni-trier.de/db/journals/tse/tse35.html#CornelissenZDMK09

[11] A. Hamou-Lhadj, T. Lethbridge, and L. Fu, "Seat: a usable trace analysis tool," in Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on, May 2005, pp. 157–160.

[12] T. Weibel. Network - umass trace repository. [accessed: 05, 2014]. [Online]. Available: http://traces.cs.umass.edu/index.php/Network/Network

[13] Apache.org. Jena framework. [accessed: 05, 2014]. [Online]. Available: http://www.apache.org/

[14] W3.org. Rdf - semantic web standards. [accessed: 05, 2014]. [Online]. Available: http://www.w3.org/RDF/

[15] W3.org. Sparql query language for rdf. [accessed: 05, 2014]. [Online]. Available: http://www.w3.org/RDF/

 ISBN: 978-1-61208-364-3

A SQL-based Context Query Language for Context-aware Systems

Penghe Chen, Shubhabrata Sen, Hung Keng Pung and Wai Choong Wong
National University of Singapore
Singapore
e-mails:{g0901858, g0701139, dcsphk, elewwcl}@nus.edu.sg

Abstract—**Context-aware computing is a typical paradigm of ubiquitous computing and aims to provide context information anywhere and anytime. Context data management which handles gathering, processing, managing, evaluating and disseminating context information is the heart of context-aware system. Context provision and acquisition, thus, become crucial for context-aware computing. In order to decouple application developers from the tedious work of managing underlying context data sources, a proper context query language should be defined to express context information requirements without considering details of underlying structure. Different types of context queries have been proposed previously and an evaluation demonstrates that SQL-based and RDF-based query languages are most powerful in expressing context queries. However, although the RDF-based languages are more suitable for expressing relations and reasoning operations, it is not as flexible as the SQL-based methods in representing user requirements. Additionally, the RDF-based language creates a large amount of overheads due to its various definitions of classes, sub-classes and relations. In order to address these issues, we propose and design this SQL-based context query language which is easy to use and very flexible to express queries with different constraints. It supports both pull and push based context data retrieval, as well as different context processing functions to generate meaningful context information.**

Keywords-context; context-awareness; context query language; SQL-based; context data management; context provision and acquisition; ubiquitous computing.

I. INTRODUCTION

The advances of wireless communication and mobile computing technologies have brought people into the era of ubiquitous computing. Context-aware computing which is a paradigm of ubiquitous computing, aims to provide information to people anywhere and anytime. As a result, applications can recognize and adapt to changes in the environment automatically. Context is usually defined as any information that can be used to characterize the users' situations [1], while context-aware systems are systems that can leverage on context information to adapt their behaviors in order to realize context-awareness [15].

An important part of a context-aware system is a context data management system which takes charge of gathering, processing, managing, evaluating and disseminating context information [15]. A properly designed context data management system could decouple the lower level data collection part from the upper layer application design, so that developers are not concerned with the actual details of the underlying data collection. Additionally, context data management systems can perform intermediate context processing to improve the quality of context data. In order to better utilize -systems, application developers should have access to programming tools that abstract the underlying details of the context data collection process.

An important part of this problem is to define and design an appropriate Context Query Language (CQL) to formally represent context data acquisition requirements through queries. A context query language is a formal language for representing queries in context-aware systems and defines the basic structure and syntax for a context query [15]. A properly defined CQL can ease the process of expressing the required context data retrieval conditions as well as the corresponding context query processing operations.

Query languages designed for traditional database systems are not suitable for context data querying due to the special characteristics of context data. First, context data is usually dynamic, which may produce frequent updating operations and easily cause data inconsistency using traditional database system. Additionally, context data is usually not well structured and can be various kinds of information, such as situation information or metadata information, which cannot be properly represented by query languages designed for well structured schema based data. Furthermore, context data can come from heterogeneous and distributed context sources [7], which makes traditional query languages hard to represent the data information. In addition, context data management requires lots of context reasoning operations to derive higher level information [9], which cannot be represented by traditional query languages. A properly designed CQL should take care of these characteristics.

In order to address these issues, different methods have been proposed to design CQL as discussed in [3][6][9]: Structure Query Language (SQL)-based, Resource Description Framework (RDF)-based, Graph-based, Extensible Markup Language (XML)-based and Application Programming Interface (API)-based query languages. By analyzing the advantages and drawbacks of each type of these query languages, the evaluation conducted by Haghighi et al. [6] demonstrates that RDF- and SQL-based CQLs are more powerful and effective. However, compared with SQL-based method, RDF-based method is more specialized for RDF or ontology based context representation, which limits

 ISBN: 978-1-61208-364-3

its capability and makes it hard to be integrated with traditional database data. Additionally, RDF is relatively complex in defining and creating various kinds of classes and subclasses as well as relations, which produce large amounts of overhead and make it difficult to implement. On the other hand, an SQL-like language is very flexible and can significantly reduce the learning curve, providing a developer has worked with standard SQL [19].

In this paper, we discuss the design of a new SQL-based CQL that helps application developers to construct context queries to acquire context data from various domains and different context sources. Our proposed CQL can also support different types of context acquisition manners as well as context operations.

The rest of the paper is organized as follows. Analysis of requirements for a well-defined CQL is given in Section II. Section III presents some discussions on existing CQLs. The proposed SQL-based CQL is described in Section IV. Section V discusses the evaluation and the paper is concluded in Section VI.

II. REQUIREMENTS

SQL is a widely utilized query language in relational database management systems. However, directly utilizing SQL in context data management is not possible as context data has its own characteristics that are different from relational database data. This is why a separate CQL is required to support queries on context data, but the powerfulness and experiences of SQL can be learnt to create a CQL upon the existing SQL structure. Compared to traditional database data, context data has its own special characteristics [6].

- **Dynamic or static**: unlike traditional database data which are usually static, most of context data can be either static or dynamic which makes the management and accessing differently.
- **Stream data**: context data may be queried and accessed in a continuous or periodic manner which is different from traditional database queries usually with specific or discrete data.
- **Temporal and spatial relationship**: different from traditional database data, context data is closely related to spatial and temporal information.
- **Situational information**: compared to simple and clear traditional database data, context data can be derived situational information based on other information.
- **Unstructured**: most of time, context data do not have a predefined uniform schema like that in traditional database data.

All these dedicated characteristics of context data should be carefully considered while designing a CQL. In addition, the inherent nature of context data management can also attribute some requirements on the design of CQL. Some characteristics of pervasive computing environment are summarized by Perich et al. [13]: autonomy, mobility, distribution and heterogeneity, which also create challenges to the design of CQL, just as discussed in [6]. Autonomy implies that each entity is an independent context source. The mobility aspect indicates that entities can appear and disappear frequently from the network. The distribution aspect implies that context data should be retrieved from different entities whereas the heterogeneity aspect results in no common data model. Also, pervasive applications need to access context information about users and their devices without dealing with the details of the data collection process [4]. Additionally, context queries should be expressed in a context model that can be converted into different data models as required.

Based on these considerations, we can list out some of the requirements in designing CQL for context-aware systems as following:

- The CQL should express context queries in an abstract level without indicating details of context sources such as locations and storage mechanism.
- The CQL should express context queries utilizing a predefined context data model and this model should be well integrated with the underlying context data representation
- The CQL should express queries acquiring context data either in a pull-based or push-based manner by specifying filters and conditions.
- The CQL should express certain advance context processing operations to filter and process context data, such as aggregating and reasoning operations.
- The CQL should express context queries being able to access context data from more than one context source as well as continuous data.
- The CQL should be able to express compound queries concerning multiple contexts domains and conditions.

III. EXISTING CONTEXT QUERY LANGUAGES

Different types of CQLs leveraging on different technologies have been surveyed and classified previously: SQL-based, RDF-based, XML-based, API-based and Graph-based CQLs [3][6][9]. The evaluation conducted in [6] has further demonstrated that SQL-based and RDF-based CQLs are more powerful and effective than others. In this section, we will review existing SQL-based CQLs and compare with RDF-based CQLs to illustrate the necessity for a new SQL-based CQL.

A. SQL-based CQL

SQL is a well-established, declarative query language that has been recognized as an effective language for accessing traditional database. As demonstrated in [6], SQL-based CQL is one of the two most effective types of query languages in context-aware computing. Various SQL-based CQLs have been proposed and designed previously and we are inspired by these works to design our new SQL-based CQL.

Contory [18] utilizes a SQL-based CQL as an interface to provide context information for applications. Contory supports both pull- and push- based methods to acquire context data, but it does not support context processing operations and has poor performance on consolidating context information from different sources. Another SQL-based query language designed to query data in an ambient

 ISBN: 978-1-61208-364-3

intelligent environment is presented by Feng [2], which uses context information to define data retrieval conditions in a relational database. This work presents some good concepts and definitions on acquiring data through SQL-based mechanism but it is limited on relational databases and has less support to context data management.

PerLa [19] is a SQL-like language designed for collecting data from different nodes in a pervasive system. PerLa can acquire data from different sources independently from the underlying structure by dividing queries into two levels. However, PerLa does not adequately consider context data processing but limits sensor data. CML [12] queries context data directly through SQL by converting its context data model which can represent different type of context data to relational database schema through its internal designed facility. However, this conversion may generate complex SQL join operations and query representations are not programming and platform independent.

Another SQL-based context querying mechanism is described by Judd and Steenkiste [8]. By defining four types of entities (i.e., device, access point, people and space), this framework utilizes a Context Service Interface-SQL (CSI-SQL) wrapper to query and access data from context sources, but the primary function of this design is on expressing attribute requirements, timely execution and meta-attribution rather than context data. Fjords architecture [10] designs some SQL-based facilities to manage and query over streaming sensor data with supporting both pull- and push-based data acquisition methods. The mechanism presented by Madden et al. [11] can also acquire sensor data either through pull queries or push queries as well as a hybrid queries.

By studying the existing SQL-based CQLs, we observe that no one has fully fulfilled the requirements we illustrated in Section II. Motivated by this, we propose and design this SQL-based CQL leveraging on these previous works.

B. *Non-SQL based CQL*

RDF-based CQL is the other most powerful CQL in context-aware computing as shown in [6]. A typical RDF-based CQL is the MUSIC CQL proposed by Reichle et al. [17] which has a good support on querying derived information. However, supporting only single entity severely limits its querying capability. Another RDF-based querying mechanism described by Perich et al. [14] decouples queries into sub-queries and process them separately to overcome some obstacles posed by mobile environment. However, this reliance on mobile devices creates availability and reliability issues. SOCAM [5] can easily provide context information about entities and relationships between entities in the smart space leveraging on the ontology technology, but it has limitations in supporting complex queries.

Although RDF-based CQLs are more suitable for expressing relationships of context sources and are as powerful as SQL-based CQLs, they are not as flexible as SQL-based CQLs and produce large amounts of overhead in creating classes, sub-classes and relationships [14]. Furthermore, a large percentage of existing systems are using SQL-based methods to manage and acquire data, which are not easy to integrate with RDF-based approach. As a result, we plan to devise this SQL-based language so that the learning curve is not steep and integration is faster.

IV. PROPOSED CONTEXT QUERY LANGUAGE

Motivated by the need of a CQL that can fulfill all the requirements discussed previously, we propose a new SQL-based CQL to express context queries. As the language is designed leveraging on SQL syntax, the query structure consists of the existing SQL constructs like SELECT… FROM… WHERE with GROUP BY…HAVING…ORDER BY…as optional instructions. Most of the existing CQLs just extend SQL with basic SELECT…FROM…WHERE structure without mentioning other optional instructions to support querying on context data. In this proposed query language, we add some constructs to the existing SQL query structure to support them. These constructs include MODE, SUBSCRIBE/UNSUBSCRIBE, ON VALUE…LIFETIME which will be discussed in detail and which syntax is given in this section. The underlying context model leverages on the concept of context domain to divide and manage different categories of context sources, and each context domain is represented by a list of context attributes to represent context data, just as shown by by Pung et al. [16]. The basic structure of context query is as follows with constructs that are enclosed by […] being optional:

MODE GLOBAL | LOCAL
SELECT | SUBSCRIBE | UNSUBSCRIBE
<attribute> | <contextEvent> | <operation (attribute)> | <operation (contextEvent)> [, attribute, …]
FROM <domain> [, domain, …]
[ON INTERVAL <interval> **LIFETIME** <timespan>**]**
WHERE <predicate> [**AND** | **OR** predicate …]
[GROUP BY <attribute**>]**
[HAVING <predicate**>]**
[ORDER BY <attribute**>] [DESC | ASC]**

A. *Description*

The **MODE** clause specifies the mode of the query which is a new construct added specially for this CQL. Nowadays, with the advances of ubiquitous computing, data about various kinds of environmental conditions and other entities can be obtained and utilized by applications. On the other hand, with the advances of sensing technology on mobile devices, many applications which only utilize context information of a host device have also been designed and implemented. We believe that a properly designed CQL should be able to support these two types of applications effectively. By analyzing the characteristics, we can see that processing context queries of local applications should be different from global applications. In order to address this issue, we divide context queries into two types and use the MODE construct in the proposed CQL with possible values LOCAL and GLOBAL respectively. LOCAL represents queries of applications which utilize context information

 ISBN: 978-1-61208-364-3

from their host devices only; while GLOBAL indicates queries of applications which demand context data from different context sources. As a result, this proposed CQL supports both application types by supporting their corresponding data acquisition conditions.

The **SELECT/SUBSCRIBE/UNSUBSCRIBE** clause specifies the required context information or actions. The SELECT construct is inherited directly from SQL and represents queries that acquiring context data in a pull-based manner. However, since context-aware applications aim to detect changes of situations and adapt to them automatically, there are many queries concerning the changes of context information. We call these context changes as context events which will be pushed to applications as notifications when they are triggered. Unfortunately, the pull-based SELECT construct cannot support those push-based kind of queries. In order to address this issue, we design this new **SUBSCRIBE** construct which lets the application developers issue push-based queries. We also design the **UNSUBSCRIBE** construct to let applications cease their reception of notifications about certain kinds of context events.

In addition, we extend the **SELECT/SUBSCRIBE /UNSUBSCRIBE** clause with three different types of expressions to indicate what type of context information is required by the query, namely: context attribute, context event, and operation involved context. *Context attribute* is inherited from SQL but extended to indicate context information of a specific context attribute of a specific context domain. In the case of traditional database, data is integrated, these attributes can also be relational attributes.

Besides the simple context attribute, we also define two new types of context information: context event and operations involving context. The *context event* represents context information about changes of context source status and triggering notifications. As discussed previously, context event is an important type of context information for context-aware applications. Autonomy of context-aware applications is realized through situation detection and event notification mechanisms. By tracking the changes of situational status, those context events can trigger corresponding notification mechanism to make applications adapt to new situation automatically. It takes the format like "*domain. contextEvent*" which includes three sections: domain name, delimiter(.) and event name. The same event name may appear in different context domains, so the context domain information is shown to solve the possible ambiguity.

The other newly defined type of context information is the *operation involved context* which indicates context information derived by applying certain context processing operations on raw context data. Unlike traditional database systems which mainly focus on data updating and retrieval, context-aware systems also need to interpret or derive higher level context information during the query answering process. In order to generate the higher level context information, appropriate functions should be applied on collected raw context data. Even though traditional database systems also provide certain aggregating functions, they are not powerful enough to handle other contextual based operations. In order to solve this problem, we propose this new type of context information to represent those operation involved contexts. The expression consists of two parts: *operation* and *context data*. Operation indicates what context processing operations are utilized, while context data specifies raw context data required for the operations. This context data can be either context attribute or context event as illustrated above. There are several types of operations that can be applied on context data and we provide a separate section to give more details in next sub section.

```
<context> ::= <context attribute> |
   <context event>  | <operational
   context>
<operational context> ::= <operations>
   (<context attribute> | <context
   event>)
<context attribute> ::= <context
   domain>.< attribute>
<context event> ::= <context domain>.<
   event>
<context domain> ::= PERSON |OFFICE |
   HOME | SHOP | CLINIC | CAR ...
<attribute> ::= name | location |
   temperature | mood | activity |
   humidity | luminanicity | ...
<event> ::= locationChange | moodIsSad
   | temperatureIsHigh | ...
<operation> ::= <aggregating function>
   | <algebraic function> |
   <contextual function>
```

The **FROM** clause is also inherited directly from SQL, but we extend it to specify the context domains involved in the context queries. Those context domains are predefined by context-aware systems. In the context of Coalition [16], which organizes context sources into different domains, the values can be PERSON, OFFICE, HOME, SHOP, CLINIC, etc. Additionally, these context domains can be easily extended to relations of traditional databases. As a result, traditional database can be easily integrated with context data to produce higher order information. In the future implementation, it can also be extended to include semantic web sources or other Internet sources.

```
<context domain> ::= PERSON |OFFICE |
     HOME | SHOP | CLINIC | CAR |...
```

The **ON INTERVAL … LIFETIME** clause is a new construct designed for supporting context queries about continuous or periodic context acquisition. Just as discussed in Section II, context can be continuous data or periodic data. There are many context-aware applications that need to acquire context data continuously. One typical example is monitoring the security status of a critical place. Additionally, periodic acquisition of context information is also common for context-aware applications and a typical example is to monitor the patient status in hospitals. However, traditional database systems utilizing SQL which usually focus on handling sporadic data retrieval cannot

 ISBN: 978-1-61208-364-3

represent and handle these continuous or periodic cases appropriately. In order to solve this problem, we design this new construct and augment it to the proposed CQL.

The **ON INTERVAL** clause of the construct specifies the sampling interval for a context query. In the design, we treat continuous context retrieval as a special case of periodic context retrieval with the interval as zero. In order to differentiate this special case, we use a special value NULL to represent the continuous nature of retrieval. On the other hand, we use normal integer value plus time unit (i.e., ms, s, min, h) to indicate periodic context queries. Irrespective of whether the context retrieval is continuous or periodic, the usual pattern is that context information is retrieved over a period of time. In order to represent this issue, we design the **LIFETIME** clause of the construct to indicate the timespan of the context retrieval. There are two extreme cases of this timespan values. One is zero which implicitly indicates that the query is neither a continuous or periodic query. The other one is everlasting which means that the retrieval will never stop. Even these two cases are rare, but we include them for completeness and represent them with NULL and EVER respectively. A normal timespan value is represented by an integer value with a corresponding time unit.

The **WHERE** construct is inherited directly from SQL but extended to represent the list of constraints on context information acquisition. Constraints are the heart of a query since they provide a guideline on how to filter out unwanted context data from the large number of available context data sources. Most of existing SQL-based CQLs utilize simple predicates only, which is restrictive in expressing complex requests. In this CQL, compound predicates that consists of more context constraints connected by AND/OR are built.

```
<compound predicate> ::= <disjunctive
   predicate> AND <disjunctive predicate>
   AND ...
<disjunctive predicate> ::= <simple
   predicate> OR <simple predicate> OR ...
<simple predicate> ::= <context
   expression> <operator> <context
   expression>
<context expression> ::= <context
attribute> | <context event> | <context
constant> | <functional expression>
```

The remaining three constructs **GROUP BY, HAVING** and **ORDER BY** are all directly inherited from SQL, but extended with context variables or predicates as the arguments. Most of the previous SQL-based CQLs do not extend SQL with those constructs. However, we think these constructs are also necessary, especially in querying context data from a large number of context sources, in which case we may need these construct to further process the final results. The **GROUP BY** clause defines how the resulting context information can be further grouped with respect to specific context information. The **HAVING** clause defines how the filtered context data can be further selected with respect to certain conditions represented by having-predicates in the definition. The **ORDER BY** clause defines how the query results should be sorted with respect to the attribute indicated by context information either in ascending order (**ASC**) or descending order (**DESC**). These three constructs are all optional.

B. Context Processing Functions

Context processing functions represents various kinds of operations that can be applied on context data to generate or derive higher level context information. In addition to the updating and reading operations done by traditional database system, context query processing also takes charge of interpreting context data and deriving higher level information, which is just realized by those context processing functions. Some of the existing CQLs have similar processing operations integrated like [1] [14] [17].

SQL currently provides five aggregating functions that can be applied on a set of data to get some insights to the data patterns and behavior, namely: MAX, MIN, SUM, AVG, and COUNT. We think these aggregating functions are also necessary for context data query processing and inherited directly from SQL.

```
<aggregating function> ::= SUM |
   COUNT | AVG | MIN | MAX.
```

Another important type of functions added for the CQL is *contextual functions* which provide the task of context information interpretation and higher level information derivations from basic context data. Traditional database systems usually do not provide any functions for data interpretation. However, this type of functionalities is very important for CQL, so an important aspect of the proposed CQL is to support contextual functions that can be used by applications for information interpretation and reasoning. It is important to provide a suite of different types of such functions to support a wide variety of applications. One set of such contextual functions can interpret the basic sensor data and draw inferences based on the data and certain predefined logical conditions. For instance, isFever(temperature) interprets whether a given temperature corresponds to a fever or not. Similarly, an isFire(temperature) function can be used to determine a fire in a building and generate appropriate functions. Another class of contextual functions provides the task of deriving relations between two or more entities. For instance, isFriend(personA, personB) checks whether two persons are friends. This function can utilize the social network and contact information of a person to make this decision. Another function nearBy(x, y) can compute whether two entities are within a certain distance from each other. There are also contextual functions aiming to derive situational information. For instance, isMeeting(location) will determine whether there is a meeting going on in the given place. These are just some of the examples of the different types of contextual functions that we propose to support as part of our context query language.

```
<contextual function> ::= isFever(t) |
 distance(a, b) | isFriend(a, b) |
```

 ISBN: 978-1-61208-364-3

```
nearby(x, y) | isFire(a) |
isMeeting(l) | ...
```

Since these contextual functions are usually situation and application dependent, it is not possible to predefine and generate an exhaustive list of all the contextual functions. Instead, the application developers can define functions according to their application requirements, which requires the CQL to be extensible with new contextual functions. In order to realize this extensibility, we propose the creation of a contextual function repository that can hold the different types of contextual functions as defined by application developers. Additionally, this approach also promotes reusability as the popular contextual functions can be shared among applications. During the query processing, whenever a contextual function is detected, the query processor can retrieve the function definition and related rules from this repository to analyze the collected data and generate result accordingly.

V. EVALUATION

We implement this proposed CQL with our context-aware system Coalition [16] and examine the querying processing performance. In this section, we focus on illustrating how the proposed CQL can represent different types of requests and fulfill those requirements presented in section II and illustrate the proposed context query language usage with a study case.

A. Discussions

The proposed CQL expresses context domains, context attributes and conditions in a conceptual level. In other words, the expressions focuses on expressing what a user wants and does not mention any details about how context sources and context data are managed in the underlying framework or system. This means the proposed CQL express context should be able to express context queries at an abstract level.

Also, the proposed CQL utilizes a generic and conceptual method to model context data wherein context sources are divided into different context domains and each domain is associated with a list of context attributes. This mechanism is consistent with the concepts of relation and attribute in traditional relational database.

The proposed CQL supports two types of queries to support context acquisition: select-based queries and subscribe-based queries, which correspond to pull-based or push-based information retrieval methods respectively. Select-based queries specify the required context information with a list of conditions to filter out unwanted context data. This type of queries is issued on demand and gets context data in real time. On the other hand, subscribe-based queries issue required context information with a list of constraints proactively and the context data will be pushed to the query issuer whenever the constraints are triggered.

Additionally, the proposed CQL defines different functions to process context data to generate higher level context information. The aggregating functions can provide some pre-processing operations on the context data, while contextual functions can apply some predefined rules on context data to derive situational information.

The design of ON INTERVAL with LIFETIME section enables the proposed CQL to design context queries for continuous or periodic data retrieval. Additionally, the complex structure design of the WHERE clause enables the proposed CQL to express complex constraints and requirements. Together with specifying different domains in the FROM clause, the proposed CQL can represent compound queries that retrieve and process context data from different domains and context sources.

Another advantage of this proposed CQL is that it can be integrated with a traditional relational database system. As the proposed CQL is actually an extension of SQL, it can easily involve relational data by replacing context domains with relations. At the abstract level, application developers actually have no idea whether the data comes from context sources or relational database. As a result, context data and relational data can be seamlessly integrated. From the above illustration and analysis, we can see that proposed CQL has fulfilled the requirements illustrated in section II.

B. Case Study

In this sub section, we are going to illustrate the usage of this proposed context query language with a shopping guide scenario.

Lisa went shopping in shopping mall X during last weekend. At the moment of entering the mall, Lisa received a message about new arrivals from one clothes shop A as she subscribed the notification service previously. After going around at shop A, Lisa received a recommendation for the bookstore shop B based on her preference of book and current location. During the time in shop B, Lisa realized her friend Emily was also shopping in the same mall, and then she contacted Emily for a coffee. By checking the real time queuing information of the several coffee shops, they chose shop C, after which they looked around together and had a good shopping time.

In this scenario, different types of context information need to be queried. First, in order to push notification of new arrivals, the context event of Lisa's entering the shopping mall X should be subscribed in advance. Secondly, in order to give proper recommendations to Lisa, we need to find shops based on both her hobby and location context. Thirdly, in order to check whether there are friends around, we need to provide the function of find nearby friends based on current location. Last, in order to provide the real time queuing information, we need to monitor the queues of each shop. Following four context queries are designed for these four points respectively:

Q1: (***Context Event Subscription***)

```
MODE        GLOBAL
SUBSCRIB    person.enterMall("Mall X")
FROM        person
WHERE       person.name = "Lisa"
ON INTERVAL 0
LIFETIME    EVER
```

Q2: (***Search Matched Shops)***

```
MODE      GLOBAL
SELECT    shop.name
FROM      shop, person
WHERE     shop.type = person.preference
    AND   isNearby(shop.location, person.location)
    AND   person.name = "Lisa"
```

Q3: (***Check Nearby Friends***)

```
MODE      GLOBAL
SELECT    person.name
FROM      person
WHERE     isNearby(person.location, "Mall X")
    AND   isFriend(person.name, "Lisa")
```

Q4: (***Query Queuing Information***)

```
MODE      GLOBAL
SELECT    shop.name, MIN(shop.queue)
FROM      shop
WHERE     shop.type = "coffee"
    AND   person.location = "Mall X"
```

VI. CONCLUSION AND FUTURE WORK

We presented a new SQL-based CQL in this paper. By exploring the properties of context data and pervasive environments, we illustrated the requirements for a well-designed CQL. Through studying existing CQLs, we observed that SQL-based CQLs are more flexible and easy to utilize, inspired by which, a new SQL-based CQL is proposed and designed. This CQL supports both pull- and push-based queries as well as continuous context retrieval by different time intervals. Additionally, different context processing functions, especially contextual functions, have been designed to generate higher-level context information. Furthermore, this CQL supports compound conditions to get context data from various context entities belonging to different context domains. Leveraging on the proposed SQL-based CQL, user requests can be better represented and supported by the underlying context data management system. One of the main future works is to further integrate context reasoning operations with this proposed CQL in which reasoning operations can be represented by certain intermediate context processing operations, so that referenced context information can be generated in runtime during context query processing.

ACKNOWLEDGMENT

This research was carried out at the SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

REFERENCES

[1] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications". Human-Computer Interaction, 2001, pp. 97-166.

[2] L. Feng, "Supporting context-aware database querying in an ambient intelligent environment". In: 3rd IEEE International Conference on Ubi-media Computing (U-Media), July 2010, pp. 161-166.

[3] L. Feng, J. Deng, Z. Song, and W. Xue, "A logic based context query language". Smart Sensing and Context, 2010, pp. 122-134.

[4] C. Fra, M. Valla, and N. Paspallis, (2011). "High level context query processing: an experience report". In: IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), March 2011, pp. 421-426.

[5] Gu, T., Pung, H. K., and Zhang, D. Q. A service-oriented middleware for building context-aware services. Journal of Network and Computer Applications, 2005, pp. 1-18.

[6] P. D. Haghighi, A. Zaslavsky, and S. Krishnaswamy, "An evaluation of query languages for context-aware computing". In: 17th International Conference on Database and Expert Systems Applications (DEXA), 2006, pp. 455-462.

[7] J. Heer, A. Newberger, C. Beckmann, and J. I. Hong, "Liquid: context-aware distributed queries". In: Ubiquitous Computing (UbiComp 2003), October 2003, pp. 140-148.

[8] G. Judd and P. Steenkiste, "Providing contextual information to pervasive computing applications". In: First IEEE International Conference on Pervasive Computing and Communications (PerCom 2003), March 2003, pp. 133-142.

[9] Y. Li, L. Feng, and L. Zhou, "Context-aware database querying: recent progress and challenges". The Book Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability, 2009, pp. 147-168.

[10] S. Madden and M. J. Franklin, "Fjording the stream: an architecture for queries over streaming sensor data". In: 18th International Conference on Data Engineering, March 2002, pp. 555-566.

[11] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "TinyDB: an acquisitional query processing system for sensor networks". ACM Transactions on Database Systems (TODS), 2005, pp. 122-173.

[12] T. McFadden, T., Henricksen, K., and Indulska, J. "Automating context-aware application development". In: UbiComp 1st International Workshop on Advanced Context Modelling, Reasoning and Management, Nottingham, September 2004, pp. 90-95.

[13] F. Perich, A. Joshi, T. Finin, and Y. Yesha, "On data management in pervasive computing environments". IEEE Transactions on Knowledge and Data Engineering, 2004, pp. 621-634.

[14] F. Perich, A. Joshi, Y. Yesha, and T. Finin, "Collaborative joins in a pervasive computing environment". The International Journal on Very Large Data Bases, 2005, pp. 182-196.

[15] M. Perttunen, J. Riekki, and O. Lassila, "Context representation and reasoning in pervasive computing: a review". International Journal of Multimedia and Ubiquitous Engineering, 2009, pp. 1-28.

[16] H. K. Pung, et al. "Context-aware middleware, for pervasive elderly homecare". IEEE Journal on Selected Areas in Communications, 2009, pp. 510-524.

[17] R. Reichle, et al. "A context query language for pervasive computing environments". In: 6th Annual IEEE International Conference on Pervasive Computing and Communications. March 2008, pp. 434-440.

[18] O. Riva anD C. di Flora, "Contory: a smart phone middleware supporting multiple context provisioning strategies". In: 26th IEEE International Conference on Distributed Computing Systems Workshops, July 2006, pp. 4-7.

[19] F. A. Schreiber, R. Camplani, M. Fortunato, M. Marelli, and G. Rota, "Perla: A language and middleware architecture for data management and integration in pervasive information systems". IEEE Transactions on Software Engineering, 2012, pp. 478-496.

 ISBN: 978-1-61208-364-3

A Mobile Learning Framework on Cloud Computing Platforms

Wei Guo, Joan Lu
School of Computing and Engineering
University of Huddersfield
Huddersfield, West Yorkshire, UK
Emails: {wei.guo, j.lu@}hud.ac.uk

***Abstract*—Cloud computing infrastructure is increasingly used for distributed applications. Mobile learning applications deployed in the cloud are a new research direction. The applications require specific development approaches for effective and reliable communication. This paper proposes an interdisciplinary approach for design and development of mobile applications in the cloud. The approach includes front service toolkit and backend service toolkit. The front service toolkit packages data and sends them to a backend deployed in a cloud computing platform. The backend service toolkit manages rules and workflow of web services, supports fault tolerance and then transmits required results to the front service toolkit. To further show feasibility of the approach, the paper introduces a case study and shows its performance.**

Keywords-cloud computing; mobile devices; service set; big data.

I. INTRODUCTION

The increasing demands for efficient resources utilisation result in the use of cloud computing. Armbrust et al. have proposed that the adoption of virtualised resources brings high scalability and availability to applications deployed in clouds [1]. Cloud computing can offer scalable storage and resources expansibility [2][3][23][24][25].

A great number of relative mobile application techniques integrated in clouds have been proposed [4][5][6][7][16]. However, they only focus on certain aspects of design and implementation of mobile applications in cloud computing platforms. Currently, few publications have taken cloud-based system-level framework and generic service framework into consideration. Therefore, this paper proposes a cloud-based system-level framework including generic front service and backend service frameworks.

In this paper, a new approach for mobile devices deployed in the cloud is advocated, which consists of two ends. 1) One is the front service toolkit, used mainly for receiving messages from end users and sending processing messages to the backend service toolkit. 2) The other one is the backend service toolkit, which is responsible for executing business data flow and rules.

This paper is organised as follows. Section 2 describes related researches. Section 3 proposes frameworks of the front and backend service toolkits. Section 4 outlines a case study implementing the above frameworks. Section 5 presents experiments of the case in Section 4 and discusses the results. Section 6 sketches the conclusions.

II. BACKGROUND

Mobile application techniques in the cloud are proposed [4][5][6][7]. Designing and development of cloud based M-Learning tools are introduced by Butoi et al. [16]. Pocatilu proposes a framework for syncing mobile applications with cloud storage services [6]. However, this has not considered security between mobile devices and the cloud computing platform. Gu et al have designed services and components for transmitting files from mobile devices to the cloud to trade-off between performance and battery life [10]. Lee and Park have introduced system layer, application layer and user layer for a mobile cloud learning system [11]. A rendering adaptation technique is proposed to enable multimedia applications on rich mobile devices [12]. The utilisation of a component-based approach for the framework results in shading the implementation details of sophisticated functionalities when it is running [13][14][15].

Meng and Lu have proposed the implementation of the emerging mobile technologies in facilitating a mobile exam system [17]. They have proposed opportunities for interactive learning systems with evolutions in mobile devices [18]. Integration of smartphone's intelligent techniques for authentication in a mobile exam login process is introduced [19].

Liu has proposed that big data drives cloud adoption in enterprise [20]. Assuncao et al. have discussed approaches and environments for perform analytics in the cloud for big data applications [22]. Bahrami has proposed a cloud template approach as a big data solution [21]. However, the granularity of the cloud template is not suitable for mobile applications.

The relative researches mainly focus on usage of computing and storage ability of cloud or mobile communication. In this paper, the frameworks of front and back-end toolkits are at the system-level perspective and built based on modules.

The reasons for a new design are:

- A system-level framework is proposed for mobile applications with the cloud so as to improve software availability and expansibility.
- Generic front service and backend service frameworks are to reduce application development costs.
- The generic front and backend service framework make applications in cloud easy to maintain.

 ISBN: 978-1-61208-364-3

Figure 1. The framework of the front service toolkit.

III. FRAMEWORKS

In this section, the frameworks of the front service toolkit and backend service toolkit are proposed. The front service toolkit and backend service toolkit can work individually and only performs the cooperation with each other when the user requirements. Cloud computing, as a new concept for hosting and managing different services, is an environment solution for big data. It can eliminate the requirements of provisioning of users and allows owners of both small and enterprises to increase resources only when there is a rise in service demand.

A. *Mobile Applications in the Cloud Computing Platform*

In this section, the overall architecture of mobile applications in the cloud is described.

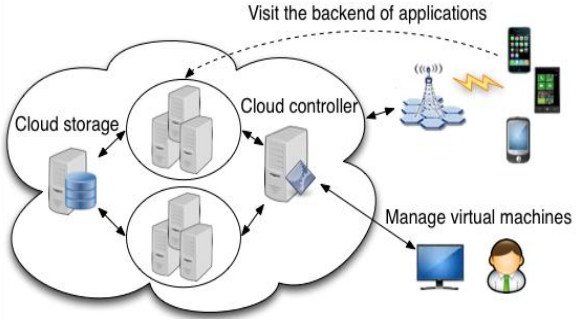


Figure 2. The overall architecture of mobile applications in the cloud.

Fig. 2 shows the overall architecture of mobile applications on a cloud computing platform. The cloud computing platform has a cloud controller, which is applied to manage virtual machines and monitor the states of physical hosts and virtual machine instances. The cloud storage, as shown in Fig. 1, stores all the information of end users and their backend of applications. It can be implemented by distributed storage framework. Administrators of cloud computing infrastructure can manage the overall cloud computing platform. A mobile application consists of the front service toolkit and back-end service toolkit, as shown in Fig. 2. These toolkits are described in the following sections.

B. *The Framework of Front Service Toolkit*

In this section, the framework of the front service toolkit is proposed. Fig. 1 is a framework and will be discussed in this section.

Packager: All of the packagers are utilised for packaging information and data from users and then the packaged messages are transmitted to backend services in the cloud. In particular, the XML packager has the responsibility for organizing data according to XML schema defined in advance.

Analyser: Analysers consists of variable analyser, array analyser and XML analyser. They are corresponding to the above packagers separately, which are applied to extract information and data from packages routed from the backend service toolkit. As the interface of receiving messages, they are required to be identical to the packagers.

Applications need to compress and decompress the messages from users with various formats to the backend side if they require communication with database to record users' information, business flow and logging messages. Therefore, the compressing and decompressing functions are packaged as packager and analyser modules.

C. *The Framework of the Backend Service Toolkit*

The backend service toolkit is the connection between the front side of applications and database. It is responsible for convert raw messages from clients into specified data format designed by developers and database designers. In this section, the framework of the backend service toolkit is proposed. The framework is mainly composed of analyser and packager, business services pool, and Infrastructure as a Service (IaaS), as shown in Fig. 3. Note that the designed framework in this paper includes business process services pool possessing rule controlling services and process controlling services. Rules can be specified and designed by developers. Based on the application business detailed services, process controlling services control and manage the overall the business flow and message flow of mobile applications.

Analyser and packager: Analysers and packagers of the backend service toolkit have the same main functions as the front service toolkit. The front service toolkit uses these functions to trigger the transmission of data and messages to backend.

Application business detailed service pool: It is used for containing and deploying components and services of applications. Note that mobile applications are divided into functions as web services in the cloud. The cloud computing platform publishes web services through a Web server. Web services implementing application business have hierarchical relationships. Web services in this layer can invoke each other and, due to specific user requirements, can work together to complete application functions.

Business process services pool: This layer is a web service container which is integrated with the rule engine and workflow engine. They offer unique standard web service interfaces to external systems and mobile devices. Workflow engine organise the relationships among web services and defines main work flow and message flow. Therefore, the rule engine can deal with decisions of work flow in terms of application requirements. In particular, when there is more than one work flow in applications, the rule engine is considered to handle the flow of information and messages.

 ISBN: 978-1-61208-364-3

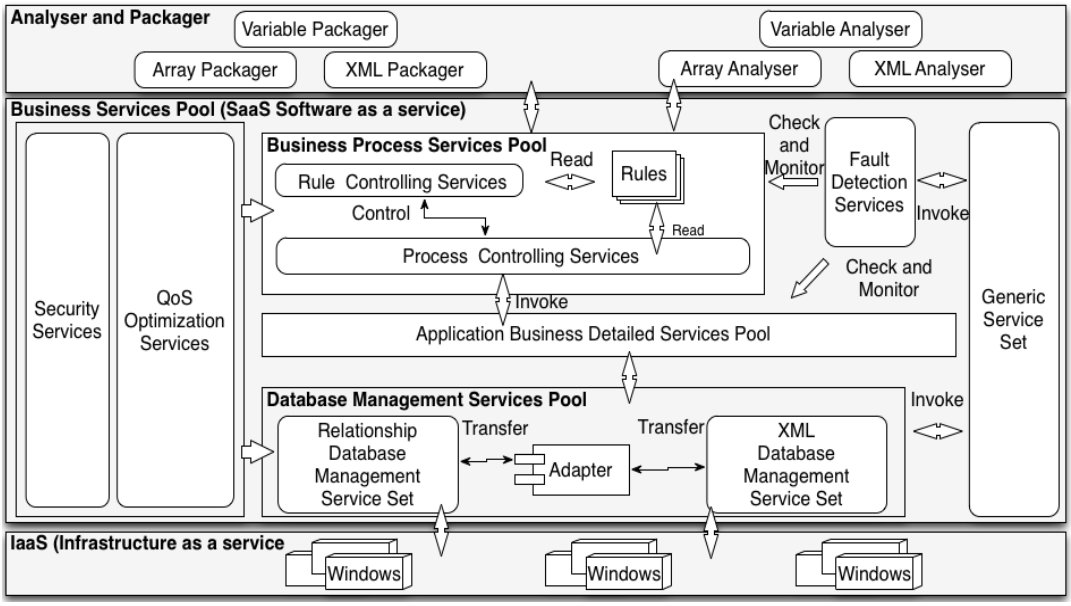


Figure 3. The framework of the backend service toolkit.

Database management services pool: This applies mapping techniques to map objects from the above layers into records in database. Here, two categories, namely, relationship database and XML database, are utilised. The backend program makes use of database management service sets to store data into databases and to retrieve record from tables. The backend service toolkit employs multiply databases to store data so as to restore storage if the main database crashes.

Fault detection services: This has the responsibility for detection and fault tolerance of faults of the front and backend service toolkits. The backend service toolkit includes service backup functions, which can back up central services and failure-prone services.

Generic service set: Common and generic services are abstracted and placed in this layer. This layer provides services to developers and other service sets. Developers can exploit libraries of services in this set to develop other applications.

Security services and QoS optimisation services: Developers and end users can define and put their security services into this module. The security module and QoS optimsation module work together to confirm that applications run and meet user requirements.

D. Fault Tolerance for Backend Service Toolkit

In this section, a fault tolerance scheme for the backend service toolkit is proposed. The fault tolerance level of services of backend toolkit is defined and specified by developers to perform the fault-tolerant execution.

1) Fault Tolerance Parameters Packaging

The front service toolkit can add a fault-tolerant identifier into the metadata of request messages.

Fault tolerance parameters definition: The front service sends the request sequence, $Request = \langle service\_id, R,\ \mathrm{P}\rangle$, to the backend service toolkit. Here, $R = \{c, n, s, \mathrm{l}\}$ is a fault-tolerant parameters set, where c, n, s and l denote computing, network, storage resources and fault tolerance level, respectively. For example, $Request = \langle service\_i, \{c_i, n_i, s_i, l_i\}\rangle$ represents front service $service\_i$ which needs c_i computing, n_i network and s_i storage resources, respectively. P is a request parameters set, which is a business request from specified functions. For a mobile application, whose backend is based on a cloud platform, the whole created fault tolerance metadata is $\mathrm{Request} = \langle \mathrm{app\_services}, \{\sum \mathrm{c}, \sum \mathrm{n}, \sum \mathrm{s}, \mathrm{I}\}\rangle$, where $\sum \mathrm{c}$, $\sum \mathrm{n}$, $\sum \mathrm{s}$ and I denote the total computing, network, storage resources and fault tolerance level sets, respectively.

2) Replica Virtual Machine Selection

The fault detection service of the backend service toolkit has responsibility for selecting a replica virtual machine according to the received fault tolerance parameters from the front service toolkit. At first, the fault tolerant service needs to collect the domain knowledge of cloud infrastructure and to generate its network topology. Then, the fault detection service selects the replica placement nodes to back up the original services.

a) Cloud Network Topology Construction

Fig. 4 lists cloud fault tolerance components, which can collect computing, network and storage resources information from virtual machines.

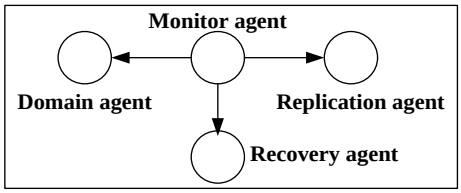


Figure 4. Cloud fault tolerance components.

 ISBN: 978-1-61208-364-3

TABLE I. CENTRAL SERVICES OF IPLAYCODE IN CLOUD

Index	Service Name	Function Descriptions
1	GameShuffleService	The web service is to disrupt the order of the question list.
2	GameMarkService	The service is used to mark students' responses.
3	GetQuestionsService	The service gets question contents and sends them to mobile devices.
4	GameCounterService	The service is to record information from user accessing mobile applications.
5	GameMathService	It includes generic and useful math libraries.
6	SetAnswersService	The service stores users' answers into database.
7	SetCommentsService	Users can submit comments on iPlayCode to help improve it.

Domain agent: The domain agent aims to execute as a domain knowledge collection management. Resources information collection management may handle constraints on selecting duplication placement nodes to create network topology. Computing resources, network capability and storage capability of virtual machine instances are collected and transformed into a resource metadata through the domain agent. Namely, $Dom = \langle vm\_id, R \rangle$, where vm_id is the instance identifier and R is a set of computing, network and storage resources. Here, $R = \{c, n, s\}$, where c, n and s denote computing, network and storage resources, respectively. For example, $\mathrm{Dom} = \langle \mathrm{vm\_i}, \{c_i, n_i, s_i\} \rangle$ represents the instance vm_i which has c_i computing, n_i network and s_i storage resources, respectively.

Monitor agent: The monitor agent is responsible for monitoring all its own agents in the cloud computing infrastructure (e.g., domain agent, replication agent and recovery agent) by means of a time-out mechanism.

Recovery agent: The recovery agent completes the goal of recovering the last checkpoint state from failures. When failures are detected, the monitor agent collects the abnormal failure node information from stable network storage and triggers the recovery mechanism.

Replication agent: The replication agent fulfills the target of replicating applications and files in virtual machines to the available calculated node with fault tolerance degree, in the presence of the client's requirements.

b) Matching Operation with the Requirements from Front Services

As mentioned earlier, the fault-tolerant and recovery component can receive the requirements from front services through the analyser and packager. Hence, the component allows front services to specify fault tolerance properties. For the fault tolerance level, it is based on the virtual machine level in this paper. It mainly depends on the computing, network and storage resources. Here the virtual machine with high computing, network and storage resources is considered as a node having higher fault-tolerant capabilities.

IV. A CASE STUDY

The section shows an application, called iPlayCode. iPlayCode is a mobile-based application and its backend applies web service techniques. All the operations of iPlayCode to database are packaged as web services.

Table I lists the central services of the backend service toolkit of iPlayCode and their function descriptions. In particular, the service, GameShuffleService, is used to disrupt the order of questions. To make students unable to remembering answers, iPlayCode needs to change the sequences of retrieved questions. In addition, the service, GameMathService, is a generic maths package. Note that, although user interfaces and functionalities of applications differ, their basic functionalities are identical. The generic service set of the backend service toolkit includes a basic development service pool. These services are packaged to offer the standard interfaces to developers.

The business flow of iPlayCode is as follows. End users login to iPlayCode typing their user name. The system would receive the physical address of physical devices and send both address and user name to the backend service toolkit. After backend services store this identity information into the database, they would send back the success message to the front applications on mobile devices. Then end users can choose a question level to select questions. After that, the system gives a response of question lists to mobile phones. After users finish questions, they can submit answers to backend services and services can store submitted answers.

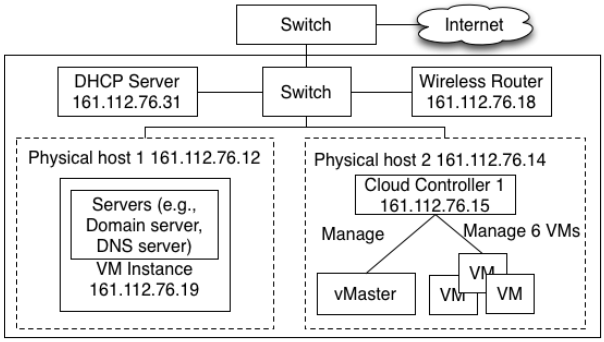


Figure 5. Network topology of cloud computing environment.

Fig. 5 introduces the network topology of the cloud computing environment which iPlayCode uses. From the graph, cloud controller 1 manages 6 virtual machines. These virtual machines deploy all the services of the backend service toolkit.

 ISBN: 978-1-61208-364-3

TABLE II PARAMETERS OF CLOUD ENVIRONMENTS

	Host A	Host B	Host C	Host D	Host E	Host F
IP	161.112.1.22	161.112.1.23	161.112.1.24	161.112.1.25	161.112.1.26	161.112.1.27
OS	Windows 2008 R2	Windows XP	Windows 7	Windows 7	Windows XP	Windows 7
CPU	Intel P-IV 2.0G	Intel P-IV 2.0G	Intel P-IV 2.0G	Intel P-IV 2.0G	Intel P-IV 2.0G	Intel P-IV 2.0G
RAM	2048 MB	1024 MB	1024 MB	1024 MB	1024 MB	1024 MB
HDD	350G	500G	500G	500G	500G	500G

TABLE III TEST RESULTS

Test Items \ Mobile Platforms		Apple iOS 7.0 (iPhone 5s)	Android 4.2 (SAMSUNG)
Interface	Information display	√	√
	Element position	√	√
	Layout of page	√	√
Functionality	Logining System	√	√
	Selecting question level	√	√
	Answering questions	√	√
	Automatic marking	√	√
	Storing marks	√	√
	Storing answers	√	√
Communication time	Logining couner	√	√
	Getting question speed	√	√
	Storing marks	√	√
	Storing answers	√	√

(Note: √ denotes correct or acceptable)

V. EXPERIMENT

The section shows actual deployment experiment of iPlayCode. First, conditions and system setup are proposed to describe the basic cloud infrastructure. Second, applications are introduced that are used in the experiment. Finally, data analysis is demonstrated.

A. Conditions and System Setup

Citrix XenServer 6.0.2 is employed as the cloud infrastructure. The virtual machines are listed in Table II. To deploy services of iPlayCode, hybrid operation systems and physical hosts are employed. All the implementation of iPlayCode is based on modular.

B. Testing of the System

1) Testing Strategy

The following list describes test mobile equipment, test cloud environment and test items. Experiment includes two parts. One is the mobile application. Here, iPhone, iPad and Android phones are used for the test. The other one is the cloud environment test. Citrix XenServer 6.02 as the cloud infrastructure is utilised to offer virtual computing, network and storage resources to backend services.

- Test mobile equipment: Apple iPhone, iPad and Android phones.
- Test cloud environment: Wi-Fi and Citrix XenServer 6.0.2.
- Test items: interface, functionality and communication time.

2) Test Results

Test results are shown in Table III. From Table III, two operation systems for mobile phones are used: Apple iOS 7.0 and Android 4.2. Test results mainly focus on interface, functionality and communication time.

C. Data of Experiment

Fig. 6 illustrates execution time of services of iPlaycode deployed on the web server and in the cloud separately. As shown from these figures, services, GameMarkService, GetQuestionsService and SetAsnwersService occupy the maximum running time.

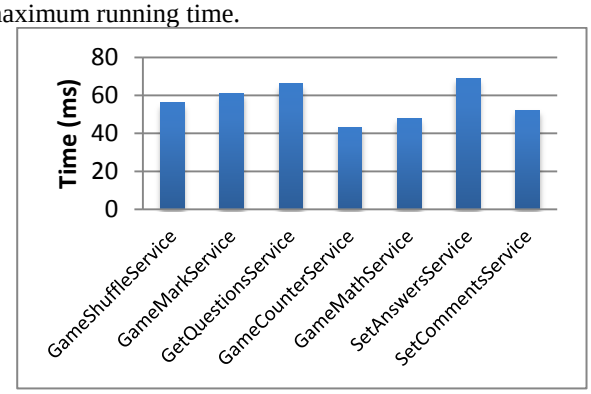


Figure 6. Execution time of services deployed on traditional Web server.

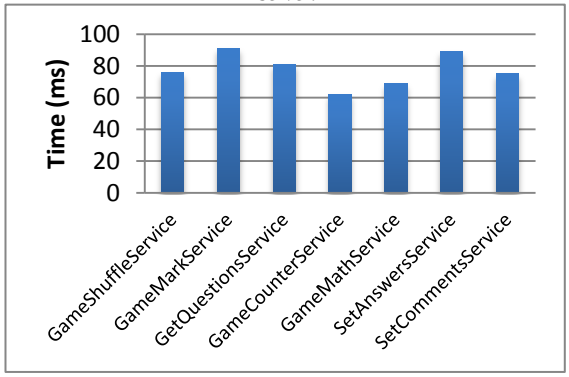


Figure 7. Execution time of services deployed in cloud.

 ISBN: 978-1-61208-364-3

Fig. 7 shows the execution time of services in the cloud, which is longer than on the web server. The deployment of the backend service toolkit is conducted on the cloud computing platform. Cloud computing platform not only offers web services deployed more security, but overhead of service communication. Web services of application, iPlayCode, inherit the interfaces of the backend service toolkit. Although web services need more time to respond to mobile devices, they can offer application availability, reliability and expansibility.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper proposed the design and development of mobile applications in the cloud. It includes front service toolkit and backend service toolkit. At the end of the paper, a case study, iPlayCode is proposed, which is deployed on Citrix XenServer 6.0.2. Execution time of services in the cloud is compared with that on a traditional web server. This investigation contributes to the following points:

- The system-level framework can improve availability and expansibility of mobile learning applications.
- The framework makes software easy to maintain.
- The toolkits reduce application development costs at the expense of some communication time.

B. Future Work

The future studies will include issues listed as follows:

- Enhance the functionality of the system according to user preference.
- Enhance analysis of the responses from students for learning performance evaluations.
- Shorten the response time between the front and the backend service toolkit.

REFERENCE

[1] M. Armbrust et al., "Above the clouds: A Berkeley view of cloud computing," EECS Dept., Univ. California, Berkeley, UCB/EECS-2009-28, Feb. 2009, pp. 1-25.

[2] Amazon Elastic Compute Cloud [retrieved: April, 2014]. Available: http://aws.amazon.com/ec2/.

[3] Eucalyptus Systems [retrieved: April, 2014]. Available: http://www.eucalyptus.com/.

[4] D. Minifie and Y. Coady, "Getting mobile with mobile devices: using the web to improve transit accessibility," W4A '09 Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A), Sept. 2009, pp. 123-126.

[5] W. Rochadel, J. P. S. Simão, J. B. D. Silva, and G. R. D, C, Alves, "Educational application of remote experimentation for mobile devices," Remote Engineering and Virtual Instrumentation (REV), 2013 10th International Conference on, Feb. 2013, pp. 1-6.

[6] P. Pocatilu, C. Boja, and C. Ciurea, "Syncing Mobile Applications with Cloud Storage Services," Informatica Economică, vol. 17, Feb. 2013, pp. 96-108.

[7] D. Popa, K. Boudaoud, M. Borda, and M. Cremene, "Mobile cloud applications and traceability," Networking in Education and Research, 2013 RoEduNet International Conference, 12th Edition, Sept. 2013, pp. 1-4.

[8] M. Armbrust et al., "A view of cloud computing," Communications of the ACM. 2010, pp. 50-58.

[9] R. Guerraoui and M. Yabandeh, "Independent faults in the cloud," in Proc. 4th Int. Workshop Large Scale Distributed Syst, Middleware, no. 6. 2010, pp. 12-17.

[10] Y. Gu, V. March, and B. Lee, "GMoCA: Green Mobile Cloud Applications," GREENS, IEEE, (2012), pp. 15-20.

[11] J. Lee and J. Park*, "Application for Mobile Cloud Learning," 16th International Conference on Network-Based Information Systems, Apr. 2013, pp. 296-299.

[12] S. Wang and S. Dey, "Adaptive Mobile Cloud Computing to Enable Rich Mobile Multimedia Applications," IEEE Transactions on Multimedia, vol. 15, Jun. 2013, pp. 870-883.

[13] M. Hiltunen and R. Schlichting, "An approach to constructing modular fault-tolerant protocols," in Proc. 12th Symp. Reliable Distributed Syst., 1993, pp. 105-114.

[14] G. Jung and K. M. Sim. "Agent-based adaptive resource allocation on the cloud computing environment," The 40th International Conference on Parallel Processing Workshops (ICPPW 2011). IEEE Computer Society, 2011, pp. 345 - 351.

[15] R. Jhawar, V. Piuri, and M. Santambrogio, "Fault tolerance management in cloud computing: a system-level perspective," IEEE. Systems Journal. Jul. 2013, pp. 288-297.

[16] A. Butoi, N. Tomai, and L. Mocean, "Cloud-Based Mobile Learning," Informatica Economică, vol. 17, Feb. 2013, pp. 27-41.

[17] Z. Meng and J. Lu, "Implementing the Emerging Mobile Technologies in Facilitating Mobile Exam System," 2nd International Conference on Networking and Information Technology, IPCSIT 25th-26th November 2011, 17 . IACSIT Press, Hong Kong, China, Nov. 2011, pp. 80-88. ISBN 978-981-07-0680-7.

[18] Z. Meng and J. Lu, "Opportunities of Interactive Learning Systems with Evolutions in Mobile Devices: A Case Study," In: Proceedings of the 2011 International Conference on Internet Computing ICOMP 2011. CSREA Press, 2011, pp. 238-244. ISBN 1601321864.

[19] Z. Meng, J. Lu, and A. Sawsaa, "Integrating Smartphone's Intelligent Techniques on Authentication in Mobile Exam Login Process," In: Computational Collective Intelligence. Technologies and Applications: 5th International Conference, ICCCI 2013, Craiova, Romania, Proceedings. Lecture Notes in Computer Science (8083). Springer, London, Sept. 2013, pp. 130-142. ISBN 978-3-642-40494-8.

[20] H. Liu, "Big Data Drives Cloud Adoption in Enterprise," Internet Computing, IEEE, July-Aug. 2013, pp. 68 – 71. ISSN 1089-7801.

[21] M. Bahrami, "Cloud Template, a Big Data Solution," International Journal of Soft Computing and Software Engineering, JSCSE, Aug. 2013, pp. 13-16. DOI 10.7321/jscse.v3.n2.2.

[22] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data Computing and Clouds: Challenges, Solutions, and Future Directions," Distributed, Parallel, and Cluster Computing, Dec. 2013, pp. 1-39.

[23] I. Kureshi et al., "Advancing Research Infrastructure Using OpenStack," International Journal of Advanced Computer Science and Applications, 2013, pp. 64-70.

[24] J. Zhao et al., "A security framework in G-Hadoop for big data computing across distributed Cloud data centres," Journal of Computer and System Sciences, 2014, pp. 1-14.

[25] Z. Xiao and Y. Xiao, "Achieving Accountable MapReduce in cloud computing," Future Generation Computer Systems, 2014, pp. 1-13.

 ISBN: 978-1-61208-364-3

Link Analysis among Sightseeing Spots based on Geo-Image Analysis

–Towards Majority-based Route Recommendation in Sightseeing–

Kohei Tashiro*, Atsushi Shimada†, Hajime Nagahara‡ and Rin-ichiro Taniguchi‡
*Graduate School of Information Science and Electrical Engineering Kyushu University Fukuoka, Japan
Email: tashiro@limu.ait.kyushu-u.ac.jp
†Faculty of Arts and Science Kyushu University Fukuoka, Japan
Email: atsushi@limu.ait.kyushu-u.ac.jp
‡Faculty of Information Science and Electrical Engineering Kyushu University Fukuoka, Japan
Email: nagahara, rin@limu.ait.kyushu-u.ac.jp

***Abstract*—In recent years, photo sharing sites such as Flickr, Picasa and others have become popular. These sites are open to public and many photographers upload their photos to share them with family, friends and people in the world. Each photo has several types of metadata including date, time, tags, geo-locations and others, which are automatically produced by a camera or manually provided by the owner. Researchers are interested in such a large-scale image database and use it for image analysis, image annotation, scene understanding and other purposes. Our research focuses on sightseeing images. Analysis of these images shows not only famous sightseeing spots but also links between several sightseeing spots, i.e., popular sightseeing routes. We extract such information through analysis of image metadata. Furthermore, we characterize the sightseeing spots by link type. We developed application software for smartphones in which recommendation information based on the above analyses are displayed. We performed several field tests of real scenes, finding that the recommendation information is useful to travelers.**

Keywords–user-generated contents; big data analysis; sightseeing spot; sightseeing links; characteristic analysis.

I. INTRODUCTION

There are many kinds of User-Generated Contents (UGCs) on the web, such as Twitter, Facebook, Picasa [1], Flickr [2], YouTube,and WiKi. UGCs provide an opportunity to share our daily activities with family, friends and/or people in the world. In recent years, many researchers have been attracted to UGCs. They wish to create new social value through analysis of the great number and types of people's activities.

Our research focused on the UGC Flickr, which has a large image database, for analyzing activities in sightseeing. Many people upload photos taken during sightseeing, on which they put text labels. Moreover, each photo has geo-location information regarding where the photo was taken. We used such " tagged images " for analysis of sightseeing activities. Among these images, our method finds sightseeing spots and link strength. Characteristics of each location are classified into four types. The analyzed information is very useful for route recommendation, navigation systems and others. We developed an application available to Android smartphones, and used it for field testing in the city of Nagasaki, Japan. In this paper, we present the potential of the large image database for application to the sightseeing recommendation system.

II. RELATED WORK

There are several research works regarding recommendation systems for sightseeing. Zheng, Zhang, Xie and Ma [3] estimated sightseeing routes from GPS logs. In addition, they determined sightseeing spots of interest to many people. The advantage of their approach is exact route estimation using GPS. However, existing GPS datasets are not as large as image datasets on the web, so there is difficulty in performing large-scale analysis.

Cao et al. [4] proposed a recommendation system for sightseeing spots at global scale. Their system receives a keyword query from a user. Then, some images related to the keyword are selected. Finally, the system recommends certain images as candidate sightseeing spots. Arase, Xie, Hara and Nishio [5] analyzed sightseeing behavior from pictures taken by tourists. These approaches focused on a relatively large area.

In contrast, our approach focuses on city or town areas to ascertain the behavior of tourists more precisely. We use photographer information, photo time and location information from the metadata. We also analyze characteristics of each sightseeing spot and relative strength (see explanation below) between spots, and assist route choices. We do not determine a sightseeing route solely to recommend it. Instead, we recommend consecutive sightseeing spots from the place where a tourist stays, by analyzing spots with strong connection to the location of that stay.

III. OVERVIEW OF THE PROPOSED IDEA

Fig. 1 shows an overview of the processing flow. First, a large number of geo-tagged images are collected from Flickr (Step 1). The images are divided into rough groups in terms of city or town levels. Second, geo-clusters are made from the images in each group (Step 2). We regard each geo-cluster as a sightseeing spot. Third, the number of people who moved from one spot to another is counted within all combinations of spots (Step 3). This number is used for calculating the strength of a link between two sightseeing spots. Furthermore, the sightseeing route of an individual is assessed by referring to the owner's ID and timestamp of each geo-tagged image. Finally, characteristics of each spot are investigated according to the strength of inflow and outflow links (Step 4).

 ISBN: 978-1-61208-364-3

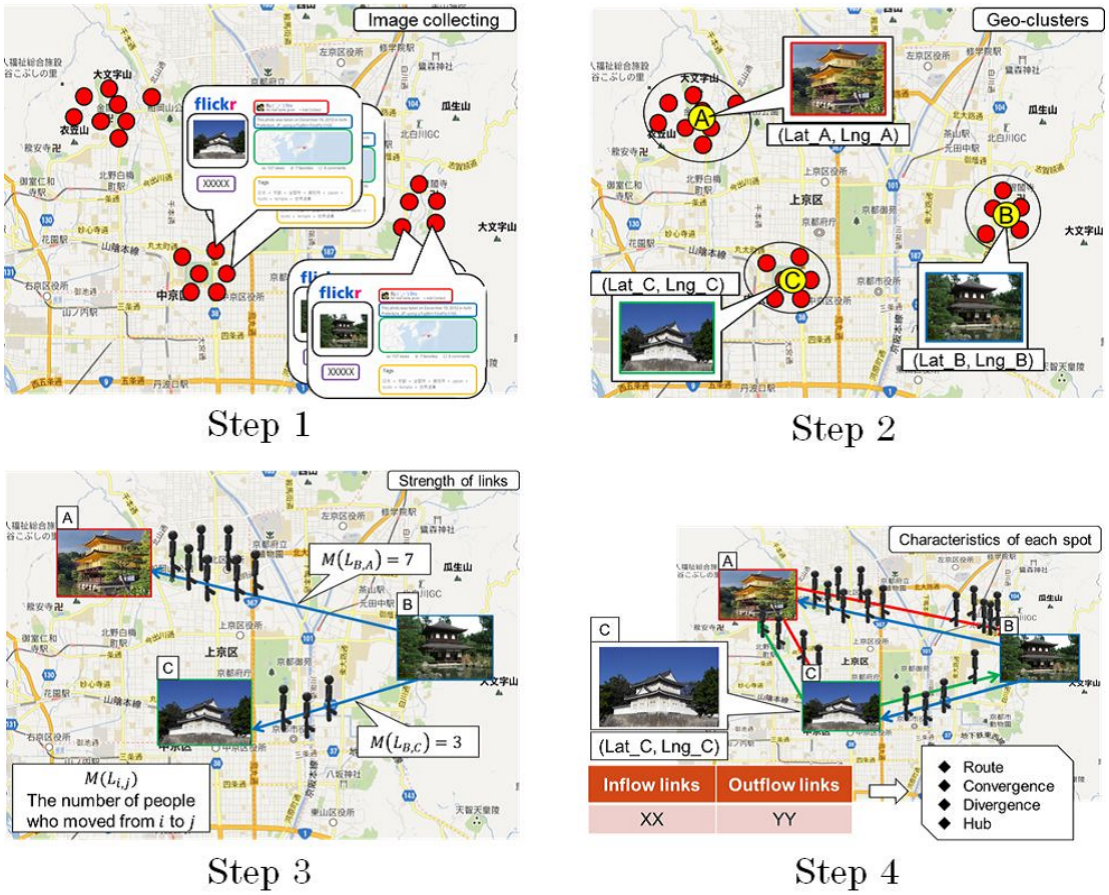


Figure 1: Flow of the proposed method

IV. Link and Characteristic Analysis

A. Dataset Property

Flickr provides not only a large image dataset but also attributes called "Exif". For example, representative attributes are owner ID, geolocation (latitude and longitude), timestamp, and others. These items were the most important information in our research. Therefore, we collected images including this information.

B. Discovery of Sightseeing Spots

Sightseeing spots are discovered by clustering the collected images. The geolocation of each image is investigated to construct geo-clusters. We cluster the geolocation in using the nearest neighbor method. This is because the number of sightseeing spots is unknown. We continue clustering until the closest distance is greater than a threshold. Although image composition should be investigated to generate more accurate clusters, based on recent research it is difficult to identify an object among images. This is why we used geolocation only for generating geo-clusters. The latter are regarded as sightseeing spots.

C. Link Strength Estimation

We represent a directed link between two sightseeing spots by $L_{i,j}$, where i and j indicate individual sightseeing spots. $L_{i,j}$ and $L_{j,i}$ are distinguished to consider the direction from i to j and j to i, respectively. $L_{i,j}$ is determined by the owner ID and timestamp of the image. For instance, if two images are taken by the same owner at sightseeing spots i and j and the timestamp of the image at i is followed by the one at j, we establish a directed link between $L_{i,j}$.

The strength of the directed link $L_{i,j}$ is defined by $M(L_{i,j})$ and is calculated by counting the number of people who moved from i to j. A larger value of $M(L_{i,j})$ indicates strong connection between the two sightseeing spots, i.e., many people tend to visit these spots consecutively.

D. Characteristic Analysis

We analyze characteristics of each sightseeing spot based on the directed links explained above. First, the number of directed links is summed. With consideration of link direction, we sum two types of links. One is inflow, which denotes all links connected from any spot to spot j. The other is outflow, or links flowing out from spot j. The calculation is done as follows.

$$S_{in}(j) = \sum_{i \in \mathcal{I}} M(L_{i,j}) \quad (1)$$

 ISBN: 978-1-61208-364-3

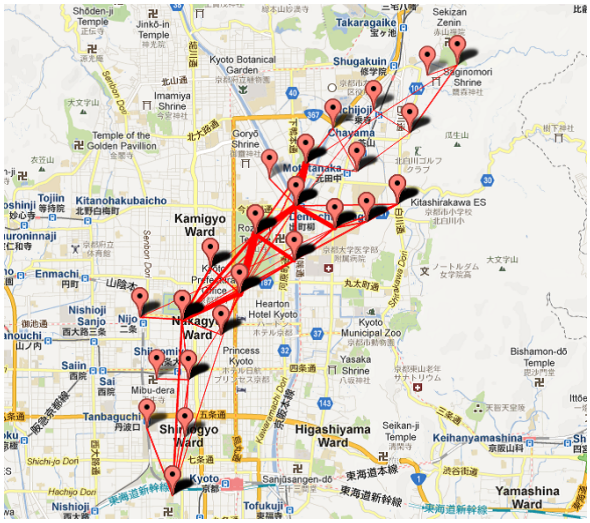


Figure 2: Estimated spots and links in Kyoto

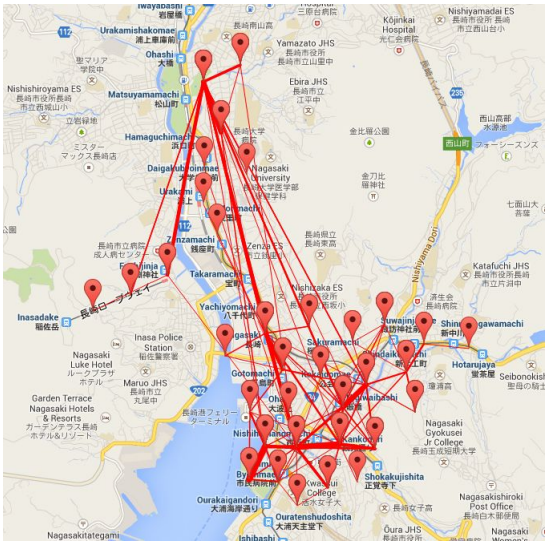


Figure 3: Estimated spots and links in Nagasaki

$$S_{out}(j) = \sum_{i \in \mathcal{O}} M(L_{i,j}) \tag{2}$$

, where $S_{in}(j)$ and $S_{out}(j)$ are inflow and outflow scores of spot j, respectively.$\mathcal{I}$ is a set of inflow links from any spot i to the focused j, and $\mathcal{O}$ is a set of outflow links from spot j to any other spot.

Next, we classify a characteristic of each spot into four types, based on $S_{in}(j)$ and $S_{out}(j)$ ”

Route Type
: A spot whose $S_{out}(j)/S_{in}(j) \approx 1$ and S_{in} is a small value, with few inflows and few outflows.

Convergence Type
: A spot whose $S_{out}(j)/S_{in}(j) \ll 1$, with many inflows and few outflows.

Divergence Type
: A spot whose $S_{out}(j)/S_{in}(j) \gg 1$, with few inflows and many outflows.

Hub Type
: A spot whose $S_{out}(j)/S_{in}(j) \approx 1$ and S_{in} is a large value, with many inflows and many outflows.

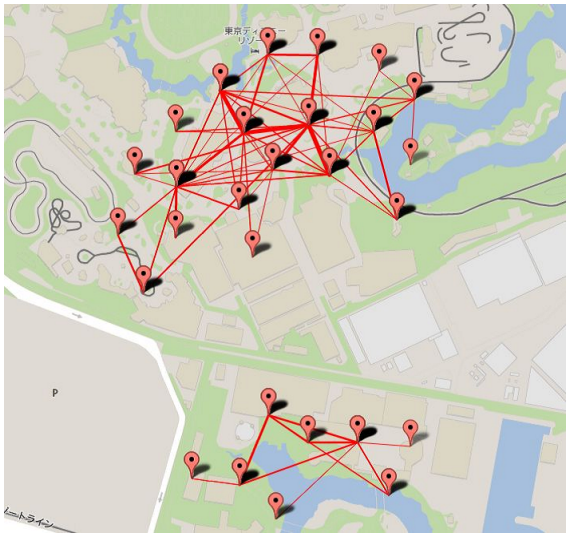


Figure 4: Estimated spots and links in amusement park

V. Experiment

A. Outline

We investigated the effectiveness of the proposed method in three situations. Two focused on sightseeing in the cities of Kyoto and Nagasaki in Japan, The other one focused on people’s activities in an amusement park (Tokyo Disney Resort [6]). We collected 16,210 images of Kyoto taken by 934 owners, 10,144 images of Nagasaki by 209 owners, and 16,250 images of the amusement park by 240 owners.

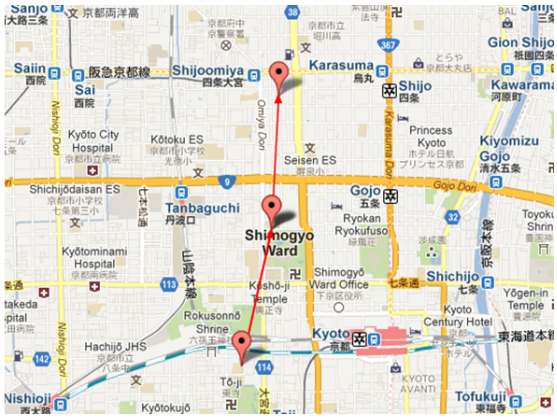
B. Estimated Spots and Links

There were 33 estimated spots and 82 links in Kyoto, whose strengths $M(L_{i,j}) \geq 2$ are shown in Fig. 2. Thick lines denote stronger links. Thirty-three sightseeing spots and 111 links were discovered in Nagasaki, as shown in Fig. 3. Twenty-eight spots and 91 links were discovered in the amusement park, as shown in Fig. 4. We found that most estimated spots were at famous sightseeing places. We discuss details of the results in the following subsection.

C. Estimated Type of Sightseeing Spots

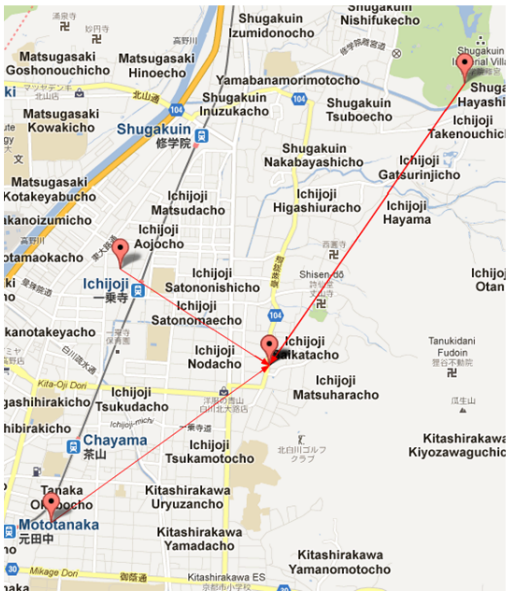
We analyzed characteristics of estimated sightseeing spots in the two cities and amusement park. Given the page limitation, we show only results for Kyoto. Representative results of each type are shown in Table I and Fig. 5. These results are discussed in the next subsection.

 ISBN: 978-1-61208-364-3

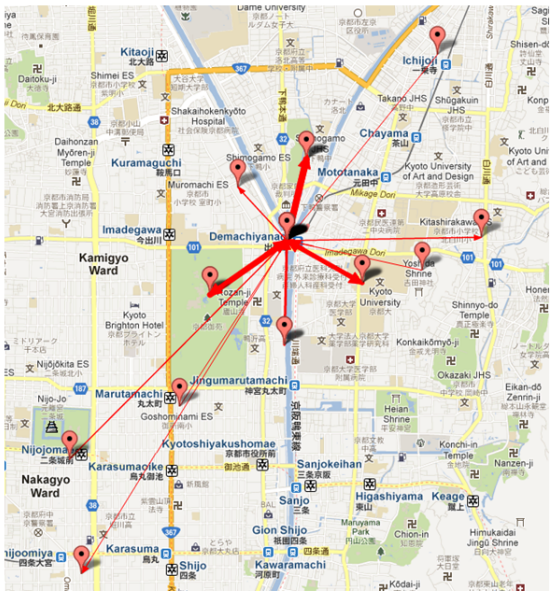


Route type: Nishi Hongan-ji Temple

Divergence type: Yoshida Shrine



Convergence type: Konpuku-ji Temple



Hub type: Demachi Bridge

Figure 5: Estimated type of sightseeing spots

TABLE I: Characteristics of sightseeing spots in Kyoto

Type	In	Out	O/I	Sightseeing spot
Route	1	1	1	Nishi Hongan-ji
Convergence	3	0	0	Kinpuku-ji
Divergence	2	4	2	Yoshida Shrine
Hub	8	7	0.875	Demachi Bridge

D. Discussion

1) Discussion of experimental results: We now address one of the results shown in Fig. 6. We put famous sightseeing spots on a map and found that the estimated spots nearly matched actual locations. A strong link was assessed between Nijo Castle and Kyoto Gosho. These spots are close, and some magazines introduce them as recommended spots in their area. This is why many people moved between the two spots. A strong link was also established between Demachi Bridge and Shimogamo Shrine. In the festival season, many people walk around these spots in Kyoto. The spot of Demachi Bridge was determined as a hub type, as shown in Fig. 4. There are bus stops and a subway station around this bridge, so this location is often used for connection between several sightseeing spots. The estimated results reflect such actual situations. Also, we address another one of the results shown in Fig. 7. A strong link was assessed between Dejima and Shian Bridge. It is possible to access by one tram. Thus, these spots likely to be visited together. The estimated results reflect such actual situations.

On the other hand, some estimated spots did not correspond to any famous places. Some are failure cases caused by images in the dataset with no relation to sightseeing. If we introduce

 ISBN: 978-1-61208-364-3

image processing for filtering out such negative samples, this kind of problem will be reduced. Other cases are "secret" spots found by tourists. Such information is not listed in sightseeing magazines. The proposed method therefore has great potential to mine such latent spots.

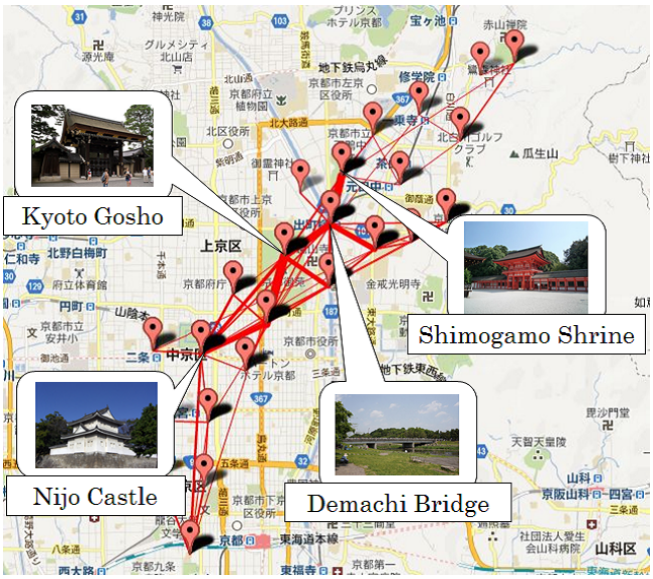


Figure 6: Strong links between sightseeing spots in Kyoto

Figure 7: Strong links between sightseeing spots in Nagasaki

2) Comparison with Other Methods: Here, we compare the related methods with ours. In most research recommending sightseeing information to tourists, strength of relationship between sightseeing spots is not taken into account. In our method, we use this strength, thereby providing more detailed information for tourists.

Okuyama and Yanai [7] recommended sightseeing routes to tourists using images with metadata. In their research, the presence of relationships among sightseeing spots was considered, but not their strength. With only presence, it is impossible to prioritize selected sightseeing spots. When we consider the selection of one spot from multiple candidates, our method is superior because it is possible to rank by a clear element of strength of relationship.

Zheng, Zhang, Xie and Ma [3] recommended sightseeing routes using strength of relationship among sightseeing spots in addition our method. However, their method differs from ours because they uniquely determined sightseeing routes. Furthermore, this method has a problem in that it may propose sightseeing spots in which tourists are not interested for part of the route. An absence of choice causes this problem. In contrast, we recommend multiple sequential sightseeing spots from the location where a tourist stays, by analyzing spots with strong connection to the stay location. Thus, tourists can select among candidates in which they are interested. As mentioned above, existing GPS datasets are not as common as image datasets on the web, making large-scale analysis difficult. Therefore, when we consider the variety of route selection and amount of data, our method is superior to that of Zheng, Zhang, Xie and Ma.

Lu [8] also recommended sightseeing routes using strength of relationship among sightseeing spots. They suggest a recommend route based on staying time of each tourist. This idea considers each tourist's characteristics in terms of staying time, but the recommended route does not always meet the demand of the tourist since it is not easy to make a plan of staying time in advance. On the other hand, our proposed approach can suggest several candidates of recommended routes for decision making.

E. Field Test

We developed a prototype route recommendation application that works on smartphones. The application is now available on the Android 4.X OS with a GPS sensor. A smartphone acquires location information (latitude and longitude) via the GPS, and then the location is sent to the server. The server retrieves several routes close to the current user location. Finally, the smartphone receives the route information and shows it as recommended routes on the display (Fig. 8). The application shows several arrows to suggest the next sightseeing spots from the current location. Thicker arrows indicate strong links (i.e., a strongly recommended place based on the analysis).

We conducted a field test in Nagasaki. About fourty people joined the experiment. The application was installed on their smartphones, and the people enjoyed sightseeing in Nagasaki. Note that the displayed arrows are just recommendations, and a user need not follow them.

We conducted questionnaires before and after the experiment. In the pre-questionnaire, we asked the subjects two questions:

Q1 Have you ever enjoyed sightseeing in Nagasaki?
Q2 Have you already planned where to go during this sightseeing?

 ISBN: 978-1-61208-364-3

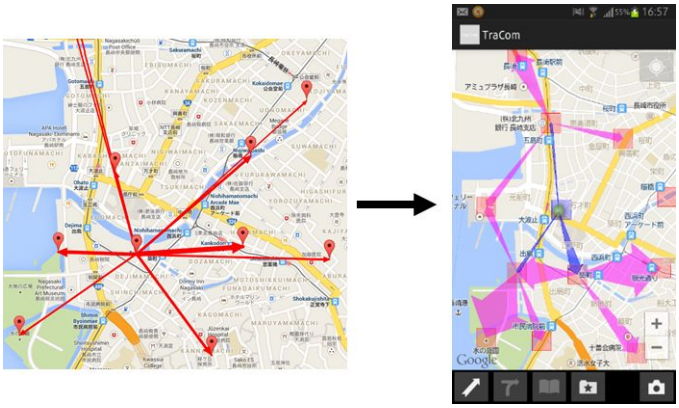


Figure 8: Application screen

TABLE II: Results of a questionnaire

Group	Q1	Q2	Ratio
Type 1	Yes	Yes	30%
Type 2	Yes	No	54%
Type 3	No	Yes	13%
Type 4	No	No	13%

We made four groups according to their answers, as shown in Table II. The rightmost column lists the ratios of people belonging to each group.

According to the post-questionnaire, we discovered that people in Type 1 and Type 3 tended to follow the recommendation more than those in Type 3 and Type 4. Please note that a displayed arrow navigates a user to a major sightseeing spot because it came from majority analysis. This is why people with no advance planning were affected by the arrows. Nonetheless, we think that sightseeing spots determined by a minority will be also useful information for people of Type 1 and Type 2, since they would like to visit “mysterious” spots that they have never seen. We are certain that the proposed method can be easily extended to find such minority-determined spots.

VI. Conclusion

In this paper, we proposed a new method to determine sightseeing spots and link their strengths. A large number of images collected from Flickr were analyzed for establishing the spots and link strengths. In addition, characteristics of the sightseeing were also assessed. This information was acquired from analyses of metadata that were attached to images shared in a large image database. We are certain that these types of information are applicable to route recommendations for sightseeing. We developed a prototype route recommendation application on smartphones and conducted onsite experiments in the city of Nagasaki. Through questionnaires and interviews of test subjects, we found that route recommendations displayed on the smartphone are useful, especially for people who did not decide sightseeing routes in advance.

In future work, we will tackle the following issues.

- We will introduce image and label processing to filter out non-useful images. The current system used all images from around the sightseeing areas. Some of these were unrelated to sightseeing. Such images can be filtered out if they are categorized using image contents and labels. Furthermore, image analyses will provide representative photos that are frequently taken around sightseeing spots. Such information will be helpful to determine subsequent visitation spots.
- We will improve the application design. The current application shows several arrows from the current location to major sightseeing spots. Through the questionnaire in the field test, we found that minority-determined spots are also useful for visitors. Therefore, it is important to change recommendation information in terms of visitor characteristics.

References

[1] “Picasa,” http://picasa.google.com/. [retrieved: May, 2014].

[2] “Flickr,” http://www.flickr.com/. [retrieved: May, 2014].

[3] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, “Mining interesting locations and travel sequences from gps trajectories,” ACM World Wide Web Conference, 2009, pp. 791–800.

[4] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T.S.Huang, “A worldwide tourism recommendation system based on geotagged web photos,” IEEE International Conference on Acoustics Spech and Signal Processing, 2010, pp. 2274–2277.

[5] Y. Arase, X. Xie, T. Hara, and S. Nishio, “Mining people’s trips from large scale geo-tagged photos,” International conference on Multimedia, 2010, pp. 133–142.

[6] “Tokyo disney resort,” http://www.tokyodisneyresort.co.jp/. [retrieved: May, 2014].

[7] K. Okuyama and K. Yanai, “A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web,” Pacific-Rim Conference on Multimedia, 2011, pp. 657–670.

[8] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, “Photo2trip: Generating travel routes from geo-tagged photos for trip planning,” ACM International Conference Multimedia, 2010, pp. 143–152.

 ISBN: 978-1-61208-364-3