IMMM 2012

The Second International Conference on Advances in Information Mining and Management

ISBN: 978-1-61208-227-1

October 21-26, 2012

Venice, Italy

IMMM 2012 Editors

Zari Dzalilov, University of Ballarat - Victoria, Australia

Petre Dini, Concordia University, Canada & China Space Agency, China

IMMM 2012

Foreword

The Second International Conference on Advances in Information Mining and Management [IMMM 2012], held between October 21-26, 2012 in Venice, Italy, 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 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a 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 2012. 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 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that IMMM 2012 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 also hope the attendees enjoyed the charm of Venice, Italy.

IMMM Chairs:

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
David Newell, Bournemouth University - Bournemouth, UK
Petre Dini, Concordia University, Canada & IARIA, USA
Andreas Holzinger, Institute for Medical Informatics, Statistics and Documentation (IMI) / Medical University Graz (MUG), Austria
Kuan-Ching Li, Providence University, Taiwan
Abdulrahman Yarali, Murray State University, USA
George Ioannidis, IN2 search interfaces development Ltd., UK
Johannes Meinecke, SAP AG / SAP Research Center Dresden, Germany
Robert Wrembel, Poznan University of Technology, Poland

Yulan He, Knowledge Media Institute / The Open University, UK
Stefan Brüggemann, OFFIS - Institute for Information Technology, Germany
Lena Strömbäck, Linköpings Universitet, Sweden
Roland Kübert, High Performance Computing Center Stuttgart / Universität Stuttgart, Germany
Zaher Al Aghbari, University of Sharjah, UAE
Alejandro Canovas Solbes, Polytechnic University of Valencia, Spain

IMMM 2012

Committee

IMMM General Chairs

Philip Davis, Bournemouth and Poole College - Bournemouth, UK
David Newell, Bournemouth University - Bournemouth, UK

IMMM Advisory Chairs

Petre Dini, Concordia University, Canada & IARIA, USA
Andreas Holzinger, Institute for Medical Informatics, Statistics and Documentation (IMI) / Medical University Graz (MUG), Austria
Kuan-Ching Li, Providence University, Taiwan
Abdulrahman Yarali, Murray State University, USA

IMMM Industry Liaison Chairs

George Ioannidis, IN2 search interfaces development Ltd., UK
Johannes Meinecke, SAP AG / SAP Research Center Dresden, Germany

IMMM Special Area Chairs on Data Management

Robert Wrembel, Poznan University of Technology, Poland

IMMM Special Area Chair on Special Mining

Yulan He, Knowledge Media Institute / The Open University, UK

IMMM Special Area Chair on Semantic Data Handling

Stefan Brüggemann, OFFIS - Institute for Information Technology, Germany

IMMM Special Area Chair on Databases

Lena Strömbäck, Linköpings Universitet, Sweden

IMMM Special Area Chair on Cloud-based Mining

Roland Kübert, High Performance Computing Center Stuttgart / Universität Stuttgart, Germany

IMMM Publicity Chairs

Zaher Al Aghbari, University of Sharjah, UAE
Alejandro Canovas Solbes, Polytechnic University of Valencia, Spain

IMMM 2012 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
Barbara Rita Barricelli, Università degli Studi di Milano, Italy
Shariq Bashir, National University of Computer and Emerging Sciences, Pakistan
Grigorios N. Beligiannis, University of Western Greece - Agrinio, Greece
Jorge Bernardino, ISEC - Polytechnic Institute of Coimbra, Portugal
Konstantinos Blekas, University of Ioannina, Greece
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Alain Casali, Aix Marseille Université, France
Sukalpa Chanda, Gjøvik University College, Norway
Chi-Hua Chen, National Chiao Tung University, Taiwan R.O.C.
Yili Chen, Monsanto Company, USA
Ronan Cummins, National University of Ireland - Galway, Ireland
Andre Ponce de Leon F. de Carvalho, University of Sao Paulo at Sao Carlos, Brazil
Sébastien Déjean, Université de Toulouse & CNRS, France
Juan José del Coz Velasco, Universidad de Oviedo - Gijón, Spain
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
Anass Elhaddadi, University of Paul Sabatier - Toulouse, France
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Ingrid Fischer, Universität Konstanz, Germany
Alessandro Giuliani, University of Cagliari, Italy
Eloy Gonzales, National Institute of Information and Communications Technology - Kyoto, Japan
Richard Gunstone, Bournemouth University, UK
Brian Harrington, Oxford University, UK
Nima Hatami, University of California - San Diego, USA
Kenji Hatano, Doshisha University, Japan
Ourania Hatzi, Harokopio University of Athens, Greece
Yulan He, The Open University, UK
Andreas Holzinger, Medical University Graz (MUG), Austria
Gilles Hubert, IRIT - University of Toulouse / Université Paul Sabatier, France
Masoumeh Izadi, McGill University Health Center - Montreal, Canada

Mansoor Zolghadri Jahromi, Shiraz University, Iran
Heng Ji, City University of New York, USA
Wei Jin, Amazon.com, Seattle, USA
Tahar Kechadi, University College Dublin, Ireland
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Frank Klawonn, Ostfalia University of Applied Sciences - Wolfenbuettel, Germany
Roland Kuebert, High Performance Computing Center Stuttgart (HLRS), Germany
Piotr Kulczycki, Polish Academy of Science | Cracow University of Technology, Poland
Rein Kuusik, Tallinn University of Technology, Estonia
Cristian Lai, CRS4, Italy
Giuliano Lancioni, Università Roma Tre, Italy
Mariusz Łapczyński, Cracow University of Economics, Poland
Hao Li, The City University of New York, USA
Kuan-Ching Li, Providence University, Taiwan
Shuying Li, University of Science and Technology of China (USTC), China
Tao Li, Florida International University, USA
Qing Liu, CSIRO, Australia
Elena Lloret Pastor, Universidad de Alicante, Spain
Qiang Ma, Kyoto University, Japan
Shuai Ma, Beihang University, China
Stéphane Maag, TELECOM SudParis, France
Thomas Mandl, Universität Hildesheim, Germany
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, Germany
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Charalampos Moschopoulos, Katholieke Universiteit Leuven, Belgium
Ulrich Norbisrath, BIOMETRY.com / University of Tartu, Estonia
Kok-Leong Ong, Deakin University, Australia
Samia Oussena, University of West London, UK
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Nathalie Pernelle, Université Paris-Sud, France
Ioannis Pratikakis, Democritus University of Thrace - Xanthi, Greece
Nishkam Ravi, NEC Labs - Princeton, USA
Daniel Romero, Cornell University, USA
Paolo Rosso, Universidad Politécnica Valencia, Spain
Igor Ruiz-Agundez, University of Deusto - Basque Country, Spain
Jörg Scheidt, University of Applied Sciences Hof, Germany
Armin Shams, University of Tehran, Iran
Hossein Sharif, University of Portsmouth, UK
Simeon Simoff, University of Western Sydney, Australia
Cristina Solimando, University Roma Tre, Italy
Tõnu Tamme, University of Tartu, Estonia
Yi Tang, Chinese Academy of Sciences, China
Xiaohui (Daniel) Tao, The University of Southern Queensland, Australia
Olivier Teste, Université de Toulouse, France

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

Negation Identification and Calculation in Sentiment Analysis 1
Amna Asmi and Tanko Ishaya

Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach 8
Nafissa Yussupova, Diana Bogdanova, and Maxim Boyko

A Novel Dependability Model to Define Normal Network Behavior 15
Maher Salem and Ulrich Buehler

Semantic Description of Text Mining Services 21
Katja Pfeifer and Alexander Schill

Semantic Tools for Forensics: A Highly Adaptable Framework 27
Michael Spranger, Stefan Schildbach, Florian Heinke, Steffen Grunert, and Dirk Labudde

Particular Requirements on Opinion Mining for the Insurance Business 32
Sven Rill, Johannes Drescher, Dirk Reinel, Jorg Scheidt, and Florian Wogenstein

How to Support Prediction of Amyloidogenic Regions - The Use of a GA-Based Wrapper Feature Selections 37
Olgierd Unold

Application of Optimisation-Based Data Mining Techniques to Medical Data Sets: A Comparative Analysis 41
Zari Dzalilov, Adil Bagirov, and Musa Mammadov

Network Monitoring Method Based on Self-learning and Multi-dimensional Analysis 47
Isao Shimokawa and Toshiaki Tarui

Structure Learning of Bayesian Networks Using a New Unrestricted Dependency Algorithm 54
Sona Taheri and Musa Mammadov

Comparison of Different Calculations of the Density-Based Local Outlier Factor 60
Vanda Vintrova, Tomas Vintr, and Hana Rezankova

Data Mining Application for Anti-Crisis Management 68
Nafisa Yusupova and Gyuzel Shakhmametova

Oracle NoSQL Database - Scalable, Transactional Key-value Store 75
Ashok Joshi, Sam Haradhvala, and Charles Lamb

Efficient Extraction of Motion Flow Data From a Repository of Three-Dimensional Trajectories Using Bi- 79

Dimensional Indexes
Antonio d'Acierno, Marco Leone, Alessia Saggese, and Mario Vento

On Biometric Verification of a User by Means of Eye Movement Data Mining 85
Youming Zhang and Martti Juhola

Extracting Transportation Information and Traffic Problems From Tweets During a Disaster 91
Aya Ishino, Shuhei Odawara, Hidetsugu Nanba, and Toshiyuki Takezawa

Classification of Time-Interval and Hybrid Sequential Temporal Patterns 97
Mohammed AL Zamil

Optimized Class Association Rule Mining using Genetic Network Programming with Automatic Termination 102
Eloy Gonzales, Bun Theang Ong, and Koji Zettsu

Information Mining Over Significant Interval on Historical Data: A Study on World Major Indexes 108
Kwan-Hua Sim

A Fast Short Read Alignment Algorithm Using Histogram-Based Features 114
Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi

AgileKDD An Agile Knowledge Discovery in Databases Process Model 118
Givanildo Nascimento and Adicineia Oliveira

An Improved Face Recognition Algorithm Using Adjacent Pixel Intensity Difference Quantization Histogram and Markov Stationary Feature 123
Feifei Lee, Koji Kotani, Qiu Chen, and Tadahiro Ohmi

Music Recommendation Based on Text Mining 129
Ziwon Hyung, MyoungA Lee, and Kyogu Lee

Automated Reference Model Generation and Utilization for Dimensional Control of Large Scale Assemblies and Assembly Processes 135
Teuvo Heimonen and Markku Manninen

ListCreator: Entity Ranking on the Web 141
Alexandros Komninos and Avi Arampatzis

Parallel Processing of Very Many Textual Customers' Reviews Freely Written Down in Natural Languages 147
Jan Zizka and Frantisek Darena

A New Algorithm for Accurate Histogram Construction 154
Zeineb Dhouioui, Wissem Labbadi, and Jalel Akaichi

Negation Identification and Calculation in Sentiment Analysis

Amna Asmi
University of Hull, UK
e-mail: A.Asmi@2008.hull.ac.uk

Tanko Ishaya
University of Jos, Nigeria
e-mail: ishayat@unijos.edu.ng

Abstract—The extensive growth of user-generated content has introduced new aspects of analysis on World Wide Web data. Sentiment analysis of written text on the web is one of the text mining aspects used to find out sentiments in a given text. The process of sentiment analysis is a task of detecting, extracting and classifying opinions and sentiments expressed in texts. It includes the identification of the meaning of words within the text through natural language processing rules. While existing research presents a number of approaches for sentiment analysis, these approaches have not quite provided an appropriate and efficient way of calculating and representing the role of negation in sentiment analysis. Therefore, this paper presents a framework for automatic identification of the presence of opinion in textual data. The proposed framework includes a description of rules for negation identification and calculation. These negation rules are designed in order to improve sentiment text analysis. Main achievement of the paper is a demonstration on an approach for automatic identification and calculations of negation in opinion and sentiment analysis.

Keywords-Negation Identification; Negation Calculation; Subjectivity Analysis; Sentiment Analysis; Opinion Mining; Social Media Mining; Text Mining.

I. INTRODUCTION

The aim of sentiment analysis is to find out the positive and negative feelings, emotions and opinions written in a text. These sentiments are based on the meaning of words used in text according to different scenarios and situations. There are a variety of ways used to express the same feeling in a written text by using different grammatical rules. These grammatical rules contain negations that are very frequently used in text that completely change the meanings of words. In other words, negation identification and detecting its scope with in a sentence (text) are necessary in finding out the sentiments from a piece of text. Although negation identification is an important aspect of sentiment analysis, it is yet to be properly addressed. In general, the efforts put into sentiment analysis of sentences having negation terms in them are less efficient with respect to general sentiment analysis. Negation identification is not a simple task and its complexity increases, since negation words such as *not, nor* etc., (syntactic negation) are not the only criterion for negation calculation. The linguistic patterns - prefixes (e.g., un-, dis-, etc.) or suffixes (e.g., -less) also introduce the context of negation in textual data [24]. Similarly, word intensifiers and diminishers (contextual valence shifter) also flip the polarity of sentiments [7, 21]. It will take a lot of efforts to enlist all such words in one list. These valence shifters do not only flip the polarity but also increase or decrease the degree to which a sentimental term is positive or negative [5]. On the other hand, negation does not restrict itself to 'not'. There are terms like; *no, not, n't, never, no longer, no more, no way, nowhere, by no means, at no time*, etc. [5, 21] that also change the meaning of a sentence. However, the precision involving the negative word "*not*" is very low, at 63% [5]. Another reason is the fact that the number of negation sentences encountered is considered insignificant during the evaluation of any sentiment analysis system as compared to the level of effort required to resolve the issues related to negation. This paper is an effort towards finding an approach to handle the syntactic negation for sentiment analysis by not only using the polarity and its intensity for words but also using the dependencies, relation within the sentences and sentence structure. The negation is handled with the diminishers, intensifiers and negation terms during the process of sentiment analysis. This research mainly focuses on the identification of negation, and identification of scope of syntactic negation. It presents a proposed framework for automatic identification of opinion in textual data. The framework has been implemented and evaluated by verifying the polarity identified by prototype system with a group of participants.

The rest of the paper has been structured as follows: Section 2 presents an analysis of related work in the area of sentient analysis. This is followed by a description of the proposed framework for sentiment analysis, and the existing resources used to generate dependencies in Section 3. Section 4 presents an application of this proposed framework for negation handling in sentiment analysis. It also explains the basic rules used in this framework for handling negation. Section 5 involves an analysis of the technique used through some example illustrations together with an analysis of the results of the prototype evaluation. Section 6 provides a conclusion and the prospective extensions to this work.

II. LITERATURE REVIEW

Text based information is broadly classified into two basic types, facts and opinions. In other words, textual analysis can be understood as classification of text either positive/negative (document or sentence level sentiment classification) or subjective/objective (sentence level subjectivity classification). Sentiment analysis is a process, which deals with the detection of sentiments, opinions, emotions, appraisals and feelings towards entities, events and properties [22]. The concept of emotion, opinion or sentiment is very broad. Different researchers have identified different spectrums of emotions in different dimensions [15, 10]. However, it is believed that all these different dimensions can be mapped to either positive or negative

 ISBN: 978-1-61208-227-1

emotions [25]. On the basis of this believe, research in sentiment analysis and opinion mining has considered positive and negative feelings. Most researchers in the field of opinion mining have used the lexicons and lists of words, with word as basic unit of expression of emotions in any language. Lexicon based negation i.e., negation introduced by suffix and/or prefix is easily handled with the help of a good lexical resource, i.e., dictionary, ontology, database etc. However, more emphasis on opinion analysis should be on how these words are joined and correlated with other words to give specific meanings in any language. This inter-relationship of words makes up sentences, which is why it is important to emphasis on finding the scope of negation, diminshers or intensifiers. Due to syntactic and semantic differences, it is difficult to interpret the intensity of polarity. While calculating a value of intensity of any sentence, there are always modifiers, which not only change the polarity of other words in the sentence but also affect the intensity. Negation is a complex thing as it changes the meaning (polarity and its intensity) if used within a clause. It is also difficult to identify which part of a clause a negation is changing in a sentence. The following sections II A – II D highlight different methods used for negation identification and how they affect sentiment analysis of text. Section II E discusses the State of Art for analysis of Negation.

A. Bag of Words

Bag of words (BOW) is a technique where each word in a document is represented by a separate variable numeric value (weight) [26]. It is the most widely used technique for sentiment analysis [3, 11]. Das and Chen [3] incorporated negation in their research for extraction of sentiments from stock market message boards. They believed that negation in a sentence reverses the meanings of the sentence. They discussed how words like “not”, “never”, “no”, etc., serve to reverse sentence meaning. They detected negation words in sentences and tag from the sentences with negation markers [3]. In 2002, researchers in [11] adopted the same technique and added the negation word with every word until the first punctuation mark following the negation word. An example that better explains this technique is “I do not NOT like NOT this NOT new NOT Nokia NOT model” [17]. From the example above, it can be seen that this technique is not an effective way to find out the negation from a written text as negation may be based on a meaning of words, whereas understanding a scope is very necessary to determine such meanings. Another limitation of this technique is that it is based on the list of words, and lists in any language can never be complete.

B. Contextual Valence Shifter

Contextual Valence Shifters or modifiers are the words, which change (boost, enhance, diminish etc.) meanings [8]. Many researchers have transferred their research on sentiment analysis from BOW to Parts of Speech (POS) especially Verbs, Adjectives and Adverbs. The pioneers in giving an understanding that there is a basic polarity associated with every word were in [12]. However, lots of contextual shifters are still needed to change or modify the valance associated with words. Negatives, intensifiers or diminishers are examples of contextual shifter [12]. For example; Negatives: John is clever versus John is not clever. Intensifier: Sam is suspicious about Anna versus Sam is deeply suspicious about Anna. Diminishers: I know what to say versus I hardly know what to say. Wiegand et al. [17] believed that the effectiveness of the model believed that the effectiveness of the model could be better judged if was evaluated. Kennedy and Inkpen [6] used the same model for Contextual Valence Shifters. They enhanced their model but still kept the scope of any negation term as immediately preceding a term. There is a need for relationship finder to define the scope of negation terms [7]. Other researchers have tried to define the scope by defining lists of verbs, adjectives and adverbs and defining their relationships for sentiment analysis [16]. Lists of positive and negative terms and a set of lists for modifiers was proposed in [8] to define the scope of these modifiers as *n*- terms before and after positive or negative terms, although this *n* remained a constant. This technique is better for negation identification in comparison to the BOW technique. However, it also considered the propagation of lists as a limitation. The lists used for this technique may grow with time and can never be complete, as in any language there might be infinite number of words and ways they can be used. Therefore, there is always a need to devise some way for the system to handle words, which are not present in the lists.

C. Semantic Relations

Semantic relations refer to the relationship between concepts or meanings for example antonym, synonym, homonym etc. It is evident from existing research that semantic relationship is also used for negation identification. It is clear that atomic words, which can provide a misleading polarity for sentences as words can be modified (weakened, strengthened, or reversed) based on lexicon, discourse, or paralinguistic contextual operators [12]. The use of linguistic structure of sentence for sentiment analysis was proposed in [9], where the polarity of a sentence is dependent upon the polarities of its parts: noun phrases (NP), verb phrases (VP) and individual parts of speech. Negation is handled by defining different intensities of negation words. In other words, the negation of words can change the polarity of an entire sentence or only parts of it [17]. Shaikh et al. [14] has used a similar approach to calculate the sentence level sentiment analysis. They performed semantic dependency analysis on the semantic verb frames of each sentence, and apply a set of rules to each dependency relation to calculate the contextual valence of the whole sentence [14]. A two-phase process was proposed in [2] as another way of compositional semantics. They identify the polarity of words in the first phrase where all the words are classified on the basis of the level of their strength in terms of the scope in the sentence. The second phase is based on the inference rules, which identify the polarity modification feature. For example, in the sentence “They could not eliminate my doubt”, the word *not* is a negater whereas *eliminate* also reverses the polarity of doubt, and not is reversing polarity of eliminate. These rules are much different as compared to the

 ISBN: 978-1-61208-227-1

ones presented in [18] and [9, 17]. This approach is working well for simple sentences in the written text but has failed for compound sentences where a sentence may have word-based or sentence-based dependencies.

D. Relations and Dependency Based

The grammatical relationships between the words within a sentence and syntactic dependencies help in extraction of textual relations. Reschke and Anand [13] have given a context aware approach for sentiment analysis where the sentiment is evaluated towards a target entity, event, or proposition. The scope of words is defined by the clauses or phrases (noun phrase, verb phrase) in the sentence and sentiment in the sentences are understood by the heuristic rules defined to join the clauses [13]. Jai and others also tried to identify the scope of different terms by using Stanford Parser tree [27]. They used simple tree based rules by identifying the dependent terms and later used some parts of speech based tools to understand the sentimental behavior of negation [5].

There is quite an extensive research undertaken for sentiment analysis and scope of different words and their relation within a sentence and on broader sense domain. However, negation is still an over looked domain probably because of the low proportion of the number of negative sentences encountered during the evaluation of sentiment analysis. Irony is a process of using words and phrases (generally positive) that are generally different if used otherwise [1]. Therefore identification, extraction and analysis of irony are difficult.

E. Analysis of Negation

For the sentence level sentiment analysis in English language, the basic structure of English sentence and its parts: clauses and phrases are necessary to be understood. These parts further divide sentences into different types of sentences (simple, complex and compound). The sentence is made more complex by adding declarative, interrogative, exclamatory and imperative sentences. In order to further complicate the problem as the comparison, contradiction, negation and irony might also be introduced in the sentences. Negation needs to identify its scope, negation can be local (e.g., not good), or it can involve longer-distance dependencies (e.g., does not look very good) or the negation of the subject (e.g., no one thinks that it's good). It even changes its roles i.e., instead of negating and can even intensify (e.g., not only good but amazing) [18]. In order to find out the scope of the negation, the sequence of words in the sentence should be identified. On the whole, it is not simply the negation of a word but negation of the sentence [19, 21].

The expression of negation within a sentence varies a lot. It can be a verb, adverb, suffix or prefix. It might also occur more than once in a sentence and rather than cancelling each other it can give negative meaning, for example; I cannot get no satisfaction [4]. Therefore, the negation analysis has been done using many different ways: Parts of Speech, Bag of Words, and Dependency Tree. However, the best results can be found by combining these approaches. A way to understand the negation by using bag of words approach and latter resolving the scope with the help of dependency tree was proposed in [20]. The following section explains the proposed framework for sentiment analysis and the approach for negation identification and calculation that helps to solve the negation problem in sentiment analysis.

III. FRAMEWORK FOR SENTIMENT ANALYSIS

This section introduces a framework for sentiment analysis and explains how it is handling negation identification, scope of negation and calculation of sentiment on sentence level. The framework presented in Figure 1, consists of a number of detection, extraction and classification components interacting at various levels.

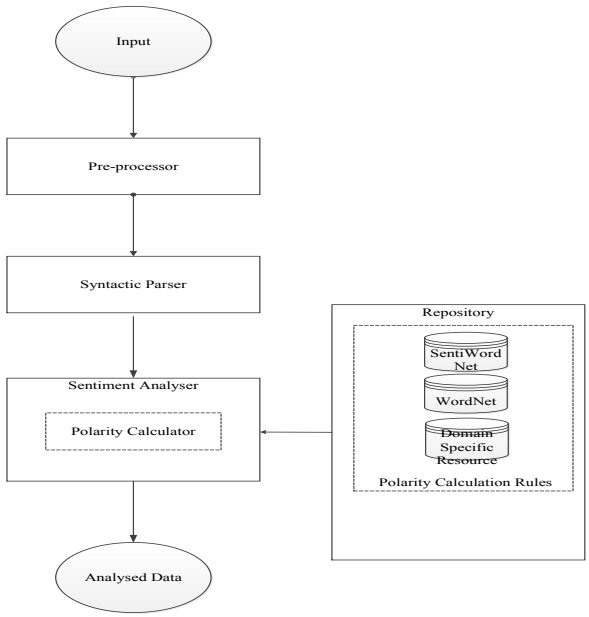


Figure 1. Framework for Sentiment Analysis

The framework shows a mixed and combined approach to lexical and syntactic analysis for sentiments. The framework uses a number of existing lexical and syntactic analysis resources for sentiment analysis. Its main components are briefly described in Sections A through C.

A. Pre-processor

The pre-processing phase of the system takes text as input and arranges all the data in required format. It splits data into sentences and forwards all the sentences to syntactic parser.

B. Syntactic Parser

The syntactic parser is an iterative parser, which uses Penn Tree Bank [30] parser to assign Parts of Speech (POS) tags to each word in the sentence. The name entities and idioms involved in a sentence are also identified in syntactic parsing. It also uses Stanford Parser [27] to identify how different words are interacting within a sentence and identifies the syntactic dependencies/relationship within a

 ISBN: 978-1-61208-227-1

sentence. The syntactic parser parses each sentence iteratively with all the identified information to the sentiment analyser after classifying the sentence as a question, an assertion, a comparison, a confirmation seeking or a confirmation providing by using the rule of sentence type identification.

C. Sentiment Analyzer

The sentiment analyzer is basically the main part of proposed framework. It uses general resources like SentiWordNet [28], WordNet [29] and any domain specific resource to extracts the sentiment-oriented words from each sentence by using the relationship information of (dependencies within) the sentence. The Sentiment Analyzer has two sub modules, which help in calculating the polarity of sentences and documents. The Polarity Calculator (PC) calculates the polarity of a sentence and assigns a score. In order to calculate polarity, PC uses SentiWordNet [28] to identify the positive and negative words and their values assigned by the SentiWordNet [28]. In this process, PC collects the synonyms of a word if its not found in SentiWordNet [28]. The PC first uses WordNet [29] to get the synonyms. The sentiment analyzer generates frames for each sentence. A frame contains the type of sentence, subject, object/feature, sentiment oriented word(s), sentiment type (absolute or relative), sentiment strength (very weak, weak, average, strong or very strong) and polarity of sentence.

IV. USAGE OF FRAMEWORK FOR NEGATION

All the sentences having negation are forwarded from the pre-processor to the syntactic parser with other sentences. There is no specific requirement for handling the negation sentences for pre-processing. However, syntactic parser identifies the negation and POS that are involved in negation with the help of Stanford Dependency Parser [27] during the syntactic parsing phase. In the negation identification process, the kind of negation i.e., no one likes his behavior where 'no' is used to determine the behavior of one, is also identified. This process also takes care of the negation in conjunction sentences. The negation identification is very import part of syntactic parser that is used for polarity calculation by the sentiment analyzer. The following section explains how sentiment analyzer uses the negation for polarity calculation.

A. Polarity Calculation

Sentiment analysis identifies the semantics involved in a sentence. The words in a sentence, their meanings, alternative words, polarity of each word and intensity associated with each word are basic elements used by sentiment analyzer for sentiment identification. The polarity of sentence is usually based on the meaning of words. However, the negation (only for negation sentences) changes the meaning of the words and polarity of the sentence. In order to calculate the polarity of a sentence, some rules are defined in Table 1. These rules are defined on the basis of POS. Most negation words are classified as adverbs, suffix, prefix or verbs. However, the nouns are generally there to determine the meaning of another noun. The scope of negation will be identified by the dependency tree, which indicates how negation is interacting with other words in the sentence. This dependency will identify the scope of the negation - whether it is a single word or a phrase / clause within a sentence. In the case of a clause or phrase, the noun phrase/ clause is first calculated for the sentiment polarity before the verb phrase or clause sentiment polarity is calculated. The negation is handled in each phrase accordingly. The intensity of polarity will not exceed (+/-) 1, where + is for positive and – is for negative polarity. The intensity of a sentence is calculated as:

*Resulting Intensity = First Word/Phrase/Clause + [(1 - Second Word/Phrase/Clause) * Second Word/Phrase/Clause]* *(1)*

TABLE 1. RULES SPECIFYING NEGATION

First Word /Phrase /Clause	Second Word /Phrase /Caluse	Negation	Result
Positive	Positive	True	Negative
Positive	Positive	False	Positive
Positive	Negative	True	Positive
Positive	Negative	False	Negative
Negative	Positive	True	Positive
Negative	Positive	False	Negative
Negative	Negative	True	Negative
Negative	Negative	False	Positive

The positive/negative value of words in the Equation 1 is extracted from the SentiWordNet [28] in order to calculate the polarity of a sentence. The extracted value from the SentiWordNet [28] is reversed during this process if negation is 'True' as presented in Table 1.

B. Algorithm for Polarity Calculation

Function CalculatePolarity Returns Polarity {

```
Double polarity = 0

For Each nounPhraseOfSentence {
  get SentiWordNet value of all Adjectives and Nouns of
  noun-phrase
  If (Sentence is Marked NEGATION by Syntax Parser) {
    Reverse the SentiWordNet values of related
    Nouns/Adjectives }
  For Each Noun and Adjective {
    polarity += [( 1 – Noun/Adjective) * Noun/Adjective]
  } }  For Each verbPhraseOfSentence {
  get SentiWordNet value of all Adverbs and Verbs of
  verb-phrase;
  If (Sentence is Marked NEGATION by Syntax Parser) {
    Reverse the SentiWordNet values of related
    Verbs/Adverbs  }
  For Each Verb and Adverb {
```

polarity += [(1 – Verb/Adverb) * Verb/Adverb]
} } Return polarity }

The syntax parser forwards each sentence to the sentiment parser as mentioned in Figure 1. The syntax parser identifies the 'negation' for negation sentences and also identifies the words identifying the negation before handing over the sentences to the sentiment parser. For the sentiment analysis, the sentiment parser calculates the polarity of each sentence through the above algorithm. In order to calculate the polarity of a sentence, all the noun and verb phrases are calculated. Polarity calculator gets the values of all nouns and adjectives involved in a noun phrase from the SentiWordNet [28]. These values are reversed by the polarity calculation in case of negation sentence depending on the negation scope. Similarly, polarity calculator obtains the values of all verbs and adverbs involved in the verb phrase from the SentiWordNet [28] and reversed these values in case of negation sentence. The whole process uses Equation 1 iteratively for polarity calculation while solving each noun and verb phrase.

V. ANALYSIS

The sentence polarity is calculated on the basis of the parts of a sentence. A sentence may contain either simple POS (Verb, Adverb, Adjectives, etc.) or complex parts of speech (Noun Phrase [Pronoun, Noun] or Verb Phrase [Verb, Noun Phrase], relations of possession, determiner, etc.). The following hierarchy is an example of POS in a complete sentence.

(Sentence
(Noun Phrase (Pronoun, Noun))
(Adverbial Phrase (Adverb))
(Verb Phrase (Verb)
(Sentence
(Verb Phrase (Verb)
(Noun Phrase (Noun))
))))

Sentiment polarity calculation is a nested process. This process calculates the sentiment of the most inner level first and then it calculates along with the next higher level, which is also called Sentiment Propagation [23]. This process calculates the polarity and intensity of the words and phrases. If there is a negation term the polarity will be calculated accordingly. The following three examples illustrate the whole process of polarity calculation.

A. Example 1

They have not succeeded, and will never succeed, in breaking the will of this valiant people.

(Sentence
(Pronoun They)
(Verb Phrase
(Verb Phrase (have not)
(Verb Phrase (Verb succeeded)))
(and)
(Verb Phrase (will)
(Adverbial Phrase (Adverb never))
(Verb Phrase (succeed)))
(Prepositional Phrase (in)
(Sentence
(Verb Phrase (breaking)
(Noun Phrase
(Noun Phrase (the will))
(Prepositional Phrase (of)
(Noun Phrase (this valiant people)))))))))

The negation word 'not' is affecting the succeeded (+) whereas never is effecting succeed (+) where succeeded and succeed are joined by and (joins same polarity). Both successes are in breaking (-) the will of people who are valiant (+) people. As they have not succeeded in doing something Negative and the polarity of sentence is Positive as shown in Figure 2.

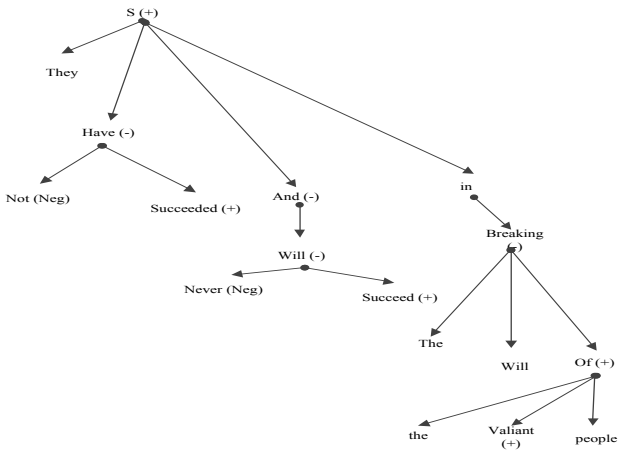


Figure 2. Dependency Tree Structure for Example 1

B. Example 2

Jhon is never successful at tennis.

(Sentence
(Noun Phrase (Jhon))
(Verb Phrase (is)
(Adverbial Phrase (never))
(Adjectival Phrase (successful)
(Prepositional Phrase (at)
(Noun Phrase (tennis))))))

Negation never is for successful (+) and this success is at tennis. This negation of positive term is a simple negation, which is presented in Figure 3.

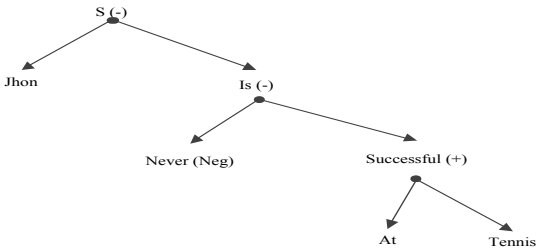


Figure 3. Dependency Tree Structure for Example 2

C. *Example 3*

The audio system on this television is not very good, but the picture is amazing.

(Sentence
(Sentence
(Noun Phrase
(Noun Phrase (the audio system))
(Prepositional Phrase (on)
(Noun Phrase (this television))))
(Verb Phrase (is not)
(Adjectival Phrase (very good))))
(, ,)
(Conjunction but)
(Sentence
(Noun Phrase (the picture))
(Verb Phrase (is)
(Adjectival (amazing)))))

Negation not is effecting the Adjectival Phrase (very good (+)) whereas the sentence also has a conjunction of 'but' which is followed by a positive clause 'the picture is amazing (+). The conjunction 'but' diminishes the meaning of first negative and gives emphasis to following positive clause as presented in Figure 4.

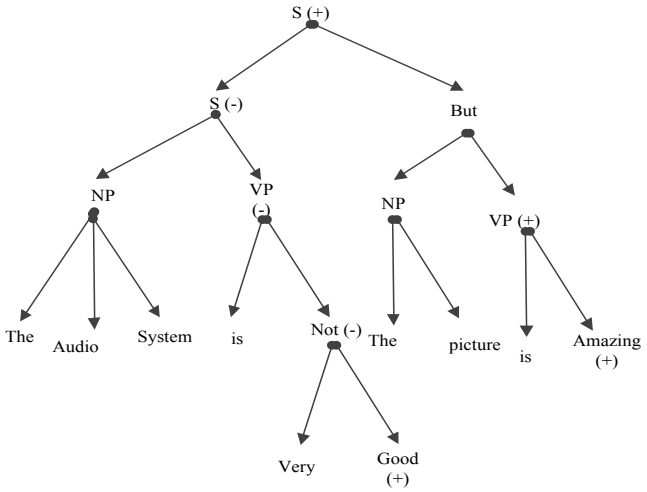


Figure 4 Dependency Tree Structure for Example 3

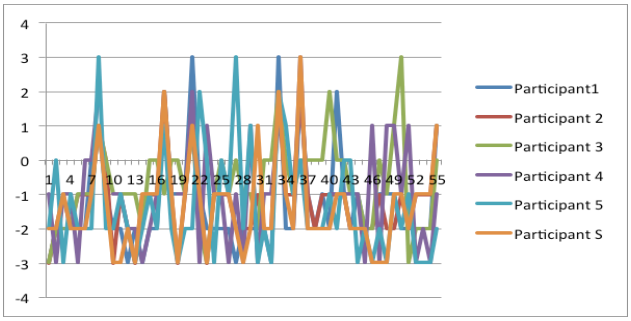


Figure 5. Graph showing Responses from Five Respondents and the Prototype System (Participant S) generated opinion scores

From the Figure 5 above, there is a clear agreement of the annotations made by all the five participants with the system with all the 55 sentences. Only two sentences (30 and 55) have more than two annotators that have given opinion polarities that are different from that of the system generated. Close analysis of these two sentences have shown why the difference.

Furthermore, the relationship between opinion polarity and intensity (as generated by the system) and all the five user generated opinion score for a sample 55 sentences was investigated using the Pearson product-moment correlation coefficient (*r*). Table 2 presents the result of the calculated multiple regression.

Table 2 Pearson Product-Moment Correlations between System and User Generated Opinion Polarity

Correlations

		Participant 1 Polarity Score	Participant 2 Polarity Score	Participant 3 Polarity Score	Participant 4 Polarity Score	Participant 5 Polarity Score	System Polarity Score
Participant 1 Polarity Score	Pearson Correlation	1	.754**	.413**	.544**	.260	.693**
	Sig. (2-tailed)		.000	.002	.000	.055	.000
	N	55	55	55	51	55	55
Participant 2 Polarity Score	Pearson Correlation	.754**	1	.412**	.342*	.329*	.872**
	Sig. (2-tailed)	.000		.002	.014	.014	.000
	N	55	55	55	51	55	55
Participant 3 Polarity Score	Pearson Correlation	.413**	.412**	1	.270	.215	.300*
	Sig. (2-tailed)	.002	.002		.056	.116	.026
	N	55	55	55	51	55	55
Participant 4 Polarity Score	Pearson Correlation	.544**	.342*	.270	1	.326*	.359**
	Sig. (2-tailed)	.000	.014	.056		.019	.010
	N	51	51	51	51	51	51
Participant 5 Polarity Score	Pearson Correlation	.260	.329*	.215	.326*	1	.327*
	Sig. (2-tailed)	.055	.014	.116	.019		.015
	N	55	55	55	51	55	55
System Polarity Score	Pearson Correlation	.693**	.872**	.300*	.359**	.327*	1
	Sig. (2-tailed)	.000	.000	.026	.010	.015	
	N	55	55	55	51	55	55

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

From the table above, we can see that the data showed no violation of normality between all the five sets. For example, *r*=.693, indicates a positive correlation between system generated polarity and opinion oriented 1, which shows a strong, positive correlation between the two variables, r=.693, n=52, p<.0005 with high levels of system generated polarity scores associated with user generated polarity scores for the sample sentences.

VI. CONCLUSION AND FUTURE WORK

Current research on sentiment analysis shows that there is a growing need to develop approaches to cope with the variety of evolving social media generated text. One aspect of research that has been identified as important, but has still received little attention is the identification of negation and its implication on the semantic understanding of sentences. This paper presents an evaluation of existing approaches to sentiment analysis and presents an approach for negation identification and calculation using a developed framework for sentiment analysis. These negation rules are designed in order to improve the sentiment text analysis.

While, there are still a number of challenges to be addressed in the field of sentiment analysis, the developed rules for negation calculation is being integrated within the general framework developed in Figure 1 within polarity

 ISBN: 978-1-61208-227-1

calculation. Further work will also include the implementation of prepositional negation calculation.

The framework is not designed by keeping any specific lexical resource in mind; therefore, by improving the precision of resources the results can easily be improved.

REFERENCES

[1] Carvalho, P., Sarmento, L., Silva, M. and Oliveira, E . (2009) Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In: *The 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. Hong Kong, China. ACM, 53-56.

[2] Choi, Y. and Cardie, C. (2008) Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: *The Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii. Association for Computational Linguistics, pp. 793-801.

[3] Das, S. and Chen, M. (2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Asia Pacific Finance Association Annual Conf. APFA (2001)*,

[4] Horn, L. R. and Kato, Y. (2000) Introduction: Negation and Polarity. at the Millennium.

[5] Jia, L., Yu, C. and Meng, W. (2009) The effect of negation on sentiment analysis and retrieval effectiveness. In: *The 18th ACM conference on Information and knowledge management*. Hong Kong, China. ACM, 1827-1830

[6] Kennedy, A. and Inkpen, D. (2005) Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. In: *FINEXIN 2005*. Ottawa.

[7] Kennedy, A. and Inkpen, D. (2006) Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. . *Computational Intelligence.* , Vol. 22 2, pp. 110-125.

[8] Li, N. and Wu, D. D. (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis. Support Syst.*, Vol. 48, 2, pp. 354-368.

[9] Moilanen, K. and Pulman, S. (2007) Sentiment Construction. In: *Recent Advances in Natural Language Processing RANLP*. Borovets, Bulgaria.

[10] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2007) Textual Affect Sensing for Sociable and Expressive Online Communication. In: *The 2nd international conference on Affective Computing and Intelligent Interaction*. Lisbon, Portugal. Springer-Verlag, 218-229.

[11] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques. In: *The ACL-02 conference on Empirical methods in natural language processing* Association for Computational Linguistics, 79-86.

[12] Polanyi, L. and Zaenen, A. (2004) Contextual Valence Shifters. In: *The AAAI Spring Symposium on Exploring Attitude and Affect in Text.* California, USA.

[13] Reschke, K. and Anand, P. (2011) Extracting contextual evaluativity. In: *The Ninth International Conference on Computational Semantics*. Oxford, United Kingdom. Association for Computational Linguistics, 370-374.

[14] Shaikh, M. A., Prendinger, H. and Mitsuru, I. (2007) Assessing Sentiment of Text by Semantic Dependency and Contextual Valence Analysis. In: *The 2nd international conference on Affective Computing and Intelligent Interaction*. Lisbon, Portugal. Springer-Verlag, 191 - 202

[15] Smith, D. B. (2010) Plutchik's Eight Primary Emotions. wordpress.com.

[16] Subrahmanian, V. S. and Reforgiato, D. (2008) AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis. *IEEE Intelligent Systems*, Vol. 23, 4, pp. 43-50.

[17] Wiegand, M., et al. (2010) A survey on the role of negation in sentiment analysis. In: *The Workshop on Negation and Speculation in Natural Language Processing*. Uppsala, Sweden. Association for Computational Linguistics, 60-68

[18] Wilson, T., Wiebe, J. and Hoffmann, P. (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: *The conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada. Association for Computational Linguistics, 347-354.

[19] Morante, R. and Daelemans, W. (2009) A metalearning approach to processing the scope of negation. In: *The Thirteenth Conference on Computational Natural Language Learning*. Boulder, Colorado. Association for Computational Linguistics, 21-29.

[20] Nakagawa, T., Inui, K. and Kurohashi, S. (2010) Dependency tree-based sentiment classification using CRFs with hidden variables. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California. Association for Computational Linguistics, 786-794.

[21] Kirkegaard, O. W. (n.d.) Negating sentences in english [WWW]. Available from: http://emilkirkegaard.dk/en/wp-content/uploads/Negating-sentences-in-english.pdf [Accessed June 9, 2012].

[22] Abbasi, A., Chen, H. and Salem, A. (2008) Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Trans. Inf. Syst.*, Vol. 26, 3, pp. 1-34.

[23] Heerschop, B., Hogenboom, A. and Frasincar, F. (2011) Sentiment Lexicon Creation from Lexical Resources. In: *14th International Conference on Business Information Systems (BIS 2011)*. Springer, 185-196.

[24] Councill, I. G., Mcdonald, R. and Velikovich, L. (2010) What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In: *The Workshop on Negation and Speculation in Natural Language Processing*. Uppsala, Sweden. Association for Computational Linguistics, 51-59.

[25] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2007) Textual Affect Sensing for Sociable and Expressive Online Communication. In: *The 2nd international conference on Affective Computing and Intelligent Interaction*. Lisbon, Portugal. Springer-Verlag, 218 – 229

[26] Grobelnik, M. and Mladenic, D. (2004) Text-Mining Tutorial. In: *Learning Methods for Text Understanding and Mining*. Grenoble, France.

[27] Esuli, A. (2008) *Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*. Ph.D., UNIVERSITÀ DI PISA.

[28] Miller, G. A. (1995) WordNet: a lexical database for English. *Communications of the ACM*, Vol. 38, 11, pp. 39-41.

[29] Penn_Treebank (1992) The Penn Treebank Project [WWW]. Available from: http://www.cis.upenn.edu/~treebank/ [Accessed Aprail 21, 2012]

 ISBN: 978-1-61208-227-1

Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach

Nafissa Yussupova
Faculty of informatics and robotics
Ufa State Aviation Technical University
Ufa, Russian Federation
yussupova@ugatu.ac.ru

Diana Bogdanova
Faculty of informatics and robotics
Ufa State Aviation Technical University
Ufa, Russian Federation
dianochka7bog@mail.ru

Maxim Boyko
Faculty of informatics and robotics
Ufa State Aviation Technical University
Ufa, Russian Federation
russian_max@inbox.ru

***Abstract*—This paper considers the problem of Sentiment classification in text messages in Russian with using Machine Learning methods - Naive Bayes classifier and the Support Vector Machine. One of the features of the Russian language is using of a wide variety of declensional endings depending on the declination, tenses, grammatical gender. Another common problem of sentiment classification for different languages is that different words can have the same meaning (synonyms) and thus give equal emotional value. Therefore, our task was to evaluate on how the lemmatization affects the sentiment classification accuracy (or another, with endings and without them), and to compare the results for Russian and English languages. For evaluating the impact of synonymy, we used the approach when the words with the same meaning are grouping into a single term. To solve these problems we used lemmatization and synonyms libraries. The results showed that using lemmatization for texts in Russian improves the accuracy of sentiment classification. On the contrary, the sentiment classification of texts in English without using lemmatization yields better result. The results also showed that the use synonymy in the model has a positive influence on accuracy.**
In the "Introduction", we describe a place Sentiment Analysis in Data Mining. In the "Approaches to the Sentiment Analysis", we tell about the main approaches of Sentiment Analysis: linguistic approach, an approach based on Machine Learning, and their combination. In the "Description of algorithms for Sentiment Analysis", we state the problem of sentiment classification and describe methods for solving it using a Naïve Bayesian classifier, Bagging, Support Vector Machine. In the "Results of experiments", we describe aims of the experiment and the features of the implementation of the algorithm and report the results of the experiment. In the "Conclusion", we present the output from the results.

Keywords-text analysis; analysis of tonality; sentiment analysis; machine learning.

I. INTRODUCTION

The present stage of human development is characterized by rapid growth of information. One of the most common forms of storage is the text in natural language. Textual form of information is natural for human beings and they readily accept it. The development of information technologies is accompanied by intense growth in the number of websites, which currently stands at more than 285 millions, and as a consequence of increasing the volume of text data. The vast amount of information collected in numerous text databases that are stored in personal computers, local and wide area networks. Average user is becoming more difficult to work with huge amounts of data. Reading the texts of the volume, manual search and analysis of relevant information in giant arrays of text data are ineffective. To solve this problem and to automatically process the information, many developments were done in the areas of natural language processing, information retrieval, machine translation, information extraction, sentiment analysis and others.

The article is devoted to the Sentiment Analysis of Russian text messages using Machine Learning [19]. Sentiment Analysis in the text is one of the directions in the analysis of natural language texts. Sentiment is the emotional score, which is expressed in the text. It can have one-dimensional emotive space (two classes of sentiments) or multivariate (more than two). Foresight sentiment of the text lies in the fact that based on textual information, it allows you to evaluate the success of the campaign, political and economic reforms, to identify relevant press and media to a certain person, to an organization for the event, to determine how consumers relate to a particular product, to services to the organization. In [1], Boyko et al. consider applying Sentiment Analysis to the study opinions of consumer of different banks.

Despite the promise of this direction, while it is not as actively used in text processing systems. The reasons are the difficulties of highlight the emotional vocabulary in the texts, a imperfection of the existing text analyzers, dependence on the domain. Therefore, the improvement and development of new analytical methods based on machine learning is an urgent task.

 ISBN: 978-1-61208-227-1

The article presents the results of a study of Sentiment classification of texts in Russian with using Machine Learning.

II. APPROACHES TO THE SENTIMENT ANALYSIS

There are three approaches of Sentiment Analysis of text messages.

1) Sentiment Analysis based on pre-defined dictionaries of tonality with linguistic analysis. Tonality dictionaries consist of elements such as words, phrases, patterns, each of which has its own emotional coloring. Tonality of the text is determined by the combination of emotive language found and evaluated in text.

2) Sentiment Analysis based on methods of Machine Learning. The text presents in vector form; the classifier is trained according to the available training data. After that, it is possible to classify the sentiments in new text message.

3) The combination of the first and the second approaches.

The first approach is rather time-consuming because of the need for a tonality of dictionaries, a list of tonality patterns and the development of language parsers, but it is more flexible. The advantage of this approach is that it allows you to see the emotional vocabulary at the level of the sentence.

In [2], Pazelskaya et al. present an algorithm for Sentiment Analysis based on the tonal dictionaries consisting of several steps: morphological analysis of text mark-up vocabulary lists for the tonality vocabulary, syntactic analysis, and directly determine the tonality. The algorithm can be estimated on the website [3].

In [4], Ermakov et al. developed the following algorithm for estimating the tonality of the text, which includes recognition of the object of tonality, parsing text, selection and classification of propositions that express the tonality, the assessment based on the general tonality of all the tonality propositions.

Abroad, there was an active search to improve the analysis of tonality on the basis of tonality dictionaries and linguistic analysis, e.g , Nasukawa et al. [5]. This work describes the analyzer, which performs the following actions: 1) remove the special terminology of the text, and 2) determine the tonality, and 3) analysis of the associative relationship. The analyzer uses two linguistic system: a dial tonality dictionary and database templates.

The approach is based on using Machine Learning, presupposes the existence of pre-marked-up the training set of data. The purpose of training in Sentiment Analysis is to get the necessary and sufficient rules, which you can use to make a classification of tonality of the new text messages, similar to those that made up the training set. The drawback of algorithms based on Machine Learning is dependence on the quality and quantity of training data. This approach does not allow an in-depth analysis of the text, to identify the object and the subject of tonality.

Machine Learning methods for solving the problem of Sentiment classification of messages are actively developing overseas. In the Russian practice of science are not yet known cases of successful application of Machine Learning to Sentiment Analysis. Therefore, we consider some of the work of foreign authors.

A great contribution to the development of Sentiment Analysis of text messages was done by researchers from Cornell University B. Pang and L. Lee [6], [7], [8]. In 2008, Pang and Lee published the book «Opinion Mining and Sentiment» [6] devoted to modern methods and approaches to Sentiment Analysis in text messages. In [7], a Sentiment classification using Machine Learning was published and they showed that this approach is superior to a simple technique based on the compilation of dictionaries of commonly used positive and negative words. Pang and Lee [8] describe an algorithm that allows us to classify sentiments using only subjective sentences. Objective proposals generally do not have the emotional coloration, but create noise in the data.

O'Keefe and Koprinska [9] consider the problem that from the training data extracts a very large number of terms. The authors describe methods for selecting the most informative terms, and evaluation of their tonality.

To address the shortcomings of the above approaches is used to combine them. Thus, in [10], the method used is based on the extracted lexical rules; training with the participation of man and machine learning are combined into a sentiment classification algorithm.

König and Brill [11], from Microsoft suggest ways to get sentiment patterns using proposed algorithm. The result is achieved through automatic extraction of informative patterns with subsequent evaluation of tonality, combining with Support Vector Machine (SVM).

The combined approach is promising, as it combines advantages of the first two approaches. Here, an important task for the study is to determine how the Linguistic approach and Machine Learning should interact with each other.

III. DESCRIPTION OF ALGORITHMS FOR SENTIMENT ANALYSIS

In this paper, we consider algorithms which are based on using Machine Learning approach. As Machine Learning algorithms we chose a Naive Bayesian classifier [12] and Support Vector Machine [20]. For improving the accuracy of classification, we considered a Meta-Machine Learning algorithm [16] - Bagging for Naive Bayesian classifier.

Mathematically, the problem of classifying of sentiment can be represented as follows. There are two classes - the class of positive messages, c_1 , and class of negative messages, c_2 , (1):

$$C = \{c_1, c_2\}, \quad (1)$$

there is a set of messages (2):

$$D = \{d_1, d_2, ..., d_n\}, \quad (2)$$

and an unknown classification function (3):

$$F : C \times D \to \{0,1\}. \quad (3)$$

We need to build a classifier F' as close to the classification function F as possible. We have a labelled set of messages for learning (4).

$$K \subset C \times D^l, \quad (4)$$

where D^l is learning set of messages.

Feature space in the this problem can be represented using the vector model. Each text message is treated as a set of words ("bag" of words). This view of a text message is presenting as a point in multidimensional space. Points lying close to each other correspond to semantically similar messages. In this model, a sequence of words is ignored. For example, the "a *good book*" and "*the book is good*" is the same. Thus, the message is a "bag" with the words.

A. Naive Bayes classifier

Let us consider Naive Bayesian classifier for sentiment classification problem. Let each message d takes the values from the dictionary V, and is described by a set of words $\{w_1, w_2, \dots, w_n\}$. There is a set of classes $C = \{c_1, c_2\}$, consisting of a class of positive messages and a class of negative messages. We need to find the most probable value of the corresponding class of the set of words (5):

$$c_{NB} = \arg\max_{c_j \in C} p(d = c_j \mid w_1, w_2, \dots, w_n) \quad (5)$$

It is known that the conditional probability of an event can be found using the Bayes theorem (6) [21]:

$$p(d = c_j \mid w_1, w_2, \dots, w_n) = \\ = \frac{p(w_1, w_2, \dots, w_n \mid d = c_j) \cdot p(d = c_j)}{p(w_1, w_2, \dots, w_n)} \quad (6)$$

Then, the expression (5) takes the form (7):

$$c_{NB} = \arg\max_{c_j \in C} \frac{p(w_1, w_2, \dots, w_n \mid d = c_j) \cdot p(d = c_j)}{p(w_1, w_2, \dots, w_n)} \quad (7)$$

From the expression (7), we are interested only in the numerator, because the denominator does not depend from the class. Thus, the denominator is a constant and can be reduced. Assuming conditional independence of attributes, we obtain the expression (8) which is using for classification:

$$c_{NB} = \arg\max_{c \in C} p(w_1, w_2, \dots, w_n \mid d = c) \cdot p(d = c) \quad (8)$$

Naive Bayesian classifier operates under the following assumptions:

- words and phrases in the message are independent from each other;
- do not takes into account the sequence of words;
- do not takes into account the length of the message.

There are two ways to implement a Naive Bayesian classifier – a Bernoulli model [12] and multinomial model [12]. The difference is that in the Bernoulli model is considering the presence of a word in a message. In the multinomial model, the number of occurrences of a word in the text is considered. Table 1 provides an example of a vector notation of the text.

TABLE I. EXAMPLE OF VECTOR FORM

	Vector description
Bernoulli model	[0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0]
Multinomial model	[0, 0, 2, 1, 0, 3, 1, 2, 0, 0, 0]

Let us consider the sentiment classification algorithm with the Bernoulli model, presented by Manning et al. [12]. In the Bernoulli model, the message is described by the vector consisting of the attributes with values 0 or 1. Thus, we consider only the presence or absence of words in the message; then, we ignore how many times it is repeated in the message.

Given a vocabulary $V = \{w_t\}_{t=1}^{|V|}$; then, the message d_i is described by the vector of length $|V|$, consisting of bits b_{it}. If a word w_t appears in the message d_i then $b_{it} = 1$, if not then $b_{it} = 0$. Then, the likelihood of belonging to a class c_j of messages d_i can be calculated by the formula (9):

$$p(d_i \mid c_j) = \prod_{t=1}^{|V|} (b_{it} \cdot p(w_t \mid c_j) + \\ + (1 - b_{it}) \cdot (1 - p(w_t \mid c_j))) \quad (9)$$

For learning a classifier it needs to find the probabilities $p(w_t|c_j)$. Let there be a training set of messages $= \{d\}_{i=1}^{|D|}$, which has labels of classes c_j, then it is possible to calculate estimates of the probabilities that a particular word occurs in a particular class (10):

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} b_{it} \cdot p(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} p(c_j \mid d_i)} \quad (10)$$

A priori probabilities of classes can be calculated by the formula (11):

$$p(c_j) = \frac{\sum_{i=1}^{|D|} p(c_j|d_i)}{|D|} \quad (11)$$

Then, the classification will be carried out by the formula (12).

$$c_{NB} = \arg\max_{c_j \in C} p(c_j) \cdot p(d_i \mid c_j) = \\ = \arg\max_{c_j \in C} [\log(\sum_{i=1}^{|D|} p(c_j \mid d_i)) + \sum_{t=1}^{|V|} \log[b_{it} \cdot \\ \cdot p(w_t \mid c_j) + (1 - b_{it})(1 - p(w_t \mid c_j))]] \quad (12)$$

From (10), it follows that the some probabilities will be zero, since that some words can be presented in one class of training data and can be absent in another. Difficulties arise with zero probabilities when they are multiplied in (12). In this case, the entire expression is zero and there is a loss of information. To avoid zero probability of obtaining used add-one, or Laplace smoothing [12], which consists of adding one to the numerator (13).

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} b_{it} \cdot p(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} p(c_j \mid d_i)} \quad (13)$$

Sentiment classification algorithm using the Bernoulli model is shown in Figures 1 and 2. It consists of learning part and classifying part. In the learning part there are input parameters is a set of labelled messages and set of classes. In this part creates a dictionary of words V, that estimates $p(c_j)$ and $p(w_t|c_j)$, sets the threshold value h which minimize the classification error. Output is a fully trained classifier with set parameters. Classifying part applies for new message, which sentiment must be determined.

In the multinomial model, see Manning et al. [12], the message is a sequence of random selection of some word

from the dictionary. This model takes into account the number of repetitions of each word in a one message, but ignores words that are absence in the message.

Given a vocabulary $V = \{w_t\}_{t=1}^{|V|}$; then, the message d_i can be described by the vector of length $|V|$, consisting of words, which is taken from the dictionary with probability $p(w_t|c_j)$. Then, the likelihood of belonging of messages d_i to a class c_i estimates by formula (14).

$$p(d_i \mid c_j) = p(|d_i|) \cdot |d_i|! \cdot \prod_{t=1}^{|V|} \frac{1}{K_{it}!} p(w_t \mid c_j)^{K_{it}}, \quad (14)$$

where K_{it} - is the number of occurrences of word w_t in the message d_i.

For learning, the classifier also needs to find the probabilities $p(w_t|c_j)$. Let there be a training set of messages $D = \{d\}_{i=1}^{|D|}$, which is distributed in classes c_i and we know the number of occurrences of words in the message K_{it}. Then, we can calculate estimates of the probabilities that a particular word occurs in a particular class (15). In this case, also apply smoothing add-one.

Input: set of documents $D = \{d_1, d_2, \dots, d_n\}$, set of classes $C = \{c_1, c_2\}$

1. Extract all terms from D to the vocabulary V
2. For each $c_j \in C$ do
   3. Count documents N^c in each class
   4. Calculate probability $p(c_j) = N^c/N$
   5. For each $w_t \in V$ do
      6. Count documents $N^c_{w_t}$ in class containing word w_t
      7. Calculate probability $p(w_t|c_j) = (N^c_{w_t} + +1)/(N^c + 2)$
8. Set threshold h with minimal classification error

Output: V, $p(c_j)$, $p(w_t|c_j)$, h

Figure 1. Algorithm of learning NB Bernoulli model

Input: document d, V, $p(c_j)$, $p(w_t|c_j)$, h

1. Extract all terms from d to the vocabulary V_d
2. For each $c_j \in C$ do
   3. score$[c_j] = \ln p(c_j)$
   4. For each $w_t \in V$ do
      5. If $w_t \in V_d$ than score$[c_j]$ += $\ln p(w_t|c_j)$
      6. else score$[c_j]$ += $\ln(1 - p(w_t|c_j)$
7. If $score[c_1] > h \cdot score[c_2]$ than $d \in c_1$ else $d \in c_2$

Output: tonality of document d

Figure 2. Algorithm of classification NB Bernoulli model

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} K_{it} \cdot p(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} K_{is} \cdot p(c_j \mid d_i)} \quad (15)$$

A priori probabilities of classes can be calculated by the formula (16).

$$p(c_j) = \frac{\sum_{i=1}^{|D|} p(c_j \mid d_i)}{|D|} \quad (16)$$

Then, the classification will be carried out by the formula (17).

$$c_{NB} = \arg\max_j \; p(c_j) \cdot p(d_i \mid c_j) = \\ = \arg\max_j [\log(\sum_{i=1}^{|D|} p(c_j \mid d_i)) + \\ + \sum_{t=1}^{|V|} K_{it} \cdot \log \; p(w_t \mid c_j)] \quad (17)$$

Classification algorithm with the Multinomial Naïve Bayes model is shown in Figures 3 and 4. It consists of learning part and Sentiment classification part. In the learning part creates a dictionary of terms V, estimates probabilities $p(c_j)$ and $p(w_t|c_j)$, set the threshold value of h, to minimize the classification error. Classifying part applies for new message, which sentiment must be determined.

B. Bagging algorithm

One of the algorithms for improving the quality of classification is called Bagging. It was proposed by L. Breinman and describes in [16], Breinman. Bagging algorithm is shown in Figure 5.

From the initial training set of D of length $|D|$ forms training subsets D_t of the same length $|D|$ with the bootstrap - a random selection with returns. However, some messages will appear in a subset of a few times, some - not even once. Next, set the control messages by subtracting D/D_t. With using training subset D_t learns classifier h_t. Classification error e_t of h_t estimates by the control subset D/D_t and then compared with the admissible error of the classification of e. If the error is less than a classifier built admissible error, then it is added to the ensemble. Sentiment classification is produced with the ensemble of classifiers by a simple voting.

C. Support Vector Machine

The main idea of Support Vector Machine algorithm is to find separating hyperplane, represented by vector $\overline{w}$ which minimize empirical error of classification and maximize margin between classes. SVM was proposed by V. Vapnik, C. Cortez and A. Chervonenkis [20]. SVM is a high effective in classification problems and has popularity among Machine Learning algorithms. In particular, it outperforms other algorithms of Machine Learning in text categorization. The finding of separating hyperplan corresponds to a constrained quadratic optimization

problem. Let $c_j \in \{1, -1\}$ be the class of document d_j; then, the solution can be written as (18):

$$\overline{w} = \sum_j \alpha_j c_j \overline{d}_j, \quad \alpha_j \geq 0 \qquad (17)$$

where α_j are obtained by solving a dual optimization problem. Those $\overline{d}_j$ such that α_j is greater than zero are called support vectors, since they are the only document vectors contributing to $\overline{w}$. Classification of message consists of determining which side of $\overline{w}$ hyperplan it fall on.

The main disadvantage of Support Vector Machine is that it has cubic complexity in the size of dataset and requires a lot of computational resources. The cause is that it have to solve quadratic optimization problem with the number of parameters equal to number of data and to compute dot product many times.

There are many modifications of SVM developed for reducing computational time. One of them is Sequentional Minimal Optimization algorithm [17], Platt. This algorithm is used in this work. It allowed receiving the. It allowed receiving the results in acceptable time.

In this work realized to variants of SVM – first variant considers only a presents/absence of features and in the second variant considers the number of occurrences of features.

Input: set of documents $D = \{d_1, d_2, \dots, d_n\}$, set of classes $C = \{c_1, c_2\}$

1. Extract all terms from D to the vocabulary V
2. For each $c_j \in C$ do
   3. Count documents N^c in each class
   4. Calculate probability $p(c_j) = N^c/N$
   5. For each $w_t \in V$ do
      6. Count number of occurrences $M^c_{w_t}$ of word w_t in each class
      7. Calculate probability $p(w_t|c_j) = (M^c_{w_t} + 1)/\sum_{t=1}^{|V|}(M^c_{w_t} + 1)$
8. Set threshold h with minimal classification error

Output: V, $p(c_j)$, $p(w_t|c_j)$, h

Figure 3. Algorithm of learning NB Multinomial model

Input: document d, V, $p(c_j)$, $p(w_t|c_j)$, h

1. Extract all terms from d to the vocabulary V_d
2. For each $c_j \in C$ do
   3. score$[c_j] = \ln p(c_j)$
   4. For each $w_t \in V$ do
      5. If $w_t \in V_d$ than $score[c_j] \mathrel{+}= \ln p(w_t|c_j)$
6. If $score[c_1] > h \cdot score[c_2]$ than $d \in c_1$ else $d \in c_2$

Output: tonality of document d

Figure 4. Algorithm of learning NB Multinomial model

Input: $(d_1, y_1), (d_2, y_2), \dots, (d_m, y_m)$, where $d_i \in D$ – set of documents;

$y_i \in Y = \{+1, -1\}$, $y_i = +1$ if positive and $y_i = -1$ negative polarity of documents;

T – number of classifiers in ensemble;

$|D|$ - length of training data;

e – admissible error of classification

1. For $t = 1$ *to* T do
   2. Choose randomly $|D|$ documents for constructing training dataset D_t from D
   3. Construct control dataset from D/D_t
   4. Construct classifier $h_t: D \to \{+1, -1\}$
   5. estimate error e_t of classifier h_t on dataset D/D_t
   6. If $e \geq e_t$ then add classifier h_t into ensemble

Output: $H(d) = sign(\sum_{t=1}^{T} h_t(d))$ final ensemble of classifiers

Figure 5. Bagging algorithm

IV. RESULTS OF EXPERIMENTS

In this research, we aimed at studying a few points:

- Evaluate the performance for Sentiment classification of text messages in Russian language;
- Compare the performance with results obtained for text messages in English language;
- Study the influence of lemmatization on the accuracy of classification;
- Study the influence of a length of word on the accuracy of classification, and
- Study the influence of the grouping words, which have equal semantic meaning on the accuracy of classification.

According these aims, a program was developed «Text Analyzer» in the programming language C#. All listed algorithms of Sentiment classification were realized in this language.

For learning and evaluation of the accuracy of the sentiment classification, we used the test set, consisting of customer reviews of a few Russian banks taken from the Internet site [13]. It includes 304 positive reviews and 850 negative reviews in Russian. An example of review with a positive sentiment is: "An application for a loan designed to quickly, no questions asked, within 20 minutes." An example of negative review: "Consideration of the application took time for two months".

For evaluation of Sentiment classification for a text in English, we used dataset that includes 1000 negative and 1000 positive reviews about films from IMDB [18].

For study of influence of lemmatization in the pre-processing text, lemmatization of all occurring words was entered. Lemmatization brings different words to their initial form; for example, the noun is the nominative case, singular. Motivation for lemmatization of the text is due to the fact that different forms of a word can often express the same meaning. In this regard, is justified to bring the words to a initial form. We used LemmaGen library written in C#

and designed for lemmatization of words. These libraries are available on the website of the developer [14].

To evaluate the generalization capability of the algorithm used by a sliding control or cross-validation we proceeded as follows. Fixed set consisting of 10 partitions of the original sample, each of which in turn consisted of two subsamples: the training and control. For each partition, configures the algorithm for the training subsample, and then evaluated its average error on the objects of the control subsample. Assessment of the sliding control was averaged over all partitions of the error on the control subsamples; for the bagging algorithm accepted allowable error of the classifier is equal to e = 25%.

To evaluate the classification accuracy of each control unit we used the indicator "classification accuracy", which is calculated by the formula (18):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%, \qquad (18)$$

where TP - the number of correctly classified positive messages; TN - the number of correctly classified negative messages; FP - the number of non correct classified positive messages; FN - the number of non correct classified negative messages.

The results of computational experiments are presented in the Table 2. Accuracy of classification lies in range 85% - 88,3%. For Naïve Bayes, the best results obtained by Multinomial model 86,83% (Bernoulli – 86,49%) or another words by considering number of occurrences of words. Using Bagging algorithm has a positive influence on the classification. It improved accuracy for NB Multinomial model by 0,86%.

For Support Vector Machine with leaner core, the best results obtained by considering presence of word in the message (87,69% vs. 85%). Using of polinomial core gave 86,73% of accuracy.

For Sentiment classification, we received better results on dataset in Russian on an average 5%. This suggests that dataset in Russian is a more constrained domain - banking. In contrast, dataset in English has a wide range of different words, because most reviews have a description of film story. Analysis of results also indicates that SVM outperforms Naïve Bayes algorithm in two cases of language.

For grouping words that have the same semantic meaning, we used a vocabulary of synonyms with 5371 strings. For example, if in message occur to different words “borrow” and “lend” then it is equivalent occurring two words “borrow”. This modification allowed to improve accuracy by 0,1% for NB, and 0,08% for SVM. It is not so much but we hope that using more bigger and specific vocabulary of synonyms can give a more significant effect.

Lemmatization has positive influence for Sentiment classification of text in Russian (87,07% without lemmatization, 87,69% with lemmatization). It could be explained that in Russian language words could have different endings. Lemmatization allows to group cognate words with one semantic meaning and different endings. In text in English the best result received without lemmatization (84,3% vs. 85,85%).

TABLE II. RESULTS OF EXPERIMENTS WITH TEXTS IN RUSSIAN

Naïve Bayes classifier	
NB Bernoulli model	86,49%
NB Multinomial model	86,83%
NB Multinomial model with synonyms	86,93%
NB Multinomial model, length > 2	86,40%
Bagging NB Bernoulli model (e=25%)	86,82%
Bagging NB Multinomial model (e=25%)	87,69%
Support Vector Machine	
SVM, presence, leaner	87,69%
SVM, occurance, leaner	85,00%
SVM, presence, leaner, without lemmatizator	87,07%
SVM, presence, leaner, length > 2	88,21%
SVM, presence, leaner, with synonyms	87,77%
SVM, presence, polinomial	86,73%
SVM, presence, leaner, with synonyms, length > 2	88,30%

TABLE III. RESULTS OF EXPERIMENTS WITH TEXTS IN ENGLISH

Naïve Bayes classifier	
NB Bernoulli model	80,25%
NB Multinomial model	81,05%
Support Vector Machine	
SVM, presence, leaner	84,3%
SVM, occurance, leaner	83,15%
SVM, presence, leaner, without lemmatizator	85,85%

By excluding prepositions and articles from feature words, the experiment considered words with length more than two letters. This modification gave a better result, with 0,52% in SVM. But, in NB accuracy descended on 0,43%. The best result of 88,30% as obtained by SVM with lemmatization, grouping synonyms and length of word > 2.

V. CONCLUSION

Based on results of Sentiment classification of texts in Russian, we obtained the following conclusions:

- Machine Learning could provide accuracy of sentiment classification 85% - 88,3% for considered texts in Russian;
- SVM confirmed that it outperforms Naïve Bayes algorithm in two cases of language;
- Multinomial model surpasses Bernoulli model in NB;
- Bagging algorithm has a positive influence on the classification but little;
- presence feature of words surpasses number of occurance in SVM;
- using synonyms has positive influence on Sentiment classification but little;
- lemmatization has positive influence for Sentiment classification of text in Russian, but not for text in English.

 ISBN: 978-1-61208-227-1

The task of sentiment classification of text messages has a complex nature and requires innovative approaches for solution. The complexity of its nature is that the initial data are the texts in natural language. Every word of this text has its meaning, and the combination of words is a complex interaction of the meaning of each word. At present there is no universal method of modeling such an interaction in the language of the machine or the language of numbers.

Despite the complexity of the problem, it attracts a large number of researchers around the world. Searches in this area are actively maintained and there are some achievements. Many of the developed algorithms achieve classification accuracy greater than 85%. But keep in mind that these results were obtained on test data under experimental conditions. Unfortunately, there is no official information about the real successful practical application of systems to solve such problems.

REFERENCES

[1] M. Boyko, A. Hilbert, N. Yussupova, and D. Bogdanova, “Marketing research of cosumer opinions with using information technologies”, Proceedings of the 13th International Workshop on Computer Science and Information Technologies, Germany, Garmisch-Paterkirchen, September 27 - October 02, 2011, pp. 103-105.

[2] A.G. Pazelskaya and A.N. Soloviev, “Method of the determination emotions in the lyrics in Russian”, Computer program linguistics and intellectual technologies, Issue 10 (17), 201, pp. 510-522.

[3] The official site of "Эр Си О". On-line sentiment classification: http://x-file.su/tm/Default.aspx [retrieved: October, 2012].

[4] A.E. Ermakov and S.L. Kiselev, “Linguistic model for the computer analysis of key media of publications”, Computational Linguistics and the intellectual technology: proceedings of the International Conference Dialog'2005, 2005, pp. 172-177.

[5] J. Yi, T. Nasukawa, W. Niblack, and R. Bunescu, “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques”, In Proceedings of the 3rd IEEE international conference on data mining, ICDM 2003 , pp. 427-434.

[6] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval”, Vol. 2, 2008, pp. 1-135.

[7] B. Pang and L. Lee, “Thumbs up? Sentiment Classification using Machine Learning Techniques”, Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[8] B. Pang and L. Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Mini-mum Cuts”, Proceedings of the ACL, 2004, pp. 271-278.

[9] T. O'Keefe and I. Koprinska, “Feature selection and weighting methods in sentiment analysis”, Australasian Document Computing Symposium, 2009, pp. 142-153.

[10] R. Prabowo and M. Thelwall, “Sentiment analysis: A combined approach”, Journal of Informatics in 2009, pp. 143-157.

[11] A. König and E. Brill, “Reducing the Human Overhead in Text Categorization”, Proceedings of KDD, 2006, pp. 598-603.

[12] C. Manning, P. Raghavan, and H. Schuetze, “An Introduction to Information Retrieval”, Cambridge University Press. Cammbridge, England, 2009, pp. 1-544.

[13] Internet portal dedicated to the Russian banks: http://banki.ru [retrieved: October, 2012].

[14] Portal dedicated to lemmatization and stemming: http://lemmatise.ijs.si/Software/Version3 [retrieved: October, 2012].

[15] Y. Freund and R. Schapire, “Experiments with New Boosting Algorithm”, Machine Learning: Proceedings of the Thirteenth International Conference, 1996, pp. 148-156.

[16] L. Breinman, “Bagging Predictors”, Machine Learning, 24, 1996, pp. 123-140.

[17] J. Platt, “Fast training of support vector machines using sequential minimal optimization”, Advances in Neural Information Processing, Vol. 12, 2000, pp. 547-553.

[18] Datasets for tests: http://www.cs.cornell.edu/People/pabo/movie-review-data/ [retrieved: October, 2012]

[19] R. Michalski, J. Carbonell, and T. Mitchell, “Machine Learning: An Artificial Intelligence Approach”, Tioga Publishing Company, ISBN 0-935382-05-4, 1983, pp. 83-138.

[20] C. Cortes and V. Vapnik, “Support-Vector Networks”, Machine Learning, 20, 1995, pp. 273-297.

[21] T. Bayes and R. Price, “An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S”, Philosophical Transactions of the Royal Society of London 53, 1763, pp. 370-418.

 ISBN: 978-1-61208-227-1

A Novel Dependability Model to Define Normal Network Behavior

Maher Salem
Network and Data Security,
Applied Computer Sciences
University of Applied Sciences Fulda
Fulda, Germany
Maher.salem@informatik.hs-fulda.de

Ulrich Buehler
Network and Data Security,
Applied Computer Sciences
University of Applied Sciences Fulda
Fulda, Germany
u.buehler@informatik.hs-fulda.de

***Abstract*—Computer networks augment in heterogeneity so that defining a normal behavior to the network becomes a severe challenge. Particularly, such a normal network behavior is essential for security issues. In addition, this behavior consolidates the intrusion detection system to significantly detect zero-day-attacks. Therefore, in this paper, we introduce a novel dependability model based on the correlation matrix of network features. Moreover, only strongly correlated features are involved in the model such that the normal connections are recognized into the online traffic in advance. The recognition is based on the distance of the incoming traffic to the linear association between the correlated features. Furthermore, the distance is compared to a threshold value to ensure correct recognition. These steps have been evaluated by the benchmark dataset NSL-KDD. The goal of this model is to build an adaptive normal network behavior that represents the intended network continuously, reduces the overhead on the classification, and supports by detecting unknown attacks respectively. The results show that the idea of dependability model in intrusion detection system promises more accuracy and preciseness in anomaly detection.**

Keywords-correlation matrix; dependability; normal network behavior; linear association.

I. INTRODUCTION

Intrusion Detection Systems (IDS) approaches can be classified into misuse detection and anomaly detection [1], misuse detection systems are using signatures to detect known attack patterns. However, they are suffering under the constantly growing number of signatures and incapability of detecting unknown attacks as well. In contrast, anomaly based intrusion detection systems are able to detect known and even unknown attacks by recognizing the deviation from the normal network behavior. Accordingly, there are two main approaches to characterize normal network behavior presented in [2], which are studying the inference of the overall network behavior through the use of network probes and the understanding of the behavior of the individual entities or nodes. Principally, IDS analyzes and studies the network traffic to establish a profile that defines a normal network behavior (NNB). Upon this profile, the IDS can detect any deviation as an anomaly and consider it most likely an attack. Presenting a significant and heuristic model that defines the normal behavior is imperative the area of networking. Therefore, we present a novel model that defines a NNB by building a dependability model from the strength of features correlations. The main idea is to capture the online traffic in the real time and match the traffic to the dependability model to investigate its normality. Moreover, for the positive strong correlations between two features a linear association is defined to the best fit of the concerned features. This is to imply that, when two feature vectors are strongly correlated there is significant of determination that ensures a linear association between these features. Thus, such relation can be exploited to determine the normality in the incoming traffic based on the linearity of correlated features. The dependability model can then be updated with the new normal traffic. This paper is structured as following; in section II, a motivation about the proposed methods in normal network behavior is discussed. On the hand, section III presents our novel methodology. Section IV describes the preparation of dataset. Section V illustrates our results and discussion. Finally, section VI concludes our work.

II. MOTIVATION

Intrusion detection system steps are summarized into feature selection, discretization, normalization, and classification. Regarding classification the IDS builds a normal profile of the network and detect the deviation from this profile.

A real time visualization platform in [3] presents a multiple visualization techniques that provide a situational understanding of real time network activity. Such platform can visualize million of records and report the network current status. However, it may not feasible in IDS research area. On the other hand, [4] proposed a modeling approach where failures and repairs of network components as well as routing and traffic handling are described by a set of stochastic activity networks (SAN). The proposed model approach serves more in the area of network routing and availability of end-to-end network components. However, it could be exploited to build a normal network behavior.

A significant work [5] explained a correlated node behavior model based on Markov process [19] for dynamic topology network. Thus, the latter classified the nodes into four categories and show that the effect from correlated failure nodes on network survivability is much severer than

 ISBN: 978-1-61208-227-1

other misbehaviors. However, this approach is investigating network nodes and builds a behavioral model accordingly. A reasonably network behavior tool in [6] exploits only the internal network traffic to monitor the internal activities on network so that a deviation from a predefined pattern model is detected as abnormal behavior. This model aims to detect anomaly indeed but it examines only the internal traffic.

A structural model in [7] utilizes web logs to analyze user behavior based on the web-context and situation-awareness. Obviously, this model focused only on the user activities and ignores the network ones. More sufficient proposals regarding analysis of system behavior are proposed in [8],[9], and [10].

So, defining a novel model for normal network behavior is needed. Therefore, we principally focus on the network traffic to build such a model and express it as linear relations.

III. PROPOSED METHODOLOGY

In this research work, we concentrate on the definition of a normal network behavior, based on its traffic, which represents the network. In contrast to [14], we build a dependability model graph based on the correlation matrix to define a normal network behavior and to predict the normal connection in real time. Thus, the proposed idea in this work exploits network traffic statistics to build a dependability model from the feature correlations; that is, from correlation matrix. In addition, the model will be able to detect normal connections based on the linear association between the correlated features. However, selecting the most valuable features out of hundreds of network features is a provocation step. Hence we declare the proposed idea in three steps:

A. Significant Network Features

Selection of the valuable features in the area of IDS is a negotiable point in data mining research. Thus, we used the improved feature selection method proposed in [11]. It presented a novel method that abstracted the valuable features in the network based on the sequential backward search and information gain. The difference between both feature sets is that features in the most valuable feature set affect definitely the detection rate, whereas features in the most valuable and relevant feature set affect definitely the detection rate and enhance it, i.e. $MVF \subset MVRF$. Moreover, the model has been evaluated on the benchmark dataset NSL-KDD [18]. The exploited features are summarized in Table I.

TABLE I. MOST VALUABLE FEATURE SET AND MOST VALUABLE AND RELEVANT FEATURE SETS

Name of feature set	**features**
Most Valuable Features (MVF)	service, src_bytes, dst_host_serror_rate, serror_rate, dst_host_srv_diff_host_rate, protocol_type,rerror_rate,srv_rerror_rate,wrong_fragment,num_compromised,num_access_files
Most Valuable and Relevant Features (MVRF)	service, src_bytes, diff_srv_rate, same_srv_rate, dst_host_srv_count, logged_in, dst_host_serror_rate, serror_rate, srv_serror_rate, dst_host_srv_diff_host_rate, protocol_type, rerror_rate, srv_rerror_rate, hot, wrong_fragment, num_compromised, num_access_files, root_shell, num_failed_logins

In principal, we build the dependability model from the strongly correlated features of these feature sets and hence define a normal behavior for the network.

B. Correlation and Dependability Model

Network traffic has several features, which are somehow sharing an association. One of the most known methods to infer these associations is the correlation between features; that is, the correlation is used to determine the degree of association between two features [12]. Hence, let us define two network features (vectors) X and Y, which are normally distributed, such that $X=\{x_1,..., x_n\}$ and $Y=\{y_1,...,y_n\}$ where $n \in \mathrm{N}$, $x_i, y_i \in \mathfrak{R}$. Then the Pearson's correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y})}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \quad (1)$$

where $\bar{x}, \bar{y}$ are the mean values of feature X and feature Y, respectively. The correlation value between two features falls between [-1,+1], so that the more positive the value, the more significant the linear association. Thus, the correlation matrix established for m network features $F_1,..,F_m$ can be built as

$$\mathbf{Corr}_{F1,..,Fm} = \begin{bmatrix} r_{11} & \dots & r_{1m} \\ \dots & \dots & \dots \\ r_{m1} & \dots & r_{mm} \end{bmatrix} \quad (2)$$

The correlation matrix is a symmetrical one, and the correlation value of the same feature is always +1. Other values of r_{ij} could be positive or negative, e.g. 0.8 means that 80% of the changes in one feature are related to the other. On the other hand, the coefficient of determination R^2 (or r_{ij}^2) of the two features F_i *and* F_j means that the percentage of variability in one feature related to variability to the other feature. In addition, R^2 gives the proportion of the variance of one feature explained by the other [13], e.g., if the value of the coefficient determination is 0.8 that indicates about 80% of the variance of one variable is explained by the other. Furthermore, it ensures about the prediction of the feature $\bar{y}$ (predicted y) from the linear association instead of the mean value. Therefore, we consider the value of coefficient of determination to assure linear association between features. Consequently, only the strongly correlated features, which reject the null hypothesis $H_0 : \rho = 0$, are considered in this research work. The Greek symbol rho is the parameter used for nonlinear correlation. However, the null hypothesis is the most common used with Pearson's correlation coefficient [20] such that the

correlation coefficient is zero and there is no linear association between the two variables. In this regard, we use the critical P-value with 0.05 of making error type 1 to check whether the correlation value between two features rejects the null hypothesis and have a linear association as well or not. Accordingly, we abstract precisely only the significant linear association between the strong correlated features.

C. Linear Association and Prediction

If two network features are strongly correlated, then they have a linear association that describes their correlation. The linear association between two strongly correlated features $X=\{x_1,..., x_n\}$ and $Y=\{y_1, ...,y_n\}$ can be defined as

$$y_i = \omega_0 + \omega_1 x_i + \varepsilon_i \quad (3)$$

where ω_0 is the intercept and ω_1 is the slope. The idea of least squares is exploited to find the choice of slope and intercept that give the best fit among the data points. In addition, the parameter ε_i is the normally distributed random error. In this research paper we abstract the linear line to the best fitting of the scatter data, i.e., the association is definitely not 100% linear, so that a percentage of error in the linearity and prediction is expected as well. Hence, suppose we have a pair (x_i,y_i) that is not fitting exactly on the linear line, so we can determine the distance of the point to the line as in [17], such that

$$d = \frac{|y_i - \omega_1 x_i - \omega_0|}{\sqrt{\omega_1^2 + 1}} \quad (4)$$

The distance from the linear line will be used to check if the incoming online traffic belongs to the linearity between the correlated two features or not based on a certain threshold, mainly the maximum distance d_{max}.

In brief, we select the valuable and significant network features, infer the correlation values between them, establish a correlation matrix, indicate the strong positive correlation values via rejecting the null hypothesis, find out the best fitting linear line between each two correlated features, and then detect, for the online traffic, the normal connections based on the distance from the linear line. Finally, detected normal connections will be used to rebuild the model.

For example, suppose we have three network features namely F_1, F_2, and F_3 and the values belonging to these features are shown in Table II.

TABLE II. FEATURE VALUES

F_1	F_2	F_3
0.1	0.01	0.3
0.2	0.03	0.46
0.9	0.09	0.37
0.3	0.035	0.08
0.6	0.063	0.011
0.71	0.073	0.93

According to the correlation coefficient in equation (1) the correlation coefficient matrix between these features is

$$\mathbf{Corr}_{F1,F2,F3} = \begin{bmatrix} 1 & 0.9936 & 0.2674 \\ ... & 1 & 0.2750 \\ ... & ... & 1 \end{bmatrix} \text{ P-Value} = \begin{bmatrix} 1 & 0.009 & 0.5678 \\ ... & 1 & 0.449 \\ ... & ... & 1 \end{bmatrix}$$

Both matrices are symmetric so that the lower region with (…) is the same as the upper region. Obviously, the correlation between F_1 and F_2 has a P-value with 0.009, i.e., the P-value<0.05 so that the null hypothesis is rejected at 5% significant level. Whereas, other correlation values have P-values greater than 0.05 which imply that no linear association is existed between them. Therefore, only one linear association is existed with an intercept of 0.0058 and a slope value of 0.095. Moreover, the maximum distance from the pairs to the line is 0.0053. Figure 1 shows the plot of this example.

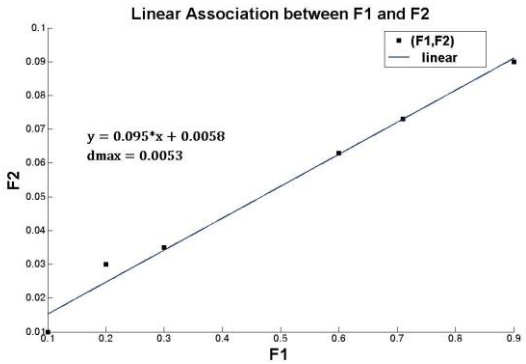


Figure 1. Linear Association between the features in the example.

Accordingly, if we receive a new feature instance in real time such that the values of F_1, F_2, and F_3 are {0.2, 0.1, 0.1} respectively then we need to calculate the distance between the pair (0.2,0.1) and the linear line then compare it to d_{max} value.

IV. DATASET PREPARATION

To evaluate our proposed method, we build a test computer network, so that only clean traffic will be aggregated. Hence, no external connections are allowed and no any USB devices will be plugged. We aggregate the traffic based on the feature set MVRF in Table I. On the other hand, a dataset with only the normal traffic from NSL-KDD is generated to establish a correlation matrix from the positively correlated features and those have linear association. Furthermore, from the correlated features a dependability model will be designed so that only features with linear association are conducted. According to the test network, we could abstract up to 15 features from the MVRF and still struggling to achieve the rest ones. Therefore, we will first test and evaluate our proposal with the normal traffic from NSL-KDD and discuss our results.

Generally, we cannot just filter a dataset out from NSL-KDD and calculate a Pearson's correlation coefficient, but we should digitize it and then normalize it. To achieve such numeric and normalized dataset we exploit the hybrid normalization method in [15] to map the nominal values into numeric and then normalize the dataset using minimum maximum normalization. Thus, a minimum maximum normalization is defined as

 ISBN: 978-1-61208-227-1

$$nv = f(v) = \frac{v - \min(v)}{\max(v) - \min(v)} \quad (5)$$

where $f: \Re \rightarrow [0,1]$ be the normalization function and $v \in \Re$ the numerical value of a feature in the feature sets, nv the normalized feature value after normalization process.

V. RESULTS AND DISCUSSION

To evaluate the novelty of the proposed method, we exploited the NSL-KDD, so that only normal traffic is abstracted and then the dataset is normalized. The selected features in our evaluation are the MVRF. Hence, a dataset with 65555 normal instances is initialized for testing and evaluation. In the following table these features are numbered to ease the explanation of our results.

TABLE III. SELECTED FEATURES IN MVRF

Feature Set	Feature number.feature name
Most Valuable and Relevant Features (MVRF)	**1**.Protocol_type, **2**. Service, **3**.scr_bytes, **4**.wrong_fragment, **5**.hot, **6**.num_failed_logins, **7**.logged_in, **8**.num_compromised, **9**.root_shell, **10**.num_access_files, **11**.serror_rate, **12**.srv_serror_rate, **13**.rerror_rate, **14**.srv_rerror_rate, **15**.same_srv_rate, **16**.diff_srv_rate, **17**.dst_host_srv_count, **18**.dst_host_srv_diff_host_rate, **19**.dst_host_serror_rate

We developed a Matlab program to calculate the correlations between features and the coefficient of determination. Due to space limitation we present a small part of the Correlation matrix and significant determination as well.

$$\mathrm{Corr}_{F1,..F19} = \begin{bmatrix} 1 & 0.5692 & ... & 0.0523 \\ ... & 1 & ... & -0.1318 \\ ... & ... & 1 & ... \\ ... & ... & ... & 1 \end{bmatrix} \quad R^2 = \begin{bmatrix} 1 & 0.3240 & ... & 0.0027 \\ ... & 1 & ... & 0.0174 \\ ... & ... & 1 & ... \\ ... & ... & ... & 1 \end{bmatrix}$$

According to the evaluated dataset MVRF, we calculate the correlation between 19 features, see Table III. Then determine the coefficient of determination, so that only the best linear association between features is considered. Table IV shows the positive correlated features from the correlation matrix, so that the correlation rejects the null hypothesis, (also, the linear equation associated between these features).

TABLE IV. STRONGLY POSITIVE CORRELATED FEATURES IN MVRF

Correlated features	r	R^2	Linear line
1↔2	0.5479	0.302	$y_i = 0.008 + 0.52x_i + \varepsilon_i$
1↔7	0.7905	0.625	$y_i = 0.23 + 1.4x_i + \varepsilon_i$
2↔7	0.5189	0.270	$y_i = 0.36 + 0.98x_i + \varepsilon_i$
11↔12	0.8748	0.765	$y_i = 0.003 + 0.8x_i + \varepsilon_i$
13↔14	0.9829	0.966	$y_i = 0.0013 + 0.98x_i + \varepsilon_i$
2↔17	0.7070	0.5	$y_i = 93 + 2.7x_i + \varepsilon_i$
15↔16	0.7620	0.58	$y_i = 0.77 - 0.77x_i + \varepsilon_i$

In addition, these equations represent the network traffic when the network is in its normal behavior. Of course, one of the drawbacks of NSL-KDD is that there is no 100% linearity between the correlated features. Therefore, we expect an error (false positive) when detecting the incoming online traffic based on these equations. Figure 2 shows a linear relation between two features. Based on this figure, most of the data fit on the linear line and other are on the area around, so we can consider the points on the line or nearby are only the related ones to this equation. To do so, a maximum distance must be determined and an error value must be defined such as $d_{max} \leq 0.5$ and $\varepsilon_i \approx 0.0005$ where d_{max} is determined from $\{(Point_{max} - Point_{min})/2\}$ and because the dataset is normalized then $Point_{max} = 1$ *and* $Point_{min} = 0$. We found after testing several cases that the maximum distance is the average between the minimum point and maximum point from the normalized dataset. On the other hand, the error is selected manually to a small value to avoid incorrect distance calculation.

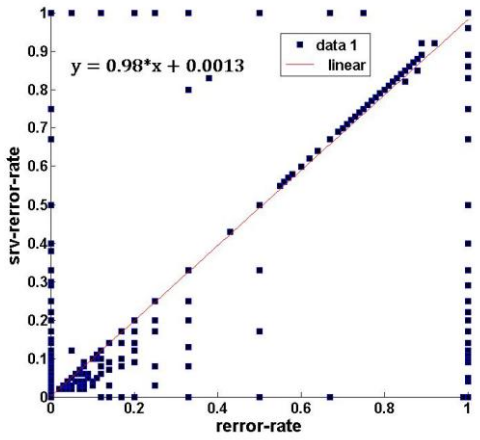


Figure 2. Linear Association between rerror_rate and srv_rerror_rate.

In contrast, Figure 3 shows more stable linear association between the feature *protocol_type* and *logged_in* so that a better detection is expected.

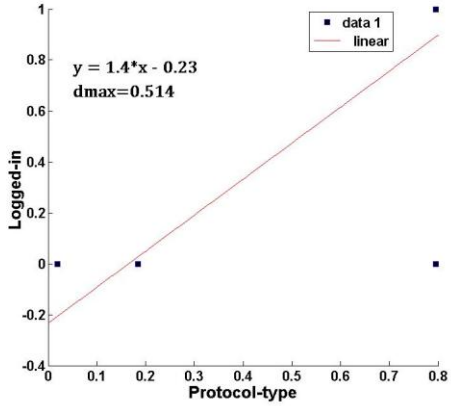


Figure 3. Linear Association between protocol_type and logged_in.

The figure shows few points because the value of the feature *logged_in* is mostly 1 or 0, so that several points are overwritten. Maximum distance is calculated from the longest distance to the line.

Moreover, although some correlated features have a high value of coefficient of determination, they could have no adequate linear association. Therefore, we prune the association with a small value of error and ignore the point on the border to achieve better linearity. So, we have derived various linear equations from the correlated features. Therefore, we present a dependability model that shows the correlated features and hence their dependencies (correlation coefficient values). Intuitively, the concerned pairs from online traffic will be matched to the related linear line. Figure 4 depicts a general dependability model of the MVRF that represents the normal network behavior based on the offline dataset NSL-KDD.

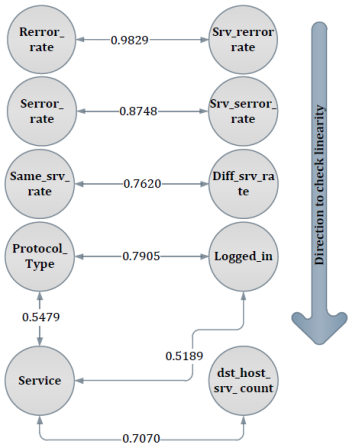


Figure 4. Dependability model of MVRF.

This model implies that, these features are strongly correlated when the network traffic is normal. That means, based on the benchmark dataset NSL-KDD and the selected feature set MVRF, this model can be used to analyze the online traffic directly and detect the normal connections or the abnormal ones. Generally, the online traffic is prepared so that firstly the distance from the pair (*rerror_rate*, *srv_rerror_rate*) and the linear line $y_i = 0.0013 + 0.98x_i + \varepsilon_i$ is calculated and compared to the value of d_{max}. Consequently, the distance from the pair (*serror_rate*, *srv_serror_rate*) and the linear line $y_i = 0.003 + 0.8x_i + \varepsilon_i$ is calculated and compared to the value of d_{max}. In the last step, distances must be evaluated so that all must fulfill the condition $d \leq d_{max}$. Finally, if the online traffic is detected as normal it will be considered in the dataset to adjust the linearity accordingly, such that the dependability model stays adaptive.

To test this model and the proposed idea of detecting the normal traffic in real time, we selected arbitrary instances from the NSL-KDD dataset so that two instances are normal and the third is anomaly. We calculated the distances as defined before. Moreover, the error value is fixed to 0.0005 for all linear lines just to ease the calculation of all distances. The instances are shown in Table V, instances values are sorted the same way as in Table III. Moreover, Table VI depicts the detection result for the test instances. It proved that the proposed idea could detect all instances significantly based on the distance from the linear line.

TABLE V. TEST INSTANCE DATA FROM NSL-KDD

Instance number	Instance values (normalized)
Instance 1	0.7957,0.5646,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0.08,0
Instance 2	0.1848,0.1343,0,0,0,0,0,0,0,0,0,0,0,1,0,0.9882,0,0
Instance 3	0.8152,0.0545,0.0037216,0,0.051948,0,1,0,0,0,0.9,0.9,0,0,1,0,0.15294,0.13,0

The first two instances are normal and the third one is anomaly. Table VI shows the results, so that linearity means the linear line between the correlated features; they are sorted according to the strong linearity in descending order. Moreover, d_{max} is the maximum distance for each linear line between features, **d** means in the table the distance of this instance from the linear line, as mentioned before, the concerned pair from each instance is selected for the suitable linear line. The following table summarizes the testing results.

TABLE VI. TEST RESULT OF SELECTED INSTANCES

Linearity	d_{max}	Instance 1 (d)	Instance 2 (d)	Instance3 (d)
13↔14	0.02	0.00092	..	0.0009
11↔12	0.02	0.001	..	**0.139**
1↔7	0.514	0.199	..	0.22
15↔16	0.2	0	..	0
2↔17	0.65	0.6	..	0.08
1↔2	0.4	0.336	..	**0.4**
2↔7	0.64	0.064	..	0.42

Obviously, the distances of instance1 are all minimum than the maximum distance of each linear association, so it is a normal traffic. In contrast, instance3 has two distances (in bold) greater than the maximum distance in the linear line for the intended correlated features, which is lead to predict this traffic as anomaly. Therefore, the new detected normal instances will be added to the normal dataset so that a new linear line and dependability model with a roughly modified correlation values will be enhanced.

Principally, we focus in this paper on the idea of dependability model and how it represents the normal network behavior. Hence, to declare this idea we have introduced a test and evaluation example from an offline dataset. But we have explained how to use this idea to predict online normal traffic using the distance threshold.

Another main discussion point is the linearity between features. We notice that when the dataset increases the features are not more correlated and they lose the linearity. The association between these features becomes nonlinear. Therefore, in the incoming research work we will exploit the idea in [16], so that a nonlinear association between features

 ISBN: 978-1-61208-227-1

will be established by exploiting the idea of Maximum Information Coefficient (MIC).

VI. CONCLUSION AND FUTURE WORK

Defining a normal network behavior is a necessary step in intrusion detection system. However, it is a challenge under research in the data mining area. In this research work, we present a novel dependability model from the positive correlations between network features. In addition, we abstract the linear associations between these correlated features and exploit them to predict the normal connection from the online traffic in the real time. The prediction is examined so that each features pair from the online traffic instance is exploited to calculate their distance from the linear line related to these pair exclusively. Furthermore, all distances of all feature pairs in the online traffic must be greater than the threshold distance (d_{max}) to consider it a linear connection. Our test results show that the model could detect the normal connection and anomaly as well from test dataset NSL-KDD. In addition, we looked to enhance the model by examine the nonlinear association in large dataset. Finally, we proved that the dependability model can represents the normal network behavior and can support the IDS to detect the attacks in online traffic. Therefore, it promises more accuracy, less overhead in classification, and enhancement in network performance.

ACKNOWLEDGMENT

This research project “SecMonet” is supported by the German Federal Ministry of Education and Research (BMBF) under the funding line “FHprofUnt”.

REFERENCES

[1] R. Karthick, V. P. Hattiwale, and B. Ravindran, “Adaptive Network Intrusion Detection System using a Hybrid Approach,” IEEE Fourth International Conference in COMSNETS, pp. 1-7, Januray, 2012.

[2] M. Thottan and C. Ji, “Anomaly Detection in IP Networks,” IEEE Transactions On Signal Processing, vol. 51, NO. 8, pp. 2191-2204, August, 2003, doi: 10.1109/TSP.2003.814797.

[3] D.M. Best, S. Bohn, D. Love, A. Wynne, and W. Pike, “Real-Time Visualization of Network Behaviors for Situational Awareness,” ACM Proceedings of the Seventh International Symposium on Visualization for Cyber Security, pp. 79-90, 2010, doi: 10.1145/1850795.1850805.

[4] Q. Gan and B. E. Helvik, “Dependability Modelling and Analysis of Networks as Taking Routing and Traffic into Account,” IEEE Conference in Next Generation Internet Design and Engineering, 2006, doi: 10.1109/NGI.2006.1678219.

[5] A.H. Aznin, R. Ahmad, Z. Muhamad, A. Basari, and B. Hussin, “Correlated Node Behavior Model based on Semi Markov Process for MANETS,” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January, 2012, ISSN:1694-0814.

[6] S. Kakuru, “Behavior Based Network Traffic Analysis Tool,” IEEE third conference in communication software and networks (ICCSN), pp. 649–652, 2011, doi: 10.1109/ICCSN.2011.6014810.

[7] W. Liu, D. Huang, and L. zhang, “Analysis of Network User Behavior” IEEE youth conference on information computing and telecommunications, pp. 126-129, November, 2010, doi: 10.1109/YCICT.2010.5713061.

[8] M. Burgess, H. Haugerud, S. Straumsnes, and T. Reitan, „Measuring System Normality“ ACM Transactions on Computer Systems, Vol. 20, No. 2, pp. 125-160, May, 2002, doi: 10.1145/507052.507054.

[9] J. M. Estevez-Tapiador, P. Gracia-Teodoro, and J. E. Diaz-Verdejo, “Measuring normality in HTTP traffic for anomaly-based intrusion detection” Elsevier Computer Networks, pp. 175-193, 2004.

[10] M. V. Mahoney and P. K. Chan, “Learning nonstationary models of normal network traffic for detecting novel attacks” ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 376-385, 2002, doi: 10.1145/775047.775102.

[11] M. Salem, U. Buehler, and S. Reissmann, “Improved Feature Selection Method using SBS-IG-Plus”, ISSE Proceeding on Securing Electronic Business Processes, pp. 352-361, 2011.

[12] A.G. Asuero, A. Sayago, and A.G. Gonzalez, “The Correlation Coefficient: An Overview” Critical Review in Analytical Chemistry, pp. 41-59, 2006, doi: 10.1080/10408340500526766.

[13] J. Freeman and T. Young, “Correlation Coefficient: Association Between Two Continuous Variables” Scope, pp. 31-33, June, 2009.

[14] N. Chen, X. Chen, B. Xiong, and H. Lu, “An Anomaly Detection and Anaylsis Method for Network Traffic Based on Correlation Coefficient Matrix,” IEEE conference on Embedded Computing, pp. 238-244, 2009, doi: 10.1109/EmbeddedCom-ScalCom.2009.50.

[15] M. Salem, U. Buehler, “ Hybrid Normalization Method in Data Mining Toward Improving the Network Intrusion Detection System” submitted to the IEEE conference on Data Mining, ICDM 2012.

[16] D. Reshef, et al. “ Detecting Novel Association in Large Data Sets”, 2011, doi: 10.1126/science.1205438.

[17] http://math.ucsd.edu/~wgarner/math4c/derivations/distance/distptline.htm.

[18] Nsl-kdd dataset: http://nsl.cs.unb.ca/NSL-KDD/, March, 2009.

[19] Stroock, D. “An Introduction to Markov Processes”. Graduate Text Series #230, Springer–Verlag, Heidelberg, 2005.

[20] Edwards, A. L. "The Correlation Coefficient." Ch. 4 in An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, pp. 33-46, 1976.

 ISBN: 978-1-61208-227-1

Semantic Description of Text Mining Services

Katja Pfeifer
SAP Research Dresden
SAP AG
01187 Dresden, Germany
katja.pfeifer01@sap.com

Alexander Schill
Computer Networks Group
Technische Universität Dresden
01062 Dresden, Germany
alexander.schill@tu-dresden.de

***Abstract*—Today, a huge amount of crucial business knowledge is hidden in unstructured text sources, such as word documents, web pages or forum entries. In order to exploit this knowledge text mining techniques were developed that are able to automatically extract or annotate entities, their relations or sentiments from textual sources. Recently, a number of text mining services that offer REST or SOAP APIs for easy consumption were published. These services differ strongly in their mining abilities and result quality and are often constructed for specific use cases. In practice, it is often desirable to combine results of multiple services to increase quality and functionality. However, this result combination is difficult since descriptions of service functionalities are often rarely documented and not standardized so that searching for specific text mining characteristics is time consuming and complex. In this paper we introduce a categorization of text mining services and provide a novel description ontology for describing functional characteristics of a text mining service. The ontology, being of interest for practitioners as well as researchers, is completed by application examples and descriptions that are made publicly available. Through the ontology and the descriptions presented in this paper the automatic use and combination of different text mining services is enabled.**

Keywords-Text Mining; Semantic Description; Service Oriented Architecture.

I. Introduction

Today, more than 80 percent of business-relevant information only exists in unstructured, mostly textual form such as web pages, office documents or forum entries as estimated in [7]. Exploiting this knowledge in business intelligence applications will be crucial for business competitiveness in future. In order to satisfy the need for knowledge extraction from text, a large quantity of text mining approaches have been developed (see [10] for an overview). These support a wide range of knowledge harvesting tasks like the classification of text documents, the recognition of entities and relationships or the identification of sentiments. Recently, more and more of these text mining solutions were made publicly available as Web - or Rest Services (e.g., OpenCalais [23] and AlchemyAPI [17]) to simplify consumption and integration.

Even though there are many text mining solutions available, some major problems remain unsolved. Text mining often still faces the problem of inaccuracy and incompleteness, and therefore, limits the confidence in information extracted by text mining solutions. Moreover, most of the solutions are developed for specific use cases and are not usable for others.

In order to alleviate these problems, we strive for a combination of multiple text mining services as described in [20]. The idea is to raise the quality of text mining by combining the strength and weaknesses of different approaches. Unfortunately, searching for specific text mining functionalities and combining these is cumbersome and often leads to a great amount of manual work.

In this paper, we address the need for a comprehensive semantic description of text mining services to simplify the search for specific mining functionalities and therefore allow to automate the combination of text mining services. In particular we make the following contributions that are of interest for practitioners as well as researchers:

- We classify existing text mining services and highlight their similarities and differences.
- We propose a novel description ontology that can be used to comprehensively describe text mining services.
- We present descriptions for common text mining services and made them publicly available.
- We show that the descriptions can be used to select text mining services based on their functionalities.

The remaining paper is structured as follows: In Section II, we further motivate the need for combining text mining results and introduce a system architecture that supports automatic combination of existing services. Related work is reviewed in Section III. To infer the information necessary for the descriptions, we classify existing text mining services in Section IV. The novel description ontology is introduced in detail in Section V and complemented by application examples in Section VI. Finally, Section VII concludes our paper.

II. Motivating Text Mining Service Combination

In order to further motivate the need for a combination of text mining services, an example is given in Figure 1. The figure shows an extract of a BBC news article together with text mining results that were extracted by four different services - in particular, OpenCalais, AlchemyAPI, FISE [12]

 ISBN: 978-1-61208-227-1

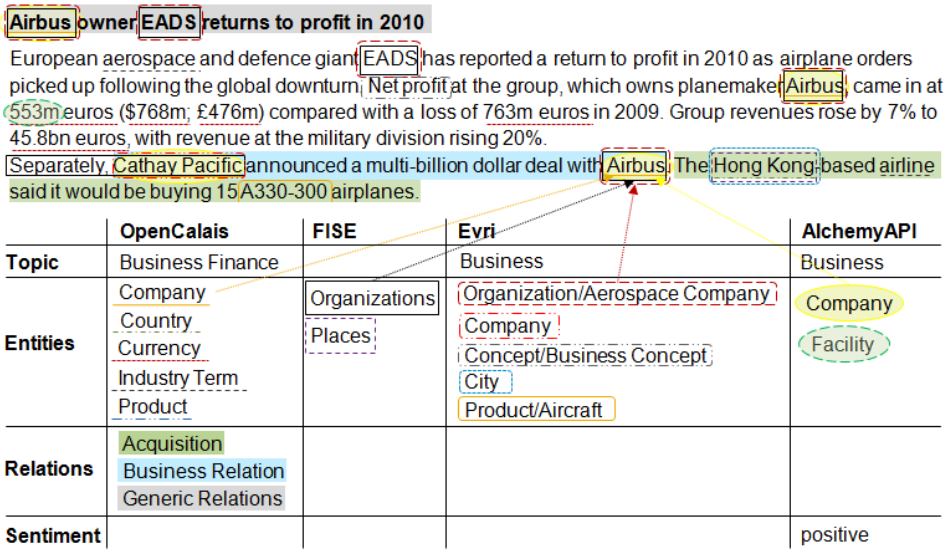


Figure 1. Analysis of a business news by several text mining services

and Evri [5] (the news article and the text mining results were retrieved on March 9, 2011). All of them are able to extract some entities such as companies, cities or products, but differ in the completeness and correctness of the results and the used annotation taxonomies. In addition, some services extracted more enhanced information such as relationships or overall topics and sentiments.

It is desirable to combine the results of these different services as proposed in [20]. Figure 2 illustrates a possible architecture of a system that is able to combine several text mining services. The lower part depicts a number of exemplary services (S1-S3) offering text mining functionalities with inconsistent interfaces and different entity taxonomies (T1-T3). We introduce a layer of wrappers that harmonize the individual service interfaces on the syntactical level and are considered in-depth in future work. These wrappers should be manually or possibly semi-automatically provided by the community or the service provider. They are simple services rewriting and adapting the original service answers to a unified format in order to facilitate the reuse and combination of the service results.

Additionally, we propose that each service functionality is semantically described using a standardized text mining ontology. This is needed since most services are often only rarely documented or documented in non machine-readable form on a website. Furthermore, available descriptions only specify the services on a syntactical level regarding their interfaces, their data types and bindings and their access modalities. The semantic description can then be put into a registry that helps to automatically find the adequate services for an envisioned text mining task. A text mining combination system is able to call multiple text mining services and combine their results.

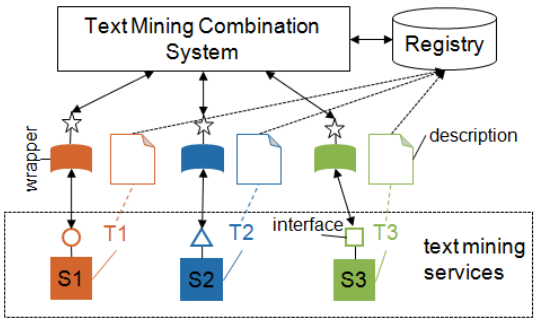


Figure 2. Architecture of service-oriented Text Mining

By extending the work of [20], the current paper focuses on the semantic description of text mining services. Before we start to introduce our work, we will give an overview on related work and distinguish it from our contributions.

III. Related Work

The first service-oriented text mining approaches, especially in the domain of information extraction, have been discussed by Habegger et al. [9] and Grover et al. [8]. Both approaches break down information extraction processes into single partial operations and offer a language or accordingly an ontology for the basic description of these service artifacts. In contrast to our ontology, none of them offers a semantic description capable to specify text mining functionalities for complex services. A first service-oriented information extraction system that uses text mining services is presented by Starlinger et al. [21]. It connects biomedical text mining services having standardized interfaces and a common taxonomy and combines their results to improve extraction quality. An automatic identification of services

 ISBN: 978-1-61208-227-1

Service	Domain	Taxonomy	Text Mining Functionalities					
			NER	RE	TC	CT	KE	SA
AIIAGMT[11]	biomedical		(en:) genes	-	-	-	-	-
AlchemyAPI	generic	list	en,fr,es,de,it,pt,ru,sv; LD, ED, QE	-	en,fr,es,de,it, pt,ru,sv	en:LD	en,fr,es,de,it, pt,ru,sv	en,fr,es,de,it,pt,ru,sv: polarity; d-, e-, k-level
BeliefNetworks [1]	generic		-	-	-	en	-	-
Evri	generic	service call (types & facets)	en; links to Evri knowledge base	-	en	-	-	-
Extractiv [6]	generic	list	en; LD, ED	en; (QE)	en	-	-	-
FISE	generic	list (DBpedia types)	en; basic LD	*planned*	-	-	-	-
OpenAmplify [15]	generic		(en: proper nouns)	(en: actions)	en	-	en	en: polarity; d-level
OpenCalais	business, finance, generic	owl (types & attr.) for en/ list (types & attr.) for es,fr	en; LD, ED es,fr	en	en: list	en: social tags	-	-
PIE [22]	biomedical		(en:) protein	en: protein interactions	-	-	-	-
uClassify [24]	generic	list (topic hierarchy)	-	-	en	-	-	en: polarity, mood; d-level

Table I
OVERVIEW OF EXISTING TEXT MINING SERVICES

based on their functionalities and corresponding descriptions is completely missing in this approach.

The CLARIN project [4] has the vision to create a research infrastructure of language resources and therefore also touches the problem of descriptive meta data for language services. In [14], a minimal set of meta data for language tools is detected. In contrast to our work, the CLARIN project mainly focuses on basic language tools (e.g., tokenizer, POS-tagger) and the (semi-)automatic build of chains between those tools. CLARIN does not review complex end user services as we do and additionally does not provide an ontology for describing the functionality of end user services like OpenCalais or AlchemyAPI.

Different web service description languages exist to describe services regarding their functionalities, the used data types, the protocols and the provided interfaces. The W3C standard Web Services Description Language (WSDL) [3] was established for the syntactic description of services. More recent approaches [13], such as the Ontology Web Language for Services (OWL-S), Web Service Modeling Ontology (WSMO) and Semantic Annotations for WSDL and XML Schema (SAWSDL) added additional semantic descriptions in order to allow the automatic selection of services based on their functionalities. Nevertheless, all semantic descriptions need a well-defined ontology for describing the service features (even the functionality description of the OWL-S profile needs complementary ontology elements to specify the input, output, precondition and effect properties). As stated above, there has been no such ontology for describing text mining services comprehensively. To the best of our knowledge, we are the first to approach this problem providing an ontology for describing text mining services.

IV. CLASSIFICATION OF TEXT MINING SERVICES

We intensively studied existing text mining services with regard to the functionalities they offer and their special characteristics and limits. We provided a first overview of existing services in [20], which we extended in Table I, focusing on the text mining specific characteristics. We selected text mining services from different domains with different functionalities (named entity recognition, relationship extraction, categorization, concept assignment, keyword extraction, sentiment analysis) and tried to cover the most established one. First of all we studied the domain the services have been designed for. We mainly discovered generic services (i.e., not being specialized for any specific domain) and services from the biomedical and business domain. We further analyzed the different text mining functionalities and identified six main types:

- *Named Entity Recognition (NER)* where entities are identified and classified into predefined categories (e.g., person, organization),
- *Relation and Interaction Extraction (RE)* for the identification of relationships between two or more entities,
- *Text Classification/Categorization (TC)* where categories are assigned to text documents,
- *Concept Tagging (CT)* for the assignment of specific terms that are derived from the text content (the terms do not have to be included in the text),
- *Keyword Extraction (KE)* where the essence of the text is extracted through the identification of the main keywords of a text and
- *Sentiment Analysis (SA)* for the extraction of any subjective information from text (e.g., polarity, attitudes, mood).

We studied these six text mining types more extensively and identified their essential properties. All types are language-dependent and most functionalities are currently only provided for English text input. The information extraction tasks NER, RE and in parts also TC are identifying elements of predefined categories. Therefore, the service providers generally release a taxonomy defining the entities, relations and classification categories (by an ontology file, a list on the service website or indirectly via some service calls). The taxonomies differ in their semantic and syntactic

 ISBN: 978-1-61208-227-1

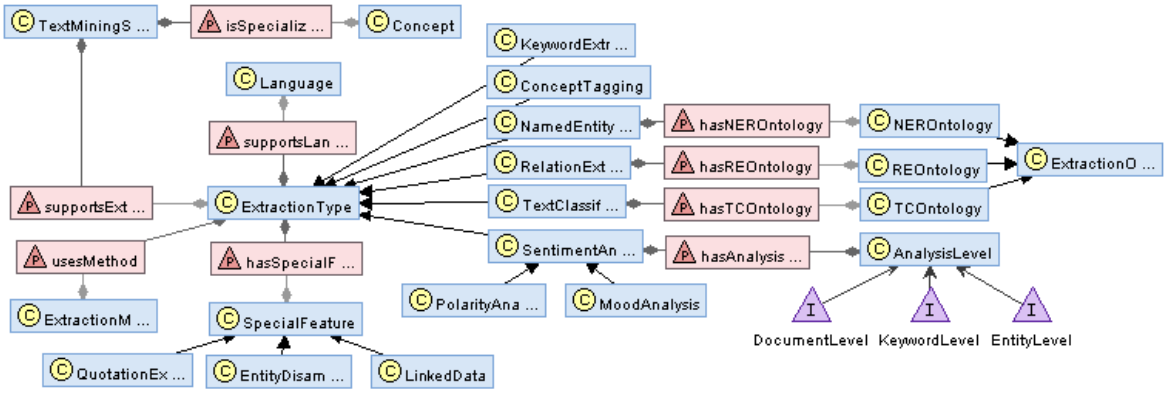


Figure 3. Ontology for the description of text mining services

granularity - some are enhanced with attributes and facets, others are only providing flat basic types. Under SA, we recapped all text mining functionalities touching subjective information. Although some of the features could also be classified under other text mining types like NER (as the subjective information is extracted on entity level), we hold that it is a text mining type on its own. For better specification, we differentiated between three analysis levels - document (d), entity (e) and keyword (k) - and indicated the exact sentiment type a service provides.

Further extraction features are offered by several services for some text mining types. The most common feature is the Linked Data (LD) support, where the extracted objects are linked to existing LD sources with additional information (e.g., a link from the person entity *Barack Obama* to a LD URL characterizing him). Another feature is the disambiguation of entities (ED) where detected instances are resolved to a unique instance (e.g., *IBM* and *IBM Corp.* are resolved to one entity). One service additionally provides a quotation extraction (QE), where person entities are complemented by quotations from this person found in the input text.

Based on the analysis of existing text mining services, we conclude that a description language for text mining services should satisfy the following requirements: describe the domain the service is designed for (R1), indicate the text mining type(s) provided by the service (R2) plus the respective languages (R3), point to the ontologies used for the predefined category types (R4) and describe the special features being available for the text mining types (R5). Additional information on the used extraction methodology, service charges and limitations can complete the description. In the following we will present an expandable ontology for the description of text mining services satisfying the mentioned requirements.

V. The Text Mining Description Ontology

Figure 3 presents our high level ontology for the description of text mining services. We especially focused on the expandability and simplicity of the ontology and modeled it with the Resource Description Framework (RDF) Schema [2]. We chose RDF since it is easy to use and provides sufficient mechanisms to define classes, properties and their relationships. RDF allows for easy extensibility and re-use of existing well-defined and more specific ontology parts and also supports user service specific extensions. In addition, it is not a problem to interlink the syntactic service description to a description provided in RDF. Before we explain this interlinking of classical descriptions with our extension, we will shortly introduce the classes and properties of our text mining description ontology.

The class *TextMiningService* is the entry point for a semantic description of a text mining service and can be used for the interlinking of the classical service description and the text mining specific description. An instance of this class represents a service with a well defined interface that offers some text mining functionalities. Via the property *supportsExtractionType*, specific text mining functionalities indicated by instances of the class *ExtractionType* are linked to the service (R2). Several subclasses of *ExtractionType* are available for the exact specification of the text mining tasks being provided. If a text mining service is specialized for a certain domain (R1), this will be indicated with the property *isSpecializedFor* that connects a *TextMiningService* with a *Concept* from the Simple Knowledge Organization System (SKOS) [25] ontology. The supported languages of a service or a concrete text mining type (R3) are given with the property *supportsLanguage* pointing to a *Language* from DBPedia. Characteristics of the *ExtractionType* can be specified with additional properties (e.g., *hasSpecialFeature*, *hasNEROntology*, *hasREOntology*, ...). In order to indicate special features provided by the services (R5), the class *SpecialFeature* and some subclasses for concrete features are provided by our ontology. The extraction ontology used by a service (R4) can be indicated with instances from the class *ExtractionOntology*. These instances are mainly used to link to the existing taxonomies of the services. The methods used

 ISBN: 978-1-61208-227-1

for the extraction can optionally be specified with instances from the class *ExtractionMethod*. We decided to model the subtypes of *SpecialFeature* and *ExtractionOntology*, as well as *ExtractionMethod* as classes, as we believe that they should be further specified with extra properties in future.

VI. Application of the Ontology

After having explained our proposed ontology in the previous section, we will now show how our ontology can be applied in practice. Semantic service descriptions using our presented ontology can for example be linked to the WSDL [3] description of a service through Semantic Annotations for WSDL and XML Schema (SAWSDL) [13]. In the following, we show an exemplary semantic description of the text mining service OpenCalais and the integration into the syntactic WSDL description.

Figure 4 shows an extract from the annotated WSDL file of OpenCalais (The original WSDL file can be found at [16]). The pointer to the semantic description can be integrated as SAWSDL *modelReference* into any service model element in the WSDL description. However, as our semantic description characterizes the service, we prefer a linkage from the WSDL service element. Other linking concepts between the syntactic and the semantic description are of course possible as our semantic descriptions build upon open standards. We extended the OpenCalais WSDL *service* element in Figure 4 with a SAWSDL *modelReference* pointing to a semantic description of the OpenCalais functionalities.

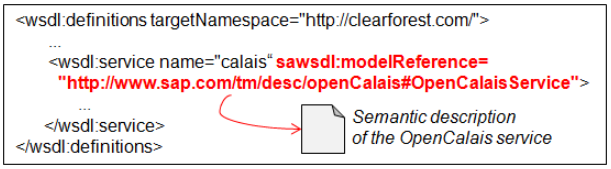


Figure 4. Extract from WSDL of OpenCalais service annotated with SAWSDL

Listing 1 displays an extract of the semantic information that can be found under the linked URI and all its connected resources. This exemplary semantic description of the OpenCalais service makes use of our previously introduced ontology (notice that this is not a complete description of all the functionalities of the OpenCalais service.).

```
@prefix oc: <http://www.sap.com/tm/descr/openCalais#> .
@prefix tm: <http://www.sap.com/tm/descr/ontology#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .

oc:OpenCalaisService a tm:TextMiningService;
  tm:isSpecializedFor dbpedia:Category:Business;
  tm:supportsExtractionType oc:NEREnglish,
                            ...
                            oc:DocumentCategorization .
oc:NEREnglish a tm:NamedEntityRecognition;
  tm:supportsLanguage dbpedia:English_language;
  tm:hasSpecialFeature   oc:EntityDisambiguation,
                         oc:LinkedData;
  tm:hasNEROntology oc:OntologyEnglish .
oc:EntityDisambiguation a tm:EntityDisambiguation .
oc:LinkedData a tm:LinkedData.
oc:OntologyEnglish a tm:NEROntology;
  nie:url http://www.opencalais.com/files/owl.opencalais
      -4.3a.xml .
...
oc:DocumentCategorization a tm:TextClassification;
  tm:supportsLanguage dbpedia:English_language .
...
```

Listing 1. Extract of a semantic description for the OpenCalais service in N3 notation

We started describing a number of text mining services with our ontology and will continuously add and extend descriptions. The ontology as well as the descriptions files can be found under [18]. The fake URIs have to be replaced for usage. The text mining service combination system makes use of the descriptions. Therefore, the describing triples are stored in a triple store like Sesame. Adequate services are then searched as follows.

A. Searching specific Services

As our descriptions of text mining services use RDF triples, it is obvious to query the descriptions with the help of SPARQL [19] a query language based on graph patterns. We will now demonstrate how to identify and select text mining services with specific functionalities that are described with our ontology. The given queries are just examples. Actual queries may be much more complex. The first query (Listing 2) selects text mining services for NER on Spanish text documents where the extracted entities are connected to Linked Data resources if possible. Figure 5 shows the corresponding query pattern for this SPARQL query.

```
PREFIX tm:<http://www.sap.com/tm/descr/ontology#>
PREFIX dbpedia:<http://dbpedia.org/page/>

SELECT DISTINCT ?service
WHERE {
  ?service tm:supportsExtractionType ?type .
  ?type a tm:NamedEntityRecognition ;
    tm:supportsLanguage dbpedia:Spanish_language ;
    tm:hasSpecialFeature ?feature .
  ?feature a tm:LinkedData .
}
```

Listing 2. SPARQL query to select services providing NER with Linked Data for Spanish text

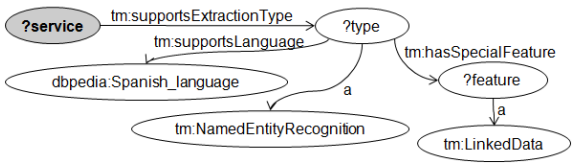


Figure 5. Query pattern for SPARQL query in Listing 2

Listing 3 shows another SPARQL query that selects services capable to analyze the mood of an English text document.

```
1  PREFIX tm:<http://www.sap.com/tm/descr/ontology#>
2  PREFIX dbpedia:<http://dbpedia.org/page/>
3
4  SELECT DISTINCT ?service
5  WHERE {
6    ?service tm:supportsExtractionType ?type .
7    ?type a tm:MoodAnalysis ;
8      tm:supportsLanguage dbpedia:English_language ;
9      tm:hasAnalysisLevel tm:DocumentLevel .
10 }
```

Listing 3. SPARQL query to select services providing sentiment analysis (mood) for English text on document level

VII. Conclusions and Future Work

In this paper, we made a number of contributions that help to simplify the description, search and reuse of text mining services and their result combination. First of all, we gave an overview on existing text mining services and classified them according to their functionalities. We further derived a novel description ontology for text mining services capable to describe complex end-user services. In contrast to previous work, we explicitly covered the real mining functionalities into the descriptions. Auxiliary, we built on open standards to easily connect descriptions using our ontology to already existing descriptions and standardizations.

As starting point for further work on the selection, reuse and combination of text mining services, we described a number of such services and made them publicly available. With this basis, the combination of text mining services as proposed in [20] is enabled. Future work will have to focus on the well-directed extension of the ontology as well as the derivation of rules and heuristics for the combination of the service results and the evaluation of the system. Another research area we investigate is the matching of service taxonomies to retrieve mappings between them and possibly even a global taxonomy. These mappings can then complement the service descriptions presented in this paper.

References

[1] BeliefNetworks. http://beliefnetworks.net/bnws/, retrieved: April, 2012.

[2] D. Brickley and R. V. Guha. Rdf vocabulary description language 1.0: Rdf schema. *W3C Recommendation*, 10, 2004.

[3] R. Chinnici, J.-J. Moreau, A. Ryman, and S. Weerawarana. Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language. Technical report, World Wide Web Consortium (W3C), 2007.

[4] CLARIN. http://www.clarin.eu/, retrieved: August, 2012.

[5] Evri. http://www.evri.com/developer/, retrieved: July, 2012.

[6] Extractiv. http://extractiv.com/, retrieved: August, 2012.

[7] S. Grimes. Unstructured Data and the 80 Percent Rule. Clarabridge Bridgepoints, 3rd quarter 2008.

[8] C. Grover, H. Halpin, E. Klein, J. L. Leidner, S. Potter, S. Riedel, S. Scrutchin, and R. Tobin. A Framework for Text Mining Services. In *AHM'04 Proc.* EPSRC, 2004.

[9] B. Habegger and M. Quafafou. Web Services for Information Extraction from the Web. In *ICWS'04 Proc.*, page 279. IEEE Computer Society, 2004.

[10] A. Hotho, A. Nürnberger, and G. Paaß. *A Brief Survey of Text Mining*, volume 20, pages 19–62. Gesellschaft für linguistische Datenverarbeitung, 2005.

[11] C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang, and I.-F. Chung. Integrating High Dimensional Bi-directional Parsing Models for Gene Mention Tagging. *Bioinformatics*, 24:286–294, 2008.

[12] Interactive Knowledge Stack. Furtwangen IKS Semantic Engine. http://wiki.iks-project.eu/index.php/FISE, retrieved: August, 2012.

[13] M. Klusch. *CASCOM - Intelligent Service Coordination in the Semantic Web*, chapter Semantic Web Service Description, pages 41–68. Birkhuser Basel, 2008.

[14] L. Lemnitze, E. Hinrichs, and A. Witt. Language Resources, Taxonomies and Metadata. In G. Heyer, editor, *Text Mining and Services Proc.*, volume XIV of *Leipziger Beiträge zur Informatik*, pages 25–39, Leipzig, 2009.

[15] OpenAmplify. http://www.openamplify.com/, retrieved: August, 2012.

[16] OpenCalais WSDL. http://api.opencalais.com/enlighten/?wsdl, retrieved: August, 2012.

[17] Orchestr8. AlchemyAPI. http://www.alchemyapi.com/, retrieved: August, 2012.

[18] K. Pfeifer. Text Mining Ontology and Descriptions. http://areca.co/20/Text-Mining-Service-Descriptions, retrieved: August, 2012.

[19] E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. Technical report, W3C, 2006.

[20] K. Seidler and A. Schill. Service-Oriented Information Extraction. In *Joint EDBT/ICDT Ph.D. Workshop'11 Proc.*, pages 25–31, New York, NY, USA, 2011. ACM.

[21] J. Starlinger, F. Leitner, A. Valencia, and U. Leser. SOA-Based Integration of Text Mining Services. In *SERVICES '09 Proc.*, pages 99–106, Washington, DC, USA, 2009. IEEE Computer Society.

[22] K. Sun, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, and B.-T. Zhang. PIE: An Online Prediction System for Protein-Protein Interactions From Text. *Nucleic Acids Research*, 36:411–415, 2008.

[23] Thomson Reuters. The OpenCalais Web Service. http://www.opencalais.com/, retrieved: August, 2012.

[24] uClassify. http://uclassify.com/, retrieved: August, 2012.

[25] World Wide Web Consortium. *SKOS Simple Knowledge Organization System Reference*, August 2009.

 ISBN: 978-1-61208-227-1

Semantic Tools for Forensics: A Highly Adaptable Framework

Michael Spranger, Stefan Schildbach, Florian Heinke, Steffen Grunert and Dirk Labudde
University of Applied Sciences Mittweida
Bioinformatics Group
Department of MNI
Germany, 09691 Mittweida
Email: {*name.surname*}@hs-mittweida.de

***Abstract*—Textual information or data annotated with textual information (meta-information) are regular targets of securing or confiscating relevant material in the field of criminal proceedings. In general evaluation of relevant material is complex, especially the manual (re)search in the increasing amount of data as a result of cheaper storage capacity available nowadays therefore the identification of valid relations are enormously complex, error-prone and slow. In addition, the adherence to time limits and data privacy protection make searching even more difficult. The development of an (semi-)automatic high modular solution for exploration of this kind of data using capabilities of computer linguistic methods and technologies is presented in this work. From a scientific perspective, the biggest challenge is the automatic handling of fragmented or defective texts and hidden semantics. A domain-specific language has been defined using the model-driven approach of the Eclipse Modeling Framework for the purpose of developing forensic taxonomies and ontologies. Based on this, role-based editors have been developed to allow the definition of case-based ontologies and taxonomies and the results of manual annotation of texts. The next steps required for further development are going to include comparison of several back-end frameworks, e.g., for indexing, information extraction, querying and the providing of a graphical representation of relations as a knowledge map. Finally, the overall process needs to be optimized and automated.**

Keywords*—*forensic; ontology; taxonomy; querying; framework.

I. INTRODUCTION

The analysis of texts retrieved from a variety of sources, e.g., secured or confiscated storage devices, computers and social networks, as well as the extraction of information, are two of the main tasks in criminal proceedings for agents or other parties involved in forensic investigations. However, the heterogeneity of data and the fast changeover of communication forms and technologies make it difficult to develop one single tool covering all possibilities. In order to address this problem, a domain framework is presented in this paper applying computer linguistic methods and technologies on forensic texts.

In this context, the term *forensics* relates to all textual information which maybe used during the procedure of taking evidence in a particular criminal proceeding. In particular, it corresponds to the hidden information and relations between entities achieved through the exploration and application of computer linguistic processing of potential texts.

Generally, there are a variety of tasks which need to be addressed:

- Recognition of texts with a case-based criminalistic relevance
- Recognition of relations in these texts
- Uncovering of relationship networks
- Uncovering of planned activities
- Identification or tracking of destructive texts
- Identification or tracking of hidden semantics

In this context, the term *hidden semantics* is synonymous with one kind of linguistic steganography, whereas such texts are defined as "...made to appear innocent in an open code." [1]. Each of these tasks can be processed and solved by combining several highly specialized services that encapsulate a problem solver based on a specific text mining technology. This problem solver can be combined and recombined like a tool kit to achieve a polymorphous behaviour depending on the kind of texts and the particular question under investigation.

Basic structural concepts of an application framework suitable to deal with these problems are presented within this paper. The previous steps of development will be outlined in the following sections.

- Development of criminalistic ontologies
- Development of criminalistic corpora
- Development of the framework's architecture
- Implementation of a prototype for manual evaluation

Specific ontologies and taxonomies are not being introduced in this paper. Case-based specific ontology and taxonomy are currently under evaluation applying the generic ontology editor developed in this work and will be released soon together with basic structures.

II. DEVELOPMENT OF CRIMINALISTIC ONTOLOGIES

The term *ontology* is commonly understood as a formal and explicit specification of a common conceptualization. In particular, it defines common classified terms and symbols referred to a syntax and a network of associate relations [2] [3]. Developing ontologies for criminalistic purposes is a prior condition for annotating texts and raise questions in this particular domain. The term *taxonomy* as a subset of ontology is used for the classification of terms (concepts) in ontologies

 ISBN: 978-1-61208-227-1

and documents. On the one hand, a criminalistic ontology is charactarised by its case-based polymorphic structure and on the other by special terms used in criminal proceedings.

A domain-specific language is necessary at the beginning to describe taxonomies and ontologies for the development of a criminalistic ontology. The domain ontologies considered need to be highly specialized by taking into account the individual nuances of the particular criminal proceeding and the legal requirements due to privacy protection. For these reasons, a vast ontology covering all areas of crime is not employable. Special case-based ontologies, in accordance to a suitable predefined ontology, are necessary and preferably developed by the person heading the investigation. Thus, it is important that the definition of the predefined ontology is easily and case specific adaptable.

The Eclipse Modeling Framework (EMF) [4] [5] has been chosen for the purposes of this work mainly because of its perfect integration into the Eclipse environment, but also for participating in the manifold advantages of the approach of a model-driven software development. To follow this paradigm, the next step required is the definition of an abstract syntax (meta-model) for describing such taxonomies and ontologies. The meta-model created that way is used for generating a concrete syntax, especially source code, that provides all model and utility classes required.

In the literature there are different approaches for representing semantics under discussion, with Topic Maps have been proven to be one of the most expressive. Topic Maps is an ISO-standardized technology for representation of knowledge and its connection to other relevant information. It enables multiple, concurrent, structurally unconstrained views on sets of information objects and is especially useful for filtering and structuring of unstructured texts [2] [6]. Therefore the Topic Maps standard has been chosen to be the starting point of the meta-model development. Since EMF already includes options for persistence as well as model searching and (strategic) traversing, only the necessary syntactical elements and paradigms from the ISO standard have been adopted. These syntactical elements provide a complete description of semantic relationships. Note, the specification given in this work takes into account the specifics of the domain with respect to slang, multilingualism and the underlying hidden semantics. The syntactical elements used for further development are defined below and examplified by Figure 1 attached.

Subject (Topic)	*red circle* represents an abstract or concrete entity in the domain to be analyzed.
Instance (Topic)	*yellow circle* is the concrete manifestation of a subject.
Descriptor (Topic)	*orange circle* typifies any other syntactical elements; i.e. adds further details related.
Association	*blue rectangle* is a relation between two topics, usually subject and instance.
Association Role	specifies the roles of the topics in an association (optional).
Occurrence	corresponds to the concrete manifestation of a topic in a resource, usually related to an Instance.
Topic Name	is the name representation of topics (container).
Name Item	denotes the name of a specific topic, associated to a Scope.
Facet	names a class of attributes of a topic and can include several Facet Values.
Facet Value	a particular attribute as distinct value, can be a topic or another Facet.
Scope	defines semantic layers; e.g. causing system to focus by filtering particular syntactical elements.

Figure 1: Use case tax fraud - an application of Topic Maps derivative as developed under this work for modeling a criminalistic ontology.

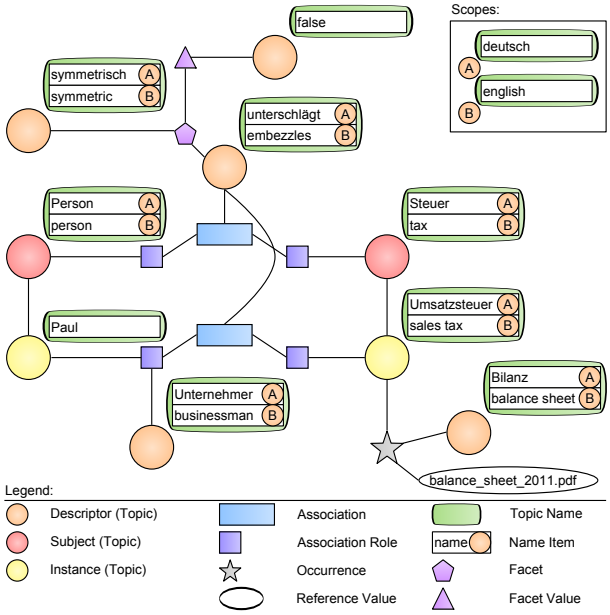


Fig. 1. Sample extract of the application "tax fraud" demonstrating a typical interaction of the different elements. The Subject *person* is described more specifically by adding the Instance *Paul*. The *person* called *Paul* then is related to the Subject *tax*, more specific the Instance *sales tax*. This by the Association, described in more detail by the Descriptor *embezzles*. *Pauls* role in this relationship is specified by the Descriptor of the Association Role *Businessman*. The Instance *sales tax* is creating an Occurence specified by the Descriptor *balance sheet* referring to the Reference Value *balance_sheet_2011.pdf* attached as evidence.

After developing the Domain Specific Language (DSL), the user interface of the ontology and taxonomy editor have been designed in this work.

At this stage of the work, the real development of a criminalistic classification and ontology has been initiated. The basis for comprehending forensic data and its relationship to

 ISBN: 978-1-61208-227-1

case-based information has been achieved in cooperation with the local criminal investigation department. In this way, a set of metadata could be established entitled to be as close to reality as possible.

III. Development of criminalistic corpora

An extensive corpus is needed for the evaluation of the implemented functionalities and development of powerful algorithms in order to detect more semantic details, especially in fragmented and defective texts and for detecting hidden semantics. Building the extensive corpus required using original data from prior preliminary investigations is not suitable because of legal requirements of data privacy protection. This data is exclusively available during the current proceedings.

An alternative method is the exploration of significant characteristics of forensic texts and generating corpora in an artificial way where it is possible to take a completely artificial creation of text into consideration. This can be realized in two ways. The first is character level based, which causes the text to be alienated by non-words and unsuitable, but proper names [7]. The second way, superior from our point of view, is based on morphemes. While the occurrences of non-words can be eliminated, the target language, in the current case German, as a non-agglutinative language, raises problems among this method in shaping and bending words [8]. In summary, the basic problem with both approaches is the possible semantic interruption of text units.

A further method is to generate texts by modifying existing sources. In this case, the internet holds numerous potential domain-specific corpora. Analyzing significant websites, ebooks or expert forums are just a few options for generating suitable texts.

Concluding, the Internet-based concept is more valuable for the project presented here. Therefore, a method for transforming texts is necessary. Common approaches, like lookup based exchanges of words (via free dictionaries), adapting typo errors (missing, wrong or twisted letters) and manipulating the orthography of words, are suitable in this case.

IV. The framework's architecture

Especially due to its platform independency, *Java* has been used for the development. The high modularity is ensured by employing the *Eclipse RCP* as a basis. Its *OSGi* [9] implementation *Equinox* allows to construct service-oriented architectures (SOA) within the *Java Virtual Machine*. The framework conceptually consists of three main modules (see Figure 2):

- **Ontology Machine** it includes all functionalities for developing criminalistic taxonomies and ontologies.
- **Indexing Machine** it includes functions and methods for extraction and annotation of forensic data.
- **Querying Machine** it includes the functions for searching and visualizing semantic coherences.

The framework is developed using the OSGi paradigm by participating in its progressive concepts of service oriented architectures, like loosely coupling, reusability, composability

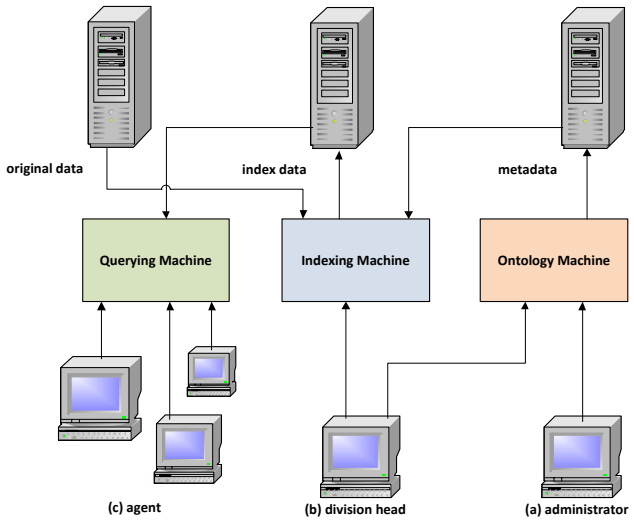


Fig. 2. A black-box view on the new multiple-role framework. (a) The *administrator* defines at least one taxonomy, in order to enable the classification of texts using the *Ontology Machine*. This data will be stored at the *metadata* server. (b) The person heading the investigation (*division head*) defines a case-based ontology using the same machine. In addition he/she can annotate the *original data* using the *Indexing Machine*. Whereas this machine combines original data and *metadata* and transforms it to index data. (c) The *agent* can access the system using the *Querying Machine*, which only has access to reading the *index data*.

and statelessness. Each service encapsulates a single computer linguistic method or technology. In this way, it is ensured that new functionalities based on actual insights of research can be added without adapting the framework's architecture. A qualitative scheme of the service landscape is depicted in Figure 3. The core of the framework is split in three service-tiers:

a) Persistence: In addition to index and metadata server the persistence-tier includes the original data server. It keeps sensitive and evidentiary data strictly separated from other parts of the system. Its interface permits read-only access for the system. The index server provides access to the processed and annotated documents in their intermediate form. The metadata server manages the ontology and taxonomy data in addition to user accounts. In contrast to the original data server, the interfaces of index and metadata server provide full access to the system.

b) Logic: Four low-level services compose the core of the logic-tier. The extracting service is mainly responsible for extracting text from numerous data types, such as .doc, .pdf, .jpg, etc. In addition, several filters for morphological analysis can be applied. The document provider service transforms the extracted data into the document-based intermediate form. It collaborates as service composition with the extracting service, therefore service consumers only need to utilize this service. The main task of the index service is to provide CRUD-operations (acronym for Create, Read, Update, Delete) for accessing index data. It will be used for annotating and querying the document's index by the high-level services of

 ISBN: 978-1-61208-227-1

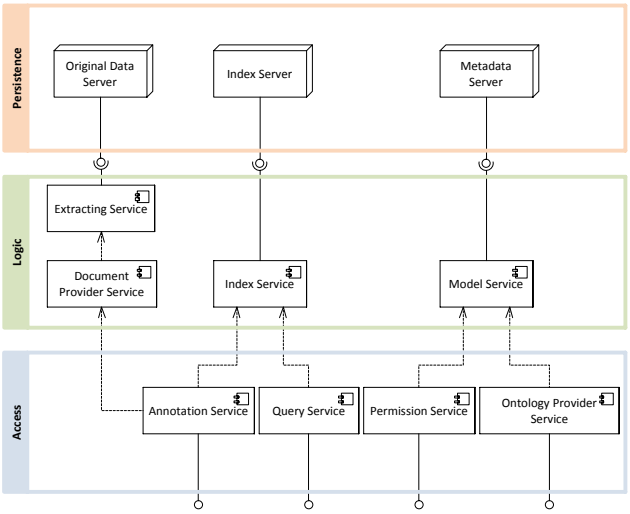


Fig. 3. Architecture overview. The framework's service landscape is divided into three tiers. Each machine (see Fig. 2) can use services from the *access-tier* directly. The *logic-tier* is providing atomic services and service compositions to solve a single problem (The figure shows only a few services exemplified). Accessing these services is only possible by utilizing services of the lower-level tier. The *persistence-tier* is responsible for keeping data and contains the three servers described in Fig. 2. It is only accessible by services of the *logic-tier*.

the access-tier (see c). In the same way, the model service is providing CRUD-operations for accessing metadata. This service is being used by higher-level services working with ontologies and user permissions.

c) Access: The access-tier contains the high-level services for using the low-level services from outside of the core. Subsequently the data is bound to the user interface. The function of high-level services is similar to the facade pattern [10]. The annotation service takes the documents from the document provider service and enriches them with additional user-specified data or data derived by other automatic information extraction services. The index service is used for transforming the data into the document-based intermediate form and pushes them to the index server. The query service fetches index data via the index service from this server, satisfying various filter criteria. The ontology provider service has to perform two tasks. On the one hand, it controls the collaborative access to the ontology model. On the other hand, it provides CRUD-operations on this on a higher level than the model service. Finally, the permission service controls the access permissions of each user to the well-defined data types (see I). Because the user data model is developed in a model-driven way analogue to the data model of ontologies and taxonomies this service collaborates with the low-level model-service. Thus, the same infrastructure as the ontology provider service can be used.

Especially the *logic-tier* is designed to include new functionality, since its services have an open architecture for extending their capabilities. For example, they provide interfaces for adding further services, such as text extraction methods, machine learning algorithms, etc.

V. Conclusion

The development of a high-modular framework for applying methods of natural language processing on forensic data is discussed In this work. Its service-oriented architecture is particularly suitable to include new functionality based on actual insights of research. In this way new knowledge will become available for the points of interest in shorter times.

The main task of the new framework is to support the criminal proceedings in evaluation of forensic data. The concept discussed in this paper is schematically summarized and illustrated in Figure 4. As elucidated, the structure mentioned gives the advantage that accessing and working with the framework is reliably ensured by using the few high-level services exclusively. In contrast, the service-compositions on the lower level can be as complex as needed and can be adapted at any time to achieve improved problem solving.

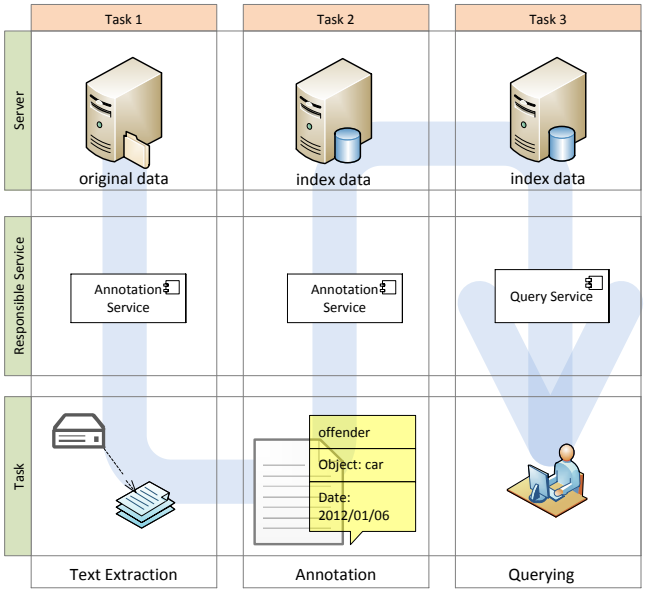


Fig. 4. Task matrix. Task 1) Texts can be extracted from the original data by using the high-level *Annotation Service*. Task 2) Subsequently, extracted texts can be annotated (semi-)automatically and indexed by the same service. To ensure proper processing of this task, an ontology and taxonomy have to be created before using the *Ontology Provider Service* (schematically depicted in Figure 3). Task 3) At this point, each agent can access the indexed data and create knowledge maps using the high-level *Query Service*.

Currently, the first prototype for manual annotation and development of criminalistic taxonomies and ontologies is evaluated in practice. In the next steps, the development of powerful algorithms for automation is emphasized. Especially, ways to extract information from defective texts and hidden semantics will be evaluated and revised.

Acknowledgment

The authors would like to thank the members of the criminal investigation department Chemnitz/Erzgebirge (Germany). We acknowledge funding by "Europäischer Sozialfonds" (ESF),

 ISBN: 978-1-61208-227-1

the Free State of Saxony and the University of Applied Sciences Mittweida.

REFERENCES

[1] Friedrich L. Bauer, *Decrypted Secrets - Methods and maxims of Cryptology*, 1st ed. Berlin, Heidelberg, Germany: Springer, 1997.

[2] Andreas Dengel, *Semantische Technologien*, 1st ed. Heidelberg, Germany: Spektrum Akademischer Verlag, 2012.

[3] Thomas R. Gruber, *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* In Nicola Guarino and Roberto Poli (Eds), *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academic Publishers, 1993.

[4] The Eclipse Foundation 2012, *Eclipse Modeling Framework Project (EMF)*, viewed 03 August 2012, http://www.eclipse.org/emf.

[5] Dave Steinberg, Frank Budinsky, Marcelo Paternostro and Ed Merks, *EMF Eclipse Modeling Framework*, 3rd ed. Boston : Addison-Wesley, 2009.

[6] JTC 1/SC 34/WG 3, *ISO/IEC 13250 - Topic Maps, Information Technology, Document Description and Processing Languages*, 2nd ed. 2002.

[7] Ilya Sutskever, James Martens, and Geoffrey Hinton, *Generating Text with Recurrent Neural Networks*, In Proceedings of the International Conference on Machine Learning (ICML), pp. 1017-1024, 2011.

[8] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde and Hagen Langer, *Computerlinguistik und Sprachtechnologie - Eine Einführung*, 3rd ed. Heidelberg, Germany: Spektrum Akademischer Verlag, 2010.

[9] OSGi^{TM} Alliance 2012, *Technology*, viewed 03 August 2012, http://www.osgi.org/Technology/HomePage.

[10] Erich Gamma, Richard Helm, Ralph Johnson and John Vlissides, *Design Patterns. Elements of Reusable Object-Oriented Software.*, 1st ed. Amsterdam : Addison-Wesley Longman, 1994.

[11] Jeff McAffer, Paul VanderLei and Simon Archer, *OSGi and Equinox - Creating Highly Modular Java Systems*, 1st ed. Boston : Addison-Wesley, 2010.

 ISBN: 978-1-61208-227-1

Particular Requirements on Opinion Mining for the Insurance Business

Sven Rill, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Florian Wogenstein
Institute of Information Systems (iisys)
University of Applied Sciences Hof
95028 Hof, Germany
{srill,jdrescher,dreinel,jscheidt,fwogenstein}@iisys.de

***Abstract*—In this paper, we discuss the work in progress of our current project focusing on opinion mining in the field of insurance business. The main purpose of this project is to improve Opinion Mining methods for the German language and optimize them with special regard to the insurance business. These improved methods make it possible to extract opinions from user-generated texts (in the insurance domain) in a better quality than today. We fetch the required text data for this study from a huge online community website for customers. There, we find a sufficient number of user reviews about insurance companies, which is necessary for our research. Besides the main purpose of this study, another aim is the development of a prototype. This could then be used to monitor the current "crowd's opinion" about insurance products and services. For this reason and in order to understand key aspects of the domain, we collaborate with the *nobisCum Deutschland GmbH*, a German company offering consulting and software development services for the insurance industry. Using one data source and limited evaluation sets, we obtained first results which look promising.**

Keywords-opinion mining; insurance; aspect-based opinion mining; domain-specific orientation.

I. Introduction

The steep rise of user-generated content over the last few years, especially user reviews, on the Web, together with its reliability and the wide public acceptance, makes it therefore important for many companies to observe these relevant sources. Since the amount of reviews, comments, posts, etc. (*user contributions* in short) from only one web platform is, sometimes, too large to be checked manually, methods are required to automate that. There are already some applications done by different groups or companies that claim to be able to reliably extract opinions from German texts. However, that does not always work very well yet. In many cases, the results are restricted to basic opinions (positive, neutral or negative) of certain texts, calculated by counting all opinion bearing words in this text (a kind of document-level sentiment classification, e.g., a bag of words approach).

Using textual data together with its meta information from different review portals, forums, blogs and/or social networks (e.g., Facebook, Twitter) together with advanced Opinion Mining methods such as *Aspect-based Opinion Mining* can help to solve this problem. Thereby, it is also very important to handle domain- and context-specific orientations on generated opinion words and phrases and, in the course of this, also to improve Opinion Mining methods for the German language. In our case, we have to improve them particularly for the insurance domain.

Beside this main purpose, another goal is to include the improved or developed methods and algorithms into a prototype. By using this prototype application, insurance companies will be able to monitor the opinion of their customers or of their whole peer group on certain aspects of their own products "in real time".

In Section II, some related work is introduced. Section III contains a brief description of the *OMVers* (Opinion Mining für die Versicherungsbranche - Opinion Mining for the insurance business) project, our identified issues and a short description about the input data. Section IV mainly depicts the different domain-specific opinion mining steps. In Section V, we want to explain our evaluation process. A short conclusion completes the paper in Section VI.

II. Related Work

Especially in the last few years, a lot of research work has been done in the area of Opinion Mining and Sentiment Analysis. A detailed overview of the whole topic has been given from Pang and Lee [1] and just recently in a survey from Liu and Zhang [2].

A good overview of the *Aspect-based Opinion Mining* approach is also given in the work from Liu and Zhang [2]. In addition to this, Liu defined a model to describe aspects in a document, called *quintuple* [3]. A method to extract the required aspects is presented in research [4].

Furthermore, there are several lists of opinion words for multiple languages. For the English language such lists are SentiWordNet [5], the Subjectivity Lexicon [6], Semantic Orientations of Words [7] and two lists of positive and negative opinion words offered by [8].

O'Hare et al. [9] analyzed blogs in the financial domain to automatically determine the sentiment of the bloggers. Zhuang et al. [10] focused their research on the movie domain and proposed an approach for review mining and summarization.

 ISBN: 978-1-61208-227-1

There is a lot of preliminary work on several aspects in the field of Opinion Mining. As a summary of this work, one can say that a general approach to Opinion Mining, applicable in many domains, does not yet give satisfactory results. On the other hand, domain specific applications are already promising.

The focus of this project is to apply Opinion Mining techniques exclusively in the insurance business.

III. The OMVers Project

OMVers [11] is a collaborative project of the *nobisCum Deutschland GmbH (nobisCum)* and the *Institute of Information Systems (iisys)*. This synergetic cooperation is very important for the project's success. While nobisCum mainly works on developer tasks such as building a front- and a back-end, creating a suitable database, providing an interface for the opinion mining analyze module, etc., iisys can completely dedicate itself to research. During the entire time, the project can benefit from the knowledge as well as the experience of nobisCum about the insurance sector.

A. Major Issues

Before we will be able to analyze user written texts in the insurance domain in such a way that suitable results can be achieved, several subtasks have to be defined and finished. In preparation for this project we identify the following major issues as such tasks (partly based on survey [2]). The final aim is to create opinion quintuples (see [3]) for every analyzed user written text in this domain.

1) Generate Opinion List: Existing lists containing opinion bearing adjectives, nouns or verbs (and phrases) for the German language are not sufficient for our project. Therefore, we need to produce an own list. Our research group considered this issue separately and published a generic approach to generate such lists [12].

2) Improve Opinion Mining for German: Currently, there are some weaknesses in Opinion Mining methods for the German language, e.g., identifying compound nouns. An example for such a noun is *"Versicherungsbetrug"* - *"insurance fraud"*.

3) Identify, extract and group aspects: The automatic extraction of aspects (also known as *features*) from a text is one of the most important parts of this project. Hu and Liu [4] present an interesting two-step-method to perform that. The accuracy of the first step was already improved by Popescu and Etzioni [13]. After extracting, the aspects have to be combined to groups. For that, the OpenThesaurus (see http://www.openthesaurus.de/) will be useful for looking up synonyms.

4) Handle domain-specific opinion words: Since we have our self-generated general opinion list, another important issue is how to handle domain-specific opinion words. For example, in the sentence *"I will change my insurance company"* the verb *change* in the insurance domain expresses a really strong opinion (negative in this case). Whereas in the sentence *"I will change my clothes"* the same verb expresses no opinion (objective sentence). Furthermore, we assume that about 80 % of the entries of our opinion list are universal, in other words domain-independent. However, this assumption has not been verified yet.

5) Handle context-dependent opinion words: Similar to the domain-specific issues mentioned above, the problem of context-dependent opinion words is really relevant for our project, too. Let us have a look at the following two sentences: *"I will change to this insurance company"* and *"I will change to my previous insurance company"*. The verb *"change"* expresses opposite opinions, positive in the first sentence, negative in the second one. There is another special case we can see in the sentence *"In any case, I will change"*. In such a case it would be impossible to determine the opinion expressed by *"change"* without looking at previously written sentences.

Ding et al. [14] proposed an approach to handle opinion words that are context-dependent.

6) Map opinion words to aspects: As soon as we have identified the aspects we can use our list of opinion bearing words to bring them together. This is a huge and important step towards our aim to create opinion quintuples. However, there are already methods for this officially called *aspect sentiment classification*, which we can test and adapt to our specific requirements [4] [14].

B. Additional Issues

In addition to the major issues there are some additional issues, such as extracting and grouping the entity (the insurance company), extracting the opinion holder and time, mining comparative opinions, handling coreference resolutions and extending the approach for multilingualism.

Although the extraction of the entity, opinion holder and time are not minor issues, we want to simplify this aspect for now. Therefore, we define that

- There is only one well-known entity per text,
- The opinion holder is always the author of the text and
- The time is always the publication time of the text.

For that reason, our opinion quintuple at the beginning will look as follows: *(e,* a_j*,* oo_j*, h, t)*. Entity *e*, opinion holder *h* and time *t* are unique per text, while aspect *a* and the opinion value *oo* are not. In a later phase of the project, this current restriction has to be improved. Please note furthermore that,

 ISBN: 978-1-61208-227-1

instead of a basic opinion orientation [3], we work with continuous opinion values between -1 and +1 [12].

After having finished all subtasks described in Section III-A, we expect very satisfactory results. This is the reason why the additional issues are not mandatory at the moment.

C. Input Data

For first investigations, we use user-generated reviews about automobile insurances from an online community for customers called *Ciao* [15]. So far, no method has been developed that would enable us to crawl such reviews automatically, so we have to fetch them manually. Besides the text, the main information of a typical *Ciao* review is a title, the author's alias plus additional user information, a rating value, the publication date, pros and cons defined by the author, an "advisable flag" as well as an average review rating retrieved from other *Ciao* members. We also use some of this additional information for this project.

IV. Domain-Specific Opinion Mining

The aim of our data analysis is to go down to the aspect level of a specific text and create the appropriate opinion quintuples. Thus, it will be possible to see a quantitative as well as qualitative summary of the opinions of different texts.

However, to see how good or bad other methods for this application case work, we have decided to use an iterative approach. This implies that we start our research with a basic *document-level sentiment classification* [1], meaning that the smallest unit is the whole review text. This method is only suitable for documents that contain just one entity (as they do in our case). Nevertheless, the granularity of this technique is probably not fine enough for our application.

After that, we test a method called *sentence-level sentiment classification*, which is an intermediate step before finally reaching the "supreme discipline" of *aspect-based sentiment analysis* or *aspect-based opinion mining*, respectively.

After each step, we perform an evaluation of the results produced by the respective algorithm with a self-defined set of sentences (see Section V). This allows us to measure the quality of every method used as well as to compare them. Furthermore it will help us to improve the methods in an iterative way.

Since the implementation of the first two main steps (*document-level sentiment classification* and *sentence-level sentiment classification*) has almost been completed, we want to describe the ongoing and future work on the *aspect-based sentiment analysis* below, which is partly based on the five tasks (except IV-A) to be performed to build opinion quintuples [2].

A. Domain-specific Opinion List Generation

As already mentioned in Section III-A1, we have published a generic approach to generate opinion lists of phrases [12]. By using this approach, we have already created such a list.

As described in Section III-A4, we now have to deal with domain-dependent opinion words. These words are currently not contained in the list or have the wrong opinion value for the insurance domain (remember the example of the verb *change*). Therefore, in the next step we add such words to our existing list manually or change their opinion value if they already exist.

B. Entity Extraction and Grouping

Currently, this task is treated as an additional issue (see Section III-B). Therefore, we start by using the *Ciao* hierarchy (*Ciao* > insurances > automobile insurances > [list of all insurances]) to determine the (unique) entity of review texts. Later on, we plan to extract entities automatically. This is still required in order to analyze blogs and forums (unstructured sources) and could be handled by using Named Entity Recognition (NER) techniques [16].

C. Aspect Extraction and Grouping

As mentioned in Section III-A3, aspect extraction is a crucial task of our project. Currently, we highly simplify this issue. That means we are currently using a manually produced list of grouped aspects as well as their synonyms prepared by *nobisCum*. Although this approach works quite well, our aim is to extract and group aspects automatically as soon as possible. In addition to that, a method to split compound nouns (a common occurrence in the German language) is needed as well (see Section III-A2).

D. Opinion Holder and Time Extraction

Similar to the entity extraction part, this task is also treated as an additional issue. As described in Section III-C, every user review from *Ciao* contains the author's alias and the publication date, among others. This meta information is presently used to determine the opinion holder as well as the time of the whole review. After finishing the major issues (see Section III-A), this task will be automated, too.

E. Aspect Sentiment Classification

As we now have our list with opinion bearing words (not yet domain-specific) and the required, at the moment static, list of aspects, we can start a simple analysis of reviews to determine the opinion on different aspects. Since our opinion list also contains opinion values for phrases like *"nicht gut"* - *"not good"*, we do not have to handle valence shifters.

We are still at the beginning of this task, i.e., by now we have neither managed context-dependent opinion words (see Section III-A5) nor "but phrases", comparative opinions, etc.

 ISBN: 978-1-61208-227-1

Currently, we use a basic approach to aggregate opinions on various aspects (see Section III-A6).

Later on, we plan to significantly improve this approach. After generating and using a domain-specific opinion list, the next intermediate aim of this task is to handle context-dependent opinion words, which are indispensable to get satisfactory results.

F. Opinion Quintuple Generation

After finishing all previous tasks, we will be able to produce simplified opinion quintuples *(e, a_j, oo_j, h, t)* (see Section III-B). As already mentioned, the quality as well as the complexity of these quintuples are to be improved iteratively by improving the other tasks.

V. Evaluation

To evaluate the quality of the applied and adjusted opinion mining methods, we create a set of sentences related to the insurance domain.

We have considered using a three-class model. The first class, which is the basic one, contains simple sentences such as *"The claim settlement of that insurance company is very good"*. Sentences of this category should be easy to analyze. The second class includes more difficult sentences (that means one subordinate clause, "but phrases", etc.), e.g., *"The employee of the customer service was friendly but not really helpful"*. The third and most intricate class contains sentences with many subordinate clauses as well as irony, e.g., *"They raised the yearly subscription again, I really love this company"*.

Every whole sentence as well as the aspects inside, if any, must be tagged in a machine-readable schema. In order to determine the opinion value of a sentence or aspect we decided to use the following six categories: strong positive (sp), weak positive (wp), neutral (n), weak negative (wn), strong negative (sn) and objective (o).

This set of tagged sentences enables us to measure the quality of our methods. Thus, we can see improvements as well as possible deterioration.

VI. Conclusion

Although user-generated content provides an enormous potential in the area of Opinion Mining, which makes it attractive for companies to pursue real time customer monitoring without relying on the usual polling techniques, it has been used little so far. The main reason for this is that Opinion Mining methods, which are available to companies (especially German ones) still do not work satisfactorily.

We aim to improve Opinion Mining methods for the German language in the course of this project. Early experiments have already shown that a domain- and, of course, a context-dependent approach is indispensable for this. As a next step, we have to check whether the assumption that about 80% of our self-generated opinion list is domain-independent is correct.

At first glance, the extraction of opinions with our adapted methods and our opinion list works quite well, but until now we have only worked with user reviews from a single source as well as within limited evaluation sets. In the near future, we would include blogs and forums and we should then see how different writing styles, text structure and a mix of topics affect our methods.

The first audited results of the project are expected in the fourth quarter of 2012.

Acknowledgment

The *OMVers* project is publicly sponsored by the Bavarian Ministry of Economic Affairs, Infrastructure, Transport and Technology (BStMWIVT).

The Institute of Information Systems is supported by the Foundation of Upper Franconia and by the State of Bavaria.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," vol. 2, p. 1, 2008.

[2] B. Liu and L. Zhang, *Mining Text Data*, 2012, ch. A Survey of Opinion Mining and Sentiment Analysis, pp. 415–463.

[3] B. Liu, *Handbook of Natural Language Processing*, 2010, vol. 2, ch. Sentiment Analysis and Subjectivity.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004, pp. 168–177.

[5] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, vol. 25, May 2010, pp. 2200–2204.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, *Human Language Technology Conference - Conference on Empirical Methods in Natural Language Processing*, 2005, ch. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.

[7] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations of words using spin model," in *Proceedings of the 43rd Annual Meeting of the ACL*, 2005, pp. 133–140.

[8] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proceedings of the 14th International World Wide Web conference (WWW-2005)*, 2005, pp. 342–351.

[9] N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. F. Smeaton, "Topic-dependent sentiment analysis of financial blogs," in *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement (TSA-09)*, 2009, p. 9.

 ISBN: 978-1-61208-227-1

[10] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM-2006)*, 2006, pp. 43–50.

[11] Institute of Information Systems (iisys), "Analytical Information Systems: Projects," http://www.iisys.de/en/research/research-groups/analytical-information-systems/projects.html, (accessed September 21, 2012).

[12] S. Rill, J. Drescher, D. Reinel, J. Scheidt, O. Schütz, F. Wogenstein, and D. Simon, "A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM-2012)*, 2012.

[13] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Oct 2005, pp. 339–346.

[14] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," *WSDM08*, pp. 231–240, 2008.

[15] Ciao GmbH, "Preisvergleich und Testberichte bei Ciao," http://www.ciao.de/, (accessed June 4, 2012).

[16] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," vol. 30, pp. 3–26, 2007.

 ISBN: 978-1-61208-227-1

How to Support Prediction of Amyloidogenic Regions - The Use of a GA-based Wrapper Feature Selections

Olgierd Unold
Institute of Computer Engineering, Control and Robotics
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
olgierd.unold@pwr.wroc.pl

***Abstract*—In this paper, we address the problem of predicting the location of amyloidogenic regions in proteins. To support this process we used a genetic algorithm-based wrapper feature subset selection. The wrapper feature subset selection approach is about choosing a minimal subset of features that satisfies an evaluation criterion. We find that most of the machine learning algorithms taken from the WEKA software achieved no worse Accuracy over reduced dataset than over the non-reduced dataset. Moreover, research has confirmed the observations of other researchers, that amino-acids have highly position-dependent propensities.**

***Keywords-Amyloid Proteins; Data Mining; Feature Subset Selection*.**

I. INTRODUCTION

In this paper, we are interested in predicting amyloid proteins. A protein becames amyloid due to an alteration in its secondary structure. A key role in conversion of proteins from their soluble state into fibrillar, beta-structured aggregates, play short sequences, named *hotspots*. Amyloid proteins cause a group of diseases called amyloidosis, such as Alzheimer's, Huntington's disease, and type II diabetes. Symptoms of amyloidosis depend on the organs and tissues amyloid affects.

A laboratory test of a large number of peptides for determining the presence of amyloid protein is in fact theoretically possible, but practically it is not feasible. Therefore, computational methods are commonly used to overcome this limitation. Over the last few years, various computational methods - among existing ones - have been developed to detect these *hotspots* in proteins, like AmylPred [1], Pafig [2], FoldAmyloid [3], and Waltz [4] (for available software dedicated to this task see [5]). However, these methods are very often time consuming algorithms (like 3D Profile method). It is useful, therefore, to use less demanding methods, such as machine learning algorithms, moreover joined with a reduction of the size of the analyzed data.

In this work, we carry an elaborate performance study of different machine learning classification algorithms and feature subset selection (FSS) method applied to Amyloidogenic dataset. All of the algorithms and the wrapper were taken from the Weka machine learning software [6]. The methods over datasets (reduced and non-reduced) were compared in terms of Accuracy.

The remainder of this paper is organized as follows. Section 2 includes state-of-the-art of the problem of predicting amyloidogenic regions, and Section 3 describes Amyloidogenic dataset, FSS method mining the data, and a set of classifiers. Section 4 shows the results obtained, and finally the conclusions are drawn in Section 5.

II. STATE OF THE ART

As established recently [7], there is the strong association between protein fibrils and amyloid diseases, such as Alzheimer's disease, Parkinson's disease, transmissible spongiform encephalopathies, and type II diabetes. It was also observed [8], that amyloids can be formed from short peptide fragments, called hotspots. These strings when exposed to the environment can cause the changeover of native proteins into amyloid state.

Since it is not possible to experimental test all possible protein sequences, several computational tools for predicting amyloid chains have emerged. Most of them are based on physicochemical grounds or structural denominators, like AmylPred [1], Pafig [2], FoldAmyloid [3], and Waltz [4]. However, to our knowledge, no one has used a genetic algorithm-based wrapper feature subset selection method to solve problem under study.

In this paper, we propose a feature subset selection to support predicting amyloid peptides. More over, we are not interested in a time-consuming investigation of physicochemical properties of the amino acids [9], [10], [11], [12] or gaining insight into aggregation propensity [13], [14], [15]. What we are trying to do is to predict amyloidogenic feature of peptide sequence, having no additional knowledge about this sequence. Feature subset selection methods are taken from general-oriented, freely available WEKA software.

III. DATA AND METHODS

A. Data

In our work, we used so-called Waltz amyloidogenic dataset [4]. This is experimentally verified database consisting of 116 amyloidogenic hexapeptides and 162 non-

 ISBN: 978-1-61208-227-1

amyloid-forming hexapeptides. According to its authors, to obtain these data more than 200 peptide sequences were inspected using different structural and biophysical methods. Advantage of Waltz dataset over the others is that it contains experimentally determined structures. Very often amyloid datasets created by various modeling methods – computationally identified – (like the 3D profile methods) [13]) are prone to producing erroneous results. Note that the RosettaDesign potential energy function used in the 3D profile methods is based on heuristic simulated annealing.

B. Classifiers

The experiments were conducted comparing the classification Accuracy of 13 classification methods implemented in the Weka software. Here we briefly list the classifiers that we used:

- *Naive Bayes* and *BayesNet* – classifiers based on the Bayesian Theorem in which it is assumed that the attributes have equal weight and are conditionally independent,
- *Support Vector Machine* – algorithm trying to find a hypersurface in the space of possible inputs,
- *C4.5*, *Random Tree*, *REPTree*, *RandomForest*, *ADTRee* – methods creating a hierarchy of nodes, each associated with a decision rule on one attribute. ADTree creates alternating decision trees, RandomTree builds a tree considers a given number of random features at each node, RandomForest builds random forests using Breiman's algorithm in which multiple random trees vote on an overall classification for the given set of inputs. REPTree uses reduced-error pruning to speed up a learning process, C4.5 algorithm improves Quinlan's method for decision tree induction,
- *JRip* – classifier generating rules, which can transformed from or in decision trees. JRip is the WEKA version of RIPPER, which is a rule-based learner that builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules,
- *MultiLayer Perceptron* – kind of simple neural network classifier, in which backpropagation algorithm calculates connection weigths given a fixed network structure,
- *KStar* – an instance-based classifier using an Entropic Distance Measure. It provides a consistent approach to handling of symbolic attributes, real valued attributes and missing values,
- *AdaBoost* – one of the most popular boosting algorithms. Boosting is an iterative method in which new model is efected by the performance of those built previously. This is achieved by assigning proper weights to learning instances in each iteration,
- *END* – a meta-classifier for handling multi-class datasets. The main idea of meta-classification is to represent the judgment of each classifier (SVM-based) for each class as a feature vector, and then to re-classify again in the new feature space. The final decision is made by the meta-classifiers instead of just linearly combining each classifiers judgment.

More information on implementing in the Weka software classifiers is presented in [6].

The quality of our predictions was evaluated using the commonly used standard value Accuracy, which is measured by the number of correct results, the sum of true positives and true negatives, in relation to the number of tests carried out

$$Accuracy = (\frac{TruePositives + TrueNegatives}{Total}) \times 100 \quad (1)$$

where True Positives are correctly (i.e., as amyloidegenic peptides) recognized positive examples, True Negatives - correctly recognized negatives (i.e., as non-amyloidogenic ones).

C. GA-based Wrapper Feature Selection

Feature selection methods can be put into two main categories from the point of view of a methody output. One category, called filter approach, comprises methods ranking features according to the same evaluation criterion; the other, called the wrapper approach, consists of methods choosing a minimum subset of features that satisfies an evaluation criterion.

It was proved that the wrapper approach produces the best results out of the feature selection methods [16], although this is a time-consuming method since each feature subset considered must be evaluated with the classifier algorithm. In the wrapper method, the attribute subset selection algorithm exists as a wrapper around the data mining algorithm and outcome evaluation. The induction algorithm is used as a black box. The feature selection algorithm conducts a search for a proper subset using the induction algorithm itself as a part of the evaluation function. GA-based wrapper methods involve a genetic algorithm (GA) as a search method of subset features.

GA is a random search method, effectively exploring large search spaces [17]. The basic idea of GA is to evolve a population of individuals (chromosemes), where individual is a possible solution to a given problem. In case of searching the appropriate subset of features, a population consists of different subsets evolved by a mutation, a crossover, and selection operations. After reaching maximum generations, algorithms returns the chromosome with the highest fitness, i.e. the subset of attributes with the highest Accuracy.

 ISBN: 978-1-61208-227-1

IV. EXPERIMENTAL RESULTS

The comparison was performed using a recently published Amyloidogenic dataset, composed by 116 hexapeptides known to induce amyloidosis and by 162 hexapeptides that do not induce amyloidosis [4]. In our experiments, we randomly split the Amyloidogenic database into 10 equally folds, and use a 10-fold cross validation method to determine the classification Accuracy.

A k-fold cross validation (k-fold cv) is a well-established statistical method of evaluating a learner, combining training and validation phases [18]. In k-fold cv the data is partioned into k folds, and next subsequently k iterations of learning and testing are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k-1$ folds are used for learning.

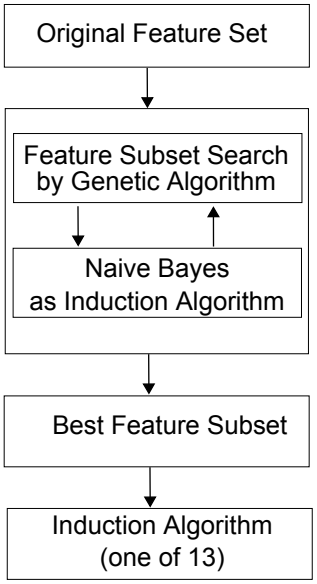


Figure 1. GA-based wrapper feature selection with Naive Bayes as an induction algorithm evaluating feature subset

We employ 13 commonly used machine learning algorithms: BayesNet, NaiveBayes, MultiLayerPerceptron (MLP), Support Vector Machine (SMO), KStar, AdaBoost, END, JRip, C4.5, Random Tree, REPTree, RandomForest, ADTree and GA-based wrapper approach for feature selection. GA-based wrapper methods involve a genetic algorithm as a search strategy of subset features, and one of the machine learning method as an induction algorithm, in this paper NaiveBayes (see Fig. 1). The study [19] noted that no significant difference exists between results achieved by various induction algorithms used in a GA-based wrapper method. All of the classification algorithms and the wrapper were taken from the Weka software [6], all of them used default parameters.

Table 1 summarizes the performances of the 13 compared methods over reduced (denoted as a Dataset 1-3-5), and non-reduced Amyloidogenic dataset (Dataset 1-2-3-4-5-6). Ten of the thirteen methods gained better results over reduced dataset, although the results were not confirmed statistically. What is interesting, the feature selection method chooses only three from six amino acids as important in hexapeptide, in positions 1, 3, and 5. Note that such observations were also made in laboratory experiments. Maurer-Stroh et al. [4] recorded the strong position-specific tendencies of the different amino acids for forming amyloid structures.

Table I
THE PERFORMANCES IN TERMS OF ACCURACY OF THE MACHINE LEARNING METHODS OVER REDUCED AND NON-REDUCED AMYLOIDOGENIC DATASET. THE HIGHER ACCURACY IN A ROW IS INDICATED IN BOLD.

Method	Dataset 1-3-5	Dataset 1-2-3-4-5-6
BayesNet	**68.02**	65.57
NaiveBayes	**66.93**	65.57
MLP	64.76	**72.34**
SMO	68.37	**73.81**
KStar	63.69	**65.50**
AdaBoost	**69.80**	68.37
END	**64.03**	60.81
JRip	**69.06**	68.74
ADTree	**73.77**	69.81
C4.5	**64.03**	60.81
RandomTree	**66.55**	65.91
REPTree	**66.90**	66.28
RandomForest	**66.56**	65.15

V. CONCLUSION

The problem of predicting the amyloidogenic regions in proteins was addressed. Our analysis showed that the use of feature subset selection can support efficiently this task. In most cases machine learning methods have achieved better results over reduced dataset. In addition, methods processed twice smaller learning set.

It is worth noticing that the overall best results have been gained by Support Vector Machine over non-reduced data (73.81 % of Accuracy), and Alternating Tree but over reduced dataset (73.77 %). If SVM is quite often used in prediction different regions in protein chains [2], the ADTree gives interpretable and understandable by human results.

The performed computational experiments confirm also laboratory studies over proteins, in which the strong position dependency of residues are observed.

REFERENCES

[1] K. Frousios, V. Iconomidou, C. Karletidi, and S. Hamodrakas, "Amyloidogenic deteminants are usually not buried," *BMC Structural Biology*, vol. 9, p. 44, 2009.

[2] J.Tian, N. Wu, J. Guo, and Y. Fan, "Prediction of amyloid fibril-forming segments based on a support vector machine," *BMC Bioinformatics*, vol. 10 (Suppl 1):S45, 2009.

 ISBN: 978-1-61208-227-1

[3] S. Garbuzynskiy, M. Lobanov, and O. Galzitskaya, "An introduction to variable and feature selection," *Bioinformatics*, vol. 26, pp. 326–332, 2010.

[4] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. L. de la Paz, I. Martins, J. Reumers, K. Morris, A. Copland, L. Serpell, L. Serrano, J. Schymkowitz, and F. Rousseau, "Exploring the sequence determinants of amyloid structure using position-specific scoring matrices," *Nat Methods*, vol. 7, pp. 237–242, 2010.

[5] S. Hamodrakas, "Protein aggregation and amyloid fibril formation prediction software from primary sequence: Towards controlling the formation of bacterial inclusion bodies," *FEBS Journal*, vol. 278, no. 14, pp. 2428–2435, 2011.

[6] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques. Third edition*. Morgan Kaufmann, 2011.

[7] L. Goldschmidt, P. Teng, R. Riek, and D. Eisenberg, "Identifying the amylome, proteins capable of forming amyloid-like fibrils," *PNAS*, vol. 107(8), pp. 3487–3492, 2010.

[8] O. Galzitskaya, S. Garbuzynskiy, and M. Lobanov, "Prediction of amyloidigenic and disordered regions in protein chains," *PLoS Computational Biology*, vol. 2(12), p. e177, 2006.

[9] G. Tartaglia, A. Cavalli, R. Pellarin, and A. Caflisch, "Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences," *Protein Sci*, vol. 14(10), pp. 2723–2734, 2005.

[10] K. DuBay, A. Pawar, F. Chiti, J. Zurdo, C. Dobson, and M. Vendruscolo, "Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains," *J Mol Biol*, vol. 341(5), pp. 1317–1326, 2004.

[11] A. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nat Biotechnol*, vol. 22(10), pp. 1302–1306, 2004.

[12] S. Idicula-Thomas and P. Balaji, "Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation," *Protein Eng Des Sel*, vol. 18(4), pp. 175–180, 2005.

[13] M. Thompson, S. Sievers, J. Karanicolas, M. Ivanova, D. Baker, and D. Eisenberg, "The 3D profile methods for identifying fibril-forming segments of proteins," *Proc Natl Acad Sci USA*, vol. 103, pp. 4074–4078, 2006.

[14] S. Yoon and W. Welsh, "Detecting hidden sequence propensity for amyloid fibril formation," *Protein Sci*, vol. 13(8), pp. 2149–2160, 2004.

[15] M. L. D. L. Paz, K. Goldie, J. Zurdo, E. Lacroix, C. Dobson, A. Hoenger, and L. Serrano, "De novo designed peptide-based amyloid fibrils," *Proc Natl Acad Sci USA*, vol. 99(25), pp. 16 052–16 057, 2002.

[16] X. Zhiwei and W. Xinghua, "Research for information extraction based on wrapper model algorithm," in *2010 Second International Conference on Computer Research and Development*, Kuala Lumpur, Malaysia, 2010, pp. 652–655.

[17] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[18] F. Mosteller and J. Turkey, *Data Analysis, Including Statistics*, ser. Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968.

[19] O. Unold, M. Dobrowolski, H. Maciejewski, P. Skrobanek, and E. Walkowicz, "A GA-based wrapper feature selection for animal breeding data mining," *Lecture Notes in Computer Science*, vol. 7209, pp. 200–209, 2012.

 ISBN: 978-1-61208-227-1

Application of Optimisation-based Data Mining Techniques to Medical Data Sets: A Comparative Analysis

Zari Dzalilov, Adil Bagirov, Musa Mammadov
The Centre for Informatics and Applied Optimisation
School of Science, Information Technology and Engineering,
University of Ballarat, Victoria, Australia
z.dzalilov@ballarat.edu.au, a.bagirov@ballarat.edu.au, m.mammadov@ballarat.edu.au

Abstract - **Computational methods have become an important tool in the analysis of medical data sets. In this paper, we apply three optimisation-based data mining methods to the following data sets: (i) a cystic fibrosis data set and (ii) a tobacco control data set. Three algorithms used in the analysis of these data sets include: the modified linear least square fit, an optimization based heuristic algorithm for feature selection and an optimization based clustering algorithm. All these methods explore the relationship between features and classes, with the aim of determining contribution of specific features to the class outcome. However, the three algorithms are based on completely different approaches. We apply these methods to solve feature selection and classification problems. We also present comparative analysis of the algorithms using computational results. Results obtained confirm that these algorithms may be effectively applied to the analysis of other (bio)medical data sets.**

Keywords – data mining; optimisation; cystic fibrosis; tobacco control.

I. INTRODUCTION

Optimization plays a fundamental role in designing efficient data mining techniques. For example, the support vector machine algorithms are among the most efficient data classification techniques [7]. Such techniques have been applied to medical data sets over the last two decades to solve wide range of problems including feature selection, data classification and prediction problems. Despite of significant developments in this area there is still a lot of evaluation work to be performed due to the fact that medical data sets are diverse and it is difficult to formulate a unique criterion for all of them. Comparison of specific medical data sets was done in [5, 8,18,19, 23].

In this paper, we present the results of application of three optimization-based data mining algorithms to two different data sets: the CF (Cystic Fibrosis) data set [14] and the Tobacco Control data set [12,13]. These algorithms can be applied to solve three different problems of data mining: data regression, data classification and clustering. All three algorithms are based on nonlinear models, and therefore, can detect nonlinear relationships between both features and instances. Data sets used in evaluation are completely different which helps to have a clear picture about efficiency and accuracy of algorithms used in the comparison. Moreover, such data sets have not been thoroughly studied using data mining techniques.

The paper is organized as follows. Section II describes the two data sets used for the analysis, Section III briefly describes the algorithms and Section IV presents the results obtained by applying these algorithms to the data sets. Section V concludes the paper.

II. DATA SET DESCRIPTION

The two data sets for analysis in this paper are the Cystic Fibrosis and Tobacco Control. These are described in more detail in this section.

A. Cystic Fibrosis data set

Cystic fibrosis is the most common fatal genetic disorder in the Caucasian population [11]. Clinical scoring systems for the assessment of Cystic fibrosis disease severity have been used for almost 50 years without being adapted to the milder phenotype of the disease in the 21st century [9,11,14, 20]. A fresh approach is needed for the development of comprehensive CF disease severity scales, which may be used as a disease predictor. The goal is to develop a scoring system to assess the longitudinal process of Cystic Fibrosis.

We propose to develop a new clinical scoring system by employing various statistical tools and optimisation methods. We previously identified an approach for developing a disease severity scale [14]. We now propose to refine this scale by using a hybrid model combining mathematical optimisation and data mining approach. We evaluate mathematical optimisation methods that can be used for the solution of feature selection problems. The advantage of these methods is that they allow one to consider datasets with an arbitrary number of classes.

The evaluation is based on the Cystic Fibrosis database from the cohort at the Royal Children's Hospital in Melbourne. The data base contains 212 subjects, with 69 features and 3 expert defined or 5 CAP defined classes. The methods applied to this data set are the *Linear Least Squares Fit (LLSF)* [21, 22] and the *Heuristic Algorithm for Feature Selection* [1,2,3]. Both of these methods explore relationships between features and classes. They allow to analyse data sets with an arbitrary number of classes. Our results show that the methods applied are helpful to determine the contribution of features to the effectiveness of disease severity classification, which is the main point for developing *a clinical scoring system.* The results obtained can be used in preparatory work for clinical trials. However, more data points are needed to finalize a clinical score, by re-running these methods in the larger data set.

 ISBN: 978-1-61208-227-1

B. Tobacco Control data set

Smoking is one of the leading causes of death around the world and as such, control of tobacco use is an important global public health issue [10]. The large detrimental impact, that smoking already has on a public health has the potential to become even greater as the population worldwide ages and dementia prevalence increases. Controlling tobacco smoking and determining effective policies is difficult because of the complexity of human nature. Nevertheless, there have been numerous attempts to describe and understand the effectiveness of tobacco control policies to smokers' quitting behaviour. Linear regression and logistic regression are currently very popular statistical techniques for modelling and analysing complex data in tobacco control systems [24]. However, in tobacco markets, numerous inter-related factors interact with tobacco control policies in non-trivial fashion, such that policies and control outcomes are non-linearly related. The use of linear and logistic regression is therefore fundamentally limited due to their inability to deal with these complex relationships. Compared with traditional statistical techniques, optimization-based methods have the potential to be more effective analysis tools of complex tobacco control systems. The Tobacco Control data set was collected in Australia for studying and evaluating the psychosocial and behavioural impact of diverse tobacco control policies to smokers' behaviour. This data set was collected in the frame of ITC project [6, 15, 17] and is shown in Figure 1.

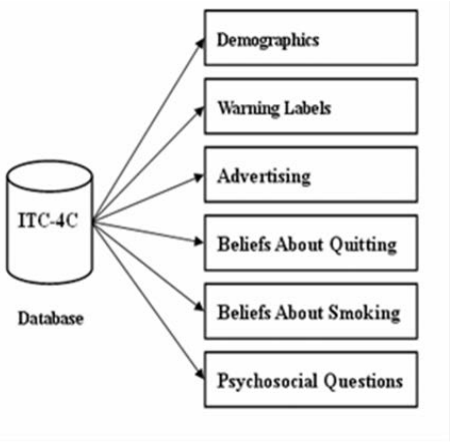


Figure 1. Description of Tobacco Control data set.

The data set was used to evaluate the optimization-based approaches in data mining [12, 13]. The aim of the exercise was to find *clusters* of smokers with similar beliefs about quitting for predicting the rate of quitting attempt. We apply the following optimization-based algorithms to the tobacco control data: the *Linear Least Squares Fit (LLSF)*, the *Heuristic Algorithm for Feature Selection*, and the *Modified Global k-means Algorithm* [4]. We have obtained some promising preliminary results for covering potential solutions in tobacco control. A brief description of algorithms is given below

III. ALGORITHM DESCRIPTION

The three algorithms used in this paper are now described in more detail.

A. The Linear Least Squares Fit (LLSF)

Let M be the number of all features and C be the number of classes. Data is given in the form of two matrices. Matrix $A=(a_{ij}),\ i=1,...,N,\ j=1,...,M,$ where N is the number of samples. Matrix $B=(b_{ik}),\ i=1,...,N,\ k=1,...,C,$ where vector $(b_{i1},...,b_{iC})$ describes class information for the row/sample i; $b_{ik}=1$ if sample i belongs to class k and $b_{ik}=0$ otherwise. Consider the matrix $X=(x_{jk}),\ j=1,...,M,\ k=1,...,C,$ that describes the relationships between features and classes. LLSF aims to find matrix X by minimizing the function $f(X)=\|AX-B\|^2$.

Take any feature p and eliminate it from the list of features. Denote

$A(p)=(a_{ij}),\ i=1,...,N,\ j=1,...,p-1,p+1,...,M$

and

$X(p)=(x_{jk}),\ j=1,...,p-1,p+1,...,M,k=1,...,C.$

Let

$X^*(p) = \arg\min\{\|AX(p)-B\|^2:\ X(p)\in R^{M\times C}\}$

Matrix $X^*(p)$ can be used to predict all samples $i=1,...,N$ using all features except p: $j=1,...,p-1,p+1,...,M.$ Denote the average accuracy obtained in this way by $E(p)$. Clearly, the inequality $E(p_1) < E(p_2)$ for some features p_1 and p_2 means that the accuracy decreases more if we eliminate feature p_1 rather than p_2. Therefore, we can say that feature p_1 is more important than p_2; we write in this case $p_1 \square p_2$, arranging all features in a way that:

$E(j_1) \le E(j_2) \le ..... \le E(j_M)$ we obtain the order of features by their importance in ascending order $j_1 \succ j_2 \succ .... \succ j_M$.

Previously, we applied LLSF to the data set on Cystic Fibrosis [14]. Our preliminary results show that the methods applied are helpful in developing a clinical scoring system on Cystic Fibrosis.

B. Heuristic Algorithm for Feature Selection

This algorithm for the solution of the feature selection problem is based on techniques of convex programming [2] and allows one to consider data sets with an arbitrary number of classes. We consider feature selection in the context of the classification problem. The algorithm calculates a subset of

 ISBN: 978-1-61208-227-1

most informative features and a smallest subset of features. The first subset provides the best description of a dataset whereas the second one provides the description which is very close to the best one. A subset of informative features is defined by using certain thresholds. The values of these thresholds depend on the objective of the task.

The purpose of the feature selection procedure is to find the smallest set of informative features possible for the object under consideration, which describes this object from a certain point of view. The following issues are very important for understanding the problem:

- It is convenient to consider (and define) informative features in the framework of classification. In other words it is possible to understand whether a certain feature is informative for a given example if we compare this example with another one from a different class.
- Our goal is to find a sufficiently small set of informative features and to remove as many superfluous features as possible. Note that this problem can have many different solutions.
- It follows from the above that the set of informative features, which describe a given object, is a categorical attribute of this object. This is also a fuzzy attribute in a certain informal sense. It leads to the following heuristic conclusion: it is useless to apply very accurate methods in order to find this set. However, if we use heuristic (not necessarily very accurate) methods we need to have experimental confirmation of the results obtained.

This algorithm proceeds as follows. First, we find centres of each class and remove a feature which gives a smallest distance among all features. Using any classification method we compute classification accuracy with the rest of features. Then, we update the centres with one removed feature and remove again the closest feature and apply again the classification method. If the classification accuracy is reduced significantly, we stop and accept the rest of the features as the most informative. Otherwise, the algorithm continues. The "significant" reduction in accuracy is defined by the user. In our calculations this reduction is 1%.

C. Modified Global k-means Algorithm

Cluster analysis, also known as unsupervised data classification, is an important subject in data mining. Its aim is to partition a collection of patterns into clusters of similar data points. In cluster analysis we assume that we have been given a finite set of points A in the n-dimensional space R^n, that is $A = \{a^1,...,a^m\}$, where $a^i \in R^n, i = 1,...,m$. There can be different types of clustering. In this paper, we consider the hard unconstrained partition clustering problem, that is the distribution of the points of the set s A into a given number *k* of disjoint subsets $A^1,...,A^k$ with respect to predefined criteria such that:

1) $A^j \neq \varnothing,\ j = 1,...k;$

2) $A^j \bigcap A^l = \varnothing,\ j,l = 1,...k,\ j \neq l;$

3) $A = \bigcup_{j=1}^{k} A^j;$

4) no constraints are imposed on clusters $A^j,\ j = 1,...k.$

The sets $A^j,\ j = 1,...k;$ are called clusters. We assume that each cluster A^j can be identified by its centre (or centroid) $x^j \in R^n,\ j = 1,...k.$ Then the clustering problem can be reduced to the following optimisation problem (see [1, 2]):

find a collection $\bar{x} = (\bar{x}^1,...,\bar{x}^p)$ of *n*-dimensional vectors, which is a solution to the following problem:

$$\min f(x^1,...,x^p) = \sum_{i=1}^{m} \min_{j=1,...,p} \left\| x^j - a^i \right\|^2 \quad \text{subject to}$$

$$x^j \in R^n,\ j = 1,...,p. \qquad (1)$$

Here, ||. || is an Euclidean norm. The problem (1) is also known as minimum sum of squares-clustering problem [16]. It is assumed that a given set of points contains *p* clusters and the solutions $\bar{x}^1,...,\bar{x}^p$ to the problem (1) are the centres of these clusters.

IV. Results and Discussion

A. Cystic Fibrosis Data Set

Statistical approaches were tested on the Cystic Fibrosis database from the cohort at the Royal Children's Hospital in Melbourne. After data preparation which included expert-opinion of an individual's clinical severity on a 3 point-scale (mild, moderate and severe disease), two multivariate techniques were used to establish a method that would have a better success in feature selection and model derivation. The methods were *Canonical Analysis of Principal Coordinates (CAP'*) and *Linear Discriminant Analysis*. A 3-step procedure was performed which included selection of features, extracting 5 severity classes from the 3 original classes as defined by medical experts and establishment of calibration datasets.

Two different methods, based on optimisation techniques have also been used for the solution of the feature selection problems. These methods are the *Linear Least Squares Fit (LLSF)* and the *Heuristic Algorithm for Feature Selection*. Since the data was already broken up into a number of classes the *Modified Global k-means Algorithm* was not used on this data set. We apply all methods to the data set containing 212 subjects, with 69 features and 3 expert defined classes or 5 CAP defined classes. The results are shown in Table 1. The methods are enumerated as follows: *LLSF*-1, *Feature Selection*-2, *Statistical*-3.

All three methods (1, 2, 3) indicate that the following features are the most significant:

 ISBN: 978-1-61208-227-1

TABLE I. CYSTIC FIBROSIS SIGNIFICANT FEATURES

Method	Significant Features
1, 2, 3	19,21,29,30, 31, 35, 47
1, 2	18,22,23, 26, 32,40, 41,48, 49, 50, 51, 55, 57

Information from Table 1 can be used in clinical practice. The highest preference should be given to the features in the first line, since they have been confirmed by all three methods. The features can be identified in the data set through the codes shown in Table 2.

TABLE II. CYSTIC FIBROSIS FEATURE CODES

Feature	Code	Feature	Code
19	NORESP	26	NTSURG4
21	NODAYS	32	CULTandBMI
29	FEV1P	40	ANTIBO
30	FVCP	41	ANTIBO
31	FEFP	48	THERAP2
35	BMIPCT	49	THERAP3
47	HERAP1	50	THERAP4
18	NOVIS	51	THERAP5
22	HTCOUR	55	NSUP1
23	HTDAYS	57	LTOXYYNU

Table 2 provides all the significant features that have been identified. The meaning of the features can be found from the data base of the Royal Children Hospital and is not included here.

Our preliminary results show that the optimization methods applied are helpful in developing a clinical scoring system. However, more data points are needed to finalize a clinical score, by re-running methods in the larger dataset.

B. Tobacco Control Dataset

As a preliminary work, we applied the *Linear Least Squares Fit*, the *Heuristic Algorithm for Feature Selection* and the *Modified Global k-means* algorithms to the four data sets containing:

- Data set 1 – 1458 subjects, with 71 features
- Data set 2 – 1477 subjects, with 69 features
- Data set 3 – 1260 subjects, with 60 features
- Data set 4 – 1350 subjects, with 60 features

1) Linear Least Squares Fit

Results obtained by LLSF algorithm for Data Set 1 are illustrated in Figure 2. This figure shows dependence of the classification accuracy on the number of features. One can see that this algorithm does not allow one to find the subset of most informative features.

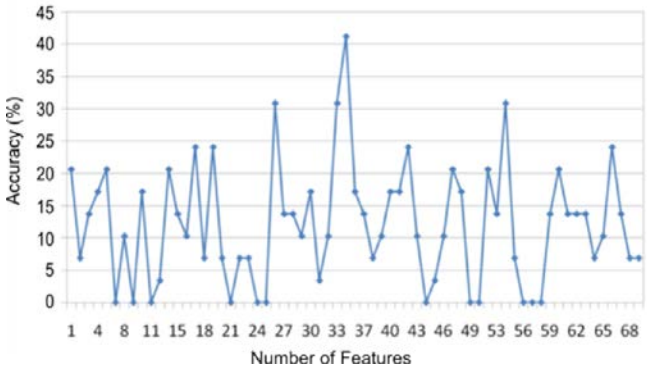


Figure 2. Results for Data Set 1 using LLSF algorithm.

2) Heuristic Algorithm for Feature Selection

Illustrations of numerical results of applications of the *Heuristic Algorithm for Feature Selection* to the Data Sets 1, 2 and 3 are shown on Figures 3-5 below. These figures show dependence of the classification accuracy on the number of features.

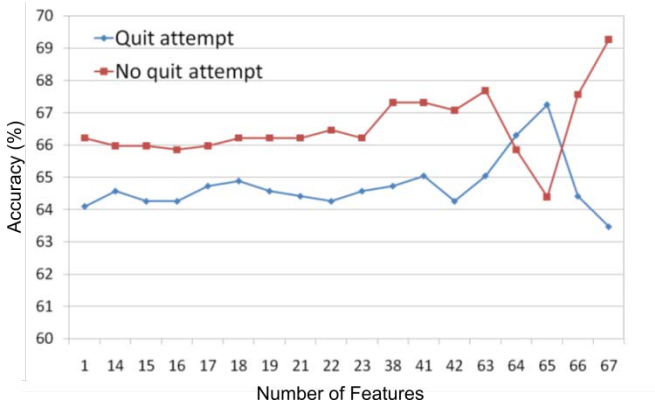


Figure 3. Results for data set 1 using Heuristic Feature Selection algorithm.

Figure 3 shows features obtained from the first survey. The red line shows the smokers with no intention to quit, with features 67 and 65 having maximal and minimal accuracies respectively. The blue line shows features associated with smokers with the intention to quit. This time the behaviour is reversed with the minimal accuracy for feature 67, and the maximal accuracy for feature 65. In general, the two feature sets exhibit complementary behaviour.

 ISBN: 978-1-61208-227-1

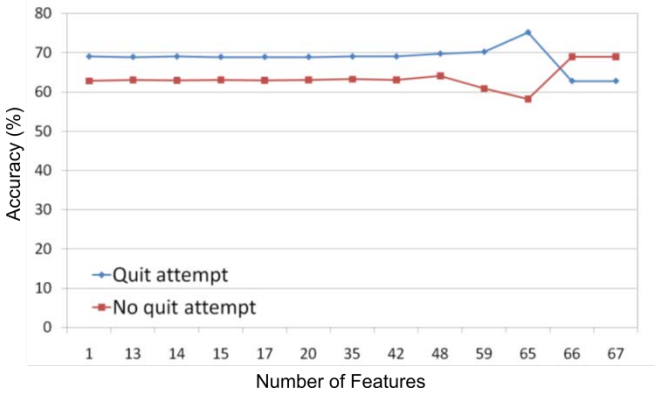


Figure 4. Results for data set 2 using Heuristic Feature Selection algorithm.

Figure 4 shows features obtained on the data from the second survey. The same maximal and minimal features are obtained for both datasets as in the previous case, however, in contrast to the data of the previous survey, the observed fluctuations in feature accuracy is much less pronounced.

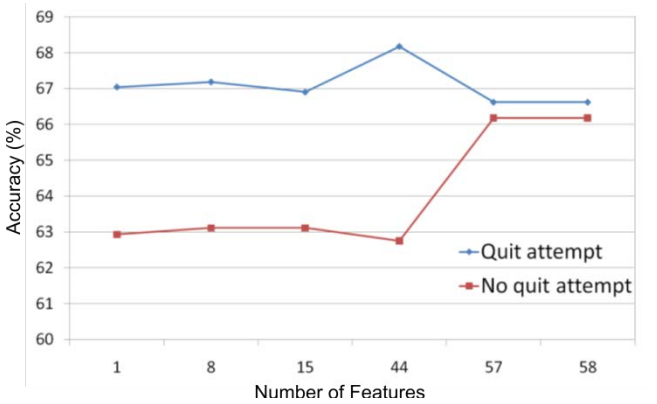


Figure 5. Results for data set 3 using Heuristic Feature Selection algorithm.

Figure 5 shows features obtained on the data from the third survey. This time the maximal feature for smokers with no intention to quit (red line) is 57 and the minimal feature is 44. For smokers with the intention to quit (blue line) the maximal feature is 44 and the minimal feature is 57. The figure once again shows complementary behaviour of these feature sets. At the same time however, the increased difference between the two feature sets, compared to previous sets, is apparent.

Some examples of the meanings of the significant features for quit attempt:

- 5 - Thoughts about danger of smoking
- 7 - Thoughts about harm to self
- 35 - Confidence to quit smoking
- 37 - Perception of quitting difficulty
- 66 - Disapproval of smoking in society
- 36 – How easy or hard to stop smoking permanently

The results obtained allow us to draw the following conclusions:

- *Heuristic Algorithm for Feature Selection* revealed complementary pattern of features for quitters and non-quitters
- Overlapping features are likely to be important in tobacco control programs

As a result of the feature selection algorithm we found a number of significant features, associated with quit attempts. The features in the neighbourhood of the maximal point can also be considered to be associated with quit attempts, with a reasonable level of accuracy. The maximum accuracy attained is 67.24. The list of smokers' response to the most significant questions in predicting the rate of quitting attempt is shown in Table 3.

TABLE III. TOBACCO CONTROL DATA SET SIGNIFICANT FEATURES

Data Set	Significant Features
Data Set 1	5, 7, 35, 37, 66
Data Set 2	35, 37, 36
Data Set 3	35, 37, 36
Data Set 4	35, 37

These findings have demonstrated that the most significant features for pushing smokers to make a quit attempt are focused on the following aspects: knowledge about the harm of smoking, worry about health, confidence about quitting and addiction to smoking. Our findings are consistent with those that we have found before using other methods. Our methods have answered that smokers' motivations, knowledge about harm of smoking, beliefs about quitting and so on are key factors for making a quitting attempt. If we consider 64 to be sufficiently accurate, then the number of significant features increases.

3) Modified Global k-means Algorithm

Computational results to the Australian tobacco control data set can be summarized as follows:

- The modified global k-means algorithm allows one to find global or near global solutions to the clustering problems in the tobacco control data set.
- Results demonstrate that the modified global k-means algorithm detects correct number of clusters and the further reduction of the tolerance $\varepsilon>0$ do not lead to the increase of the number of clusters. This means that the modified global k-means algorithm is able to find stable cluster structure of the tobacco control data set.
- The clustering algorithm allows one to find stable clusters in the data set and these clusters do not change as the number of clusters increases. Moreover, we demonstrate that the cluster structure is not changing if one removes stable clusters one by one. This structure changes only when all stable clusters are removed from data set.

Our results show that the modified global k-means algorithm is efficient and robust for solving clustering

 ISBN: 978-1-61208-227-1

problems in the tobacco control data sets. Future work in this area includes classification of the set of clusters associated with each data set. Our methods aim to answer a key question: "How can we predict the response of smokers within the clusters to tobacco control policies?" Compared with the traditional statistical techniques, the new methods have potential to become a good theoretical and methodological framework for modelling and analysing complex tobacco control systems. The results of analysis of the given data set are most likely to develop new models for a new survey, more accurate than the previous one.

V. CONCLUSION

We evaluated three optimization-based data mining methods on two distinct medical data sets. The Cystic Fibrosis is a medical data set built around measurements of disease severity. The results show that all three methods worked equally well and may consequently be used for analysis of similar medical data sets.

The Tobacco Control data set is a massive survey for studying and evaluating the psychosocial and behavioural impact of diverse tobacco control policies to smokers from many countries. This kind of data sets tend to be noisy and the design of the survey may not be optimally suited to evaluate the accuracy of the outcome. The results demonstrate that the LLSF is very sensitive to the noise whereas other two optimization-based methods (both clustering and classification) perform well in the analysis of these types of data sets. More informative data sets enriched by health parameters will help to find the links from smoking to the risk of diseases such as dementia, stroke, lung cancer, vascular dementia, oxidative stress and inflammation. The reference to the research outcome could greatly impact the health choices of smokers.

We conclude by noting the usefulness of optimisation-based methods (both clustering and classification) in the analysis of distinct types of medical data sets.

REFERENCES

[1] A. M. Bagirov, A. M. Rubinov, and J. Yearwood, "A global optimisation approach to classification," Optimization and Engineering Journal, vol. 3, no. 2, 2002, pp. 129-155.

[2] A. M. Bagirov, A. M. Rubinov, and J. Yearwood, "A heuristic algorithm for feature selection based on optimisation techniques," in Heuristic and Optimization for Knowledge Discovery, C. Newton, H. Abbas and R. Sarker, Eds. Idea Group Publishing, 2002, pp. 13-26.

[3] A. M. Bagirov, A. M. Rubinov, N. V. Soukhoroukova, and J. Yearwood, "Supervised and unsupervised data classification via nonsmooth and global optimisation," TOP: Spanish Operations Research Journal, vol. 11, no. 1, 2003, pp. 1-93.

[4] A. M. Bagirov, "Modified global k-means algorithm for sum-of-squares clustering problems," Pattern Recognition, vol. 41, no. 10, 2008, pp. 3192-3199.

[5] A. M. Bagirov, A. M. Rubinov and J. Yearwood, "Using global optimisation to improve classification for medical diagnosis and prognosis," Topics in Health Information Management, vol. 22, no. 1, 2001, pp. 65-74.

[6] R. Borland, H. H. Yong, N. W. Geoffrey, G. T. Fong, D. Hammond, K. M. Cummings, W. Hosking, and A. McNeill, "How reaction to cigarette packet health warnings influence quitting: findings from the ITC four country survey," 2009, preprint.

[7] C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, 1998, pp. 121-167.

[8] W. P. Chang and D. M. Liou, "Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data," Journal of Telemedicine and Telecare, 2008.

[9] S. P. Conway and J. M. Littlewood, "Cystic fibrosis clinical scoring systems," in Cystic fibrosis - Current Topics, vol. 3, J. H. Widdicombe Ed. New York: John Wiley & Sons Ltd, 1996, pp. 339-358.

[10] R. M. David, "50 years of reporting on tobacco and health," British Medical Journal, 2000, vol. 320, no. 74.

[11] P. B. Davis, "Cystic fibrosis since 1938," American Journal of Respiratory and Critical Care Medicine, vol. 173, no. 5, 2006, pp. 475-482.

[12] Z. Dzalilov and A. M. Bagirov, "Cluster analysis of a tobacco control data set," International Journal of Lean Thinking, vol. 1, no. 2, pp. 40-45.

[13] Z. Dzalilov, J. Zhang, A. M. Bagirov, and M. A. Mammadov, "Application of optimisation–based data mining technique to tobacco control dataset," International Journal of Lean Thinking, vol. 1, no. 1, pp. 27-41.

[14] G. Hafen, C. Hurst, J. Yearwood, M. A. Mammadov, J. Smith, Z. Dzalilov, and P. Robinson, "A new clinical scoring system in cystic fibrosis: statistical tools for database analysis – a preliminary report," BMC Medical Informatics and Decision Making, 2008, 8: 44.

[15] D. Hammond, G. T. Fong, R. Borland, K. M. Cummings, A. McNeill, and P. Driezen, "Text and graphic warnings on cigarette packages: findings from the international tobacco control four country study," American Journal of Preventive Medicine, vol. 32, 2007, pp. 202-209.

[16] P. Hansen, E. Ngai, B. K. Cheung, and N. Mladenovic, "Analysis of global k-means, an incremental heuristic for minimum sum-of-squares clustering," Journal of Classification, vol. 22, no. 2, 2005, pp. 287-310.

[17] ITC International tobacco control policy evaluation survey. 2002-2010; http://www.itcproject.org, [retrieved: August, 2011]

[18] M. A. Mammadov, A. M. Rubinov, and J. Yearwood, "The study of drug-reaction relationships using global optimization techniques," Optimization Methods and Software, vol. 22, no. 1, 2007, pp. 99-126.

[19] M. A. Mammadov, A. M. Rubinov and J. Yearwood, "An optimization approach to identify the relationship between features and output of a multi-label classifier," Data Mining in Biomedicine, vol. 7, P. Pardalos, V. Boginski and A. Vazacopoulos, Eds. Series: Springer Optimization and its Applications, 2007, pp. 141-168.

[20] M. H. Schoeni, "Presentation and critical comparison of clinical scoring systems in patients with cystic fibrosis," Klin Paediatr, vol. 205, 1993, pp. 3-8.

[21] Y. Yang, "An evaluation of statistical approaches to text categorization," Information Retrieval, vol. 1, 1999, pp. 69-90.

[22] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, 1999, pp. 42-49.

[23] D. D. Walker and G. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, no. 2, 2005, pp. 113-127.

[24] J. Y. Zhang, D. Young, K. Coghill., S. Petrovic-Lazarevic, R. Borland, C. H. Yeh, and S. Bedingfield, "A new theoretical framework for modelling and analysing complex tobacco control systems," Proceedings of the Ninth Global Business and Technology Association's Annual International Conference, 2007, pp. 811-817.

 ISBN: 978-1-61208-227-1

Network-monitoring Method based on Self-learning and Multi-dimensional Analysis

Isao Shimokawa and Toshiaki Tarui
Network Systems Research Department, Hitachi, Ltd.
292 Yoshida, Totsuka, Yokohama, Kanagawa 244-0817, Japan
{isao.shimokawa.sd, toshiaki.tarui.my}@hitachi.com

***Abstract*—A novel network-monitoring system for detecting abnormal network conditions (such as hidden network congestion) is proposed. The proposed monitoring system is based on self-learning and multi-dimensional analysis. It analyzes multiple parameters such as consumed bandwidth, packet size, and arrival interval of network packets simultaneously. By executing high-quality network monitoring it thereby achieves multi-dimensional analysis by use of Mahalanobis distance. A prototype monitoring system was constructed and evaluated. The evaluation results indicate that the monitoring system can accurately detect a hidden change of network-traffic conditions and reduce the number of unnecessary alerts for monitoring excess bandwidth according to a set threshold.**

Keyword*-*Monitor; Network Fault; Mahalanobis distance.

I. INTRODUCTION

To reduce environment load from the viewpoint of energy efficiency, lowering the power consumption of cloud-service systems is attracting much interest. In a current cloud-service system, to reduce power consumption of the system, a management server triggers migration of a virtual machine (VM), aggregates virtual servers from one server to another, and switches off unused physical servers. A power-saving information and communication technology (ICT) platform is previously proposed. [1]

The ICT platform has to guarantee network bandwidth for cloud-service systems. If a system-management server executes VM migration without considering network-link capacity, volume of network traffic may surpass network-link capacity (because network flows connected to the VM are also moved from one network to another). As a result, unexpected network congestion may occur. Moreover, quality of service (QoS) such as bandwidth guarantee may not be maintained. It is therefore important to rapidly and accurately monitor the network and to execute VM migration according to the monitored network conditions.

On the contrary, if the management server switches off a preliminary server used for redundancy in order to lower the power consumption of the cloud-service system, a "cloud-service fault" may occur because the redundant server is switched off. Accordingly, to run a cloud-service system 24 hours a day all year and maintain QoS, network faults must be rapidly detected.

To address the above-mentioned issues, so-called "feedback control" [2] by monitoring a system is expected to provide stable and high-quality cloud services. For finding network faults, it is especially critical that network monitoring rapidly detects abnormal increases or decreases of network traffic. In the present work, to meet that need, a novel network-monitoring method for rapidly detecting abnormal changes of network-traffic conditions was devised.

The rest of this paper is organized as follows. Section II summarizes issues about network operation and management. Section III outlines a proposed method. Sections IV describe a prototype network-monitoring system. Section V evaluates the prototype network-monitoring system. Section VI concludes this paper.

II. ISSUES CONCERNING NETWORK OPERATIONS AND MANAGEMENT

Network-traffic conditions are typically monitored by a method for network operations and management such as simple network management protocol (SNMP). In addition, the monitored data is analyzed according to a one-dimensional threshold (such as consumed network bandwidth) without distinguishing different network flows. If network traffic fluctuates around a predefined alarm threshold such as network bandwidth, however, an alarm may occur frequently. In that case, the administrator of the network will receive many alarms even if no fault or problem has occurred in the network. In other words, applying a one-dimensional judgment such as bandwidth threshold for detecting abnormal network conditions may raise too many alarms. As a result, it is difficult to accurately monitor network conditions. Consequently, it is necessary to establish a monitoring system that detects network congestion or faults without generating too many alarms.

III. PROPOSED METHOD FOR MONITORING AND ANALYZING NETWORK TRAFFIC

To address the issue described in the previous section, a novel network monitoring system—based on self-learning and multi-dimensional analysis (SLMDA) [3]—is proposed here. The system monitors all network flows in real time from dimensions such as bandwidth, packet size, and packet interval. It is composed of an analyzing part and a monitoring part equipped with a node called "aggregated flow mining" (AFM) [4] implemented in each node. With regard to the analyzing part, a new evaluation scheme based on the multi-dimensional Mahalanobis distance [5] is applied.

 ISBN: 978-1-61208-227-1

A. *Aggregated flow mining (AFM)*

When an administrator of a network monitors network-traffic conditions, it is necessary to distinguish many network flows and analyze them in detail. For that purpose, AFM (which distinguishes many kinds of network flows and provides statistical information about those flows) is used. The network administrator finds anomalous flows or conditions (or both) by analyzing of statistical information provided by AFM. A flow is conventionally defined as a collection of packets with five tuples (source IP address, destination IP address, source port, destination port, and protocol). In regard to AFM, the concept of the flow is extended, and an "aggregated flow" is defined by an arbitrary combination of each tuple. For example, one aggregated flow is defined by only the destination IP address irrespective the other tuples. Statistical information (such as number of packets and bytes) about flows that have the same destination IP address is therefore produced as statistics about one aggregated flow. As described above, if the concept of aggregated flow is introduced, flows that travel between one host and multiple servers are regarded as one aggregated flow. It is therefore possible to analyze network traffic or flows.

B. *Algorithm for self-learning multi-dimensional analysis*

Hereafter, the node at which "integrated mining of flow" is performed is simply referred to as "the IMF". The proposed system for monitoring network-traffic flow is shown in Figure. 1 , where several users are connected to a data center. The IMF collects statistical information from multiple AFMs for analyzing network traffic and send alarms to a network-management server when it detects abnormal network conditions.

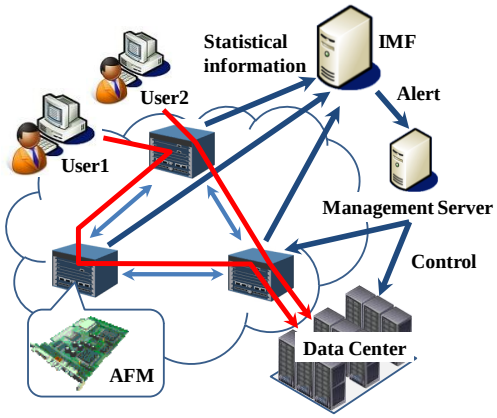


Figure. 1 Proposed monitoring system

In this system, network traffic is analyzed by using statistical information gathered from all AFMs in the network. Specifically, SLMDA using the Mahalanobis distance is used to analyze network conditions in detail. Although the system uses the same analysis parameters originally implemented in the AFM, the proposed SLMDA method is applicable to various analysis parameters. This method follows the procedure described below:

1. The standard distribution for each analysis parameter is defined.
2. A Statistical distribution for each analysis parameter is measured by an AFM in real time.
3. The Mahalanobis distance is calculated by comparing the distance between the defined standard distribution and the measured statistical distribution of each analysis parameter in real time, and the occurrence of abnormal network traffic conditions is judged.
4. The standard distribution for each analysis parameter is updated by using the measured statistical distribution (step 2) in real time.
5. If a rapid change of network traffic condition is detected, its cause is identified by analyzing the conditions in detail.
6. Return to step 2.

C. *One-dimensional judgment based on Mahalanobis distance*

To detect an abnormal network condition, it is necessary to analyze the changes of network traffic (such as bandwidth) in detail. Accordingly, a method for determining whether the network condition changes is proposed here. This one-dimensional judgment method based on the Mahalanobis distance is explained in Figure. 2 .

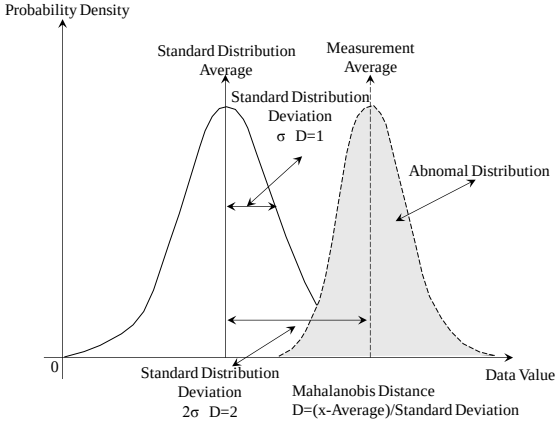


Figure. 2 One-dimensional judgment based on Mahalanobis distance

To analyse a rapid change of network-traffic distribution, it is necessary to define the standard distribution as a normal condition and compare that standard distribution and a statistical distribution measured by AFM. The comparison procedure is explained as follows. An initial value is set to define the standard distribution as a target for comparison with a measured distribution. As the parameters for comparison, average and standard deviation of the standard distribution are used. These parameters are set according to the administrator's experience or knowledge. To analyse a rapid change of network-traffic conditions, statistical information (such as data throughput) from the AFM of each router is measured in real time. Mahalanobis distance of the standard distribution is then calculated by using the

 ISBN: 978-1-61208-227-1

throughput distribution measured by the AFM in real time. The Mahalanobis distance is defined as

D=(x-average)/standard deviation [a.u.: arbitrary unit] (1)

If the Mahalanobis distance is very large, it is considered that an abnormal traffic condition exists. For example, if the calculated Mahalanobis distance is larger than 2 and the measured distribution follows a normal distribution, it is judged that the throughput distribution (namely, data rate/bandwidth of a traffic flow) is not significant according to a 5% significance level. As a result, it can be regarded as an unusual distribution that occurs at a probability of 5%.

D. Multi-dimensional judgment method based on Mahalanobias distance

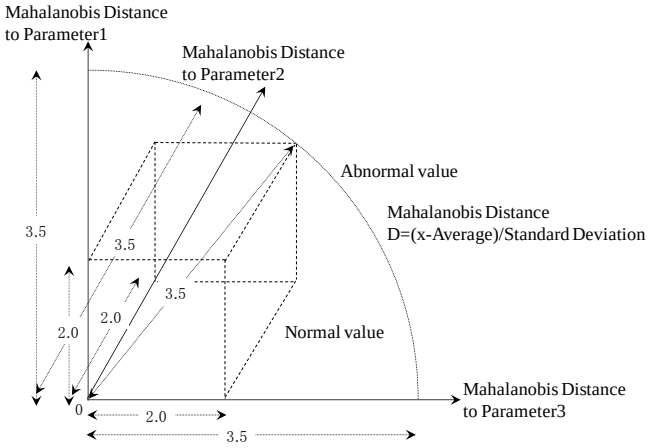


Figure 3. Three-dimensional Mahalanobis distance

If a one-dimensional judgement method is used to judge whether the measured network condition is normal or abnormal, an erroneous decision might occur frequently owing to the limited amount of information available for analyzing traffic flow. In the case that erroneous decisions occur, a management server may receive too many alerts, which might cause control errors. Accordingly, to improve the accuracy of the judgment, a multi-dimensional judgment method (as shown in Figure 3) is proposed here. This method involves several steps. First, each analysis parameter is assigned to each axis in the figure as a dimension for analyzing 3D Mahalanobis distance. On each axis, Mahalanobis distance is calculated. Multi-dimensional Mahalanobis distances are then calculated on the basis of multiple one-dimensional distances as follows.

Mahalanobis distance with three dimensions=

$$\mathrm{sqrt}(\alpha*x^2+\beta*y^2+\gamma*z^2) \quad (2a)$$

$$\alpha+\beta+\gamma=3 \quad (2b)$$

α,β,γ is not unique because reasons of network fault are not always same. So It is necessary to investigate α,β,γ from past data and system condition. In this report each of α,β,γ values is 1.

E. Updating standard distribution on the basis of feedback.

To analyze a measured network-traffic condition, a standard distribution as a target for comparison should be defined correctly. However, it is not easy to define a normal network traffic condition that changes day by day. A new self-learning method, by which a normal standard distribution is dynamically updated by using feedback data, is therefore proposed here. Basically, the network-traffic condition is monitored, and its changes are analyzed in real time. The standard distribution is then updated according to the change of the average data value of a measured distribution.

The example of a standard distribution updating method is shown in Figure 4. With the proposed method, average and standard deviation of the standard distribution are updated dynamically by calculating a moving average according to the following formula:

Moving average of average on standard distribution= (average on standard distribution + average on measured distribution)/2 (3a)

Moving average of deviation= (deviation on standard distribution + deviation on measured distribution)/2 (3b)

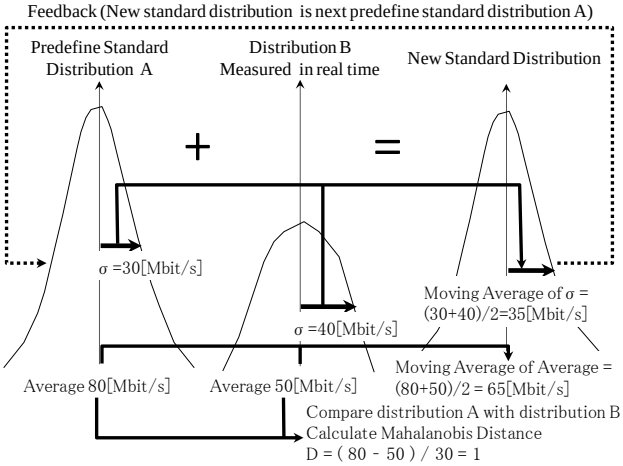


Figure 4. Example of standard distribution updating method.

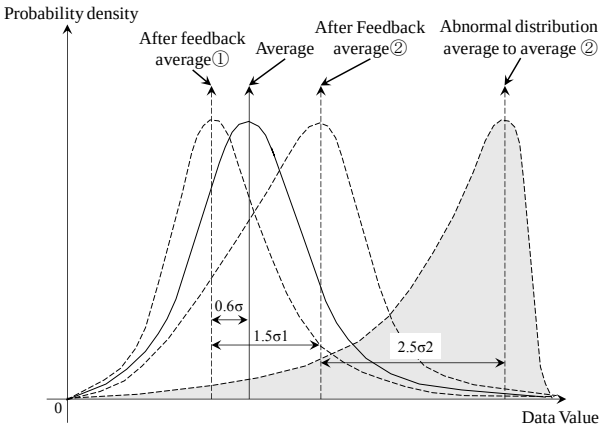


Figure 5. Abnormal distribution to standard distribution

F. *Detection of factors for changing network condition*

Detection of factors that change a network condition is explained in Figure 6. The criterion for detecting an abnormal condition is as follows:

average of standard distribution + two standard deviations < measured throughput of network flow (4)

When a rapid change of a network-traffic condition is detected, a flow that is further than two sigmas from the average value of the standard distribution is considered as a peculiar flow. Although two standard deviations is set as the threshold for detecting a peculiar flow, the threshold is set by the administrator of a network. If the threshold is two standard deviations and the measured data distribution follows a normal distribution, this condition is equivalent to a significance level of 5%, and it only occurs at a probability of 5%. In addition, flows that cause such a condition are judged as peculiar flows.

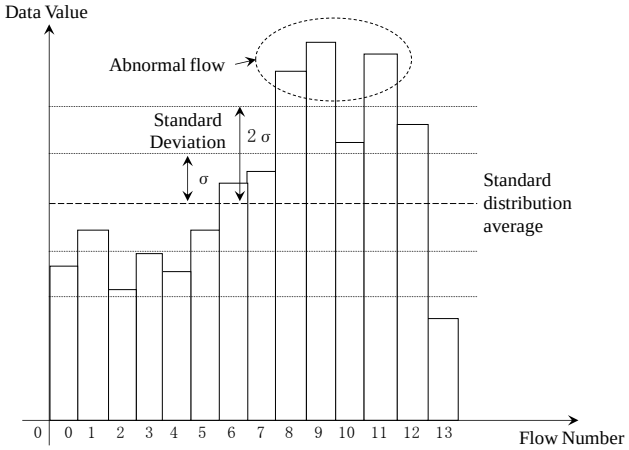


Figure 6. Detection of flows that cause abnormal condition

IV. IMPLEMENTATION OF PROTOTYPE TRAFFIC-MONITORING SYSTEM

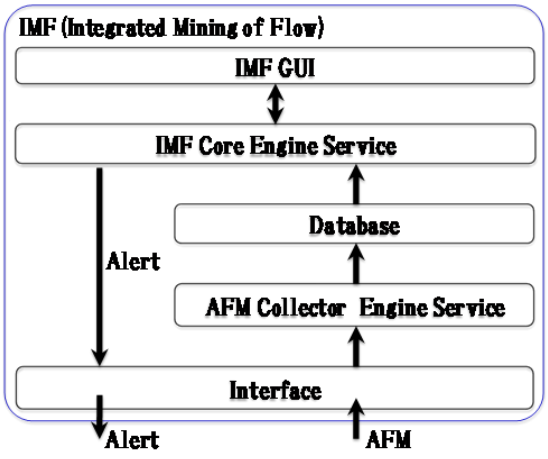


Figure 7. Block diagram of IMF

Figure 8. Example of IMF GUI

A block diagram of the IMF is shown in Figure 7. The IMF provides two functions (IMF core-engine service and AFM collector-engine service) for analyzing monitored data. The AFM collector engine service collects statistical information through a interface (such as Ethernet) and stores it in a database. The IMF core engine service then reads the statistical information from the database and analyzes it. If necessary, it sends an alert message to a network management server.

The GUI of the IMF is shown in Figure 8. The IMF analyzes statistical data from the AFM and shows real-time conditions on the GUI. The network administrator can check the flow that is presumed to be the factor causing an abnormal condition and the time of occurrence, when the change of condition is detected on GUI.

V. EVALUATION OF PROTOTYPE MONITORING SYSTEM

A. *Verification of proposed method*

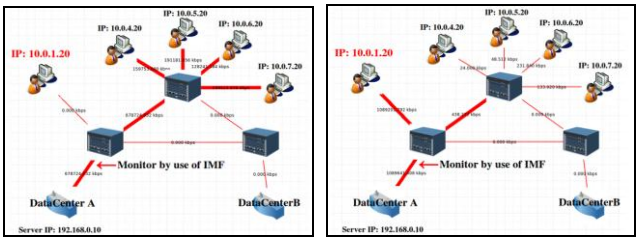


Figure 9. GUI screen of CORE

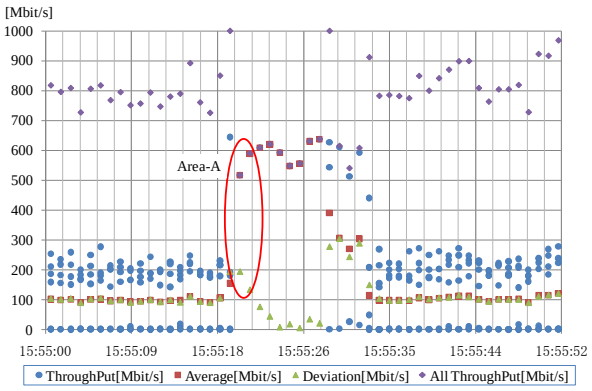


Figure 10. Throughput per network flow

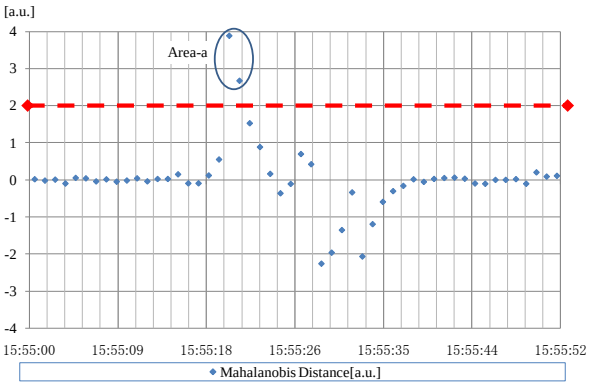


Figure 11. Mahalanobis distance with throughput per network flow

To evaluate whether detect a rapid change of network traffic by an SLMDA-based function, network congestion (as a change of network condition) is focused on, and the SLMDA-based function for detecting it is verified. Specifically, to produce network congestion, user datagram protocol (UDP) traffic is intentionally generated under the condition that only transmission control protocol (TCP) traffic is present. A rapid change of the network-traffic condition is then generated since transmission rate of TCP is rapidly decreased when TCP detects network congestion. The Common Open Research Emulator (CORE) [6] (which enables network emulation) was used to produce the above-described condition. The GUI screen of CORE, representing both conditions (before and after inserting UDP traffic), is shown in Figure 9.

An experiment to investigate traffic condition was performed by means of AFM. In the experiment, a rapid change of network condition is intentionally produced by inserting UDP traffic under the condition that only TCP traffic is flowing. As shown in Figure 9, multiple users connect to a data center by TCP communications. Network traffic via the router connected to data center A is monitored by the AFM. A user (IP:10.0.1.20) connects to data center A for a certain period by UDP communication. TCP controls the window size for data transmission when it detects congestion. On the other hand, UDP does not control transmission rate. Therefore, when UDP traffic is inserted into the TCP traffic, UDP traffic occupies most of the network link. As shown on the right of Figure 9, when UDP traffic is generated, the traffic (IP:10.0.1.20) occupies most of the network link. The throughput measured by AFM is shown in Figure 10. As shown in area A in the figure, a rapid change of network condition occurred at 15:55:20. Average throughput is not high until 15:55:20, since only multiple TCP communications share the network link. However, the average throughput increases rapidly, since a UDP communication occupies the network link from 15:55:20. As shown in Figure 11, a significant change of network-traffic condition is detected at 15:55:20, since the Mahalanobis distance is over 2 at 15:55:20. It is thus possible to detect abnormal traffic conditions that the network administrator cannot recognize by a conventional method. Moreover, the IMF could detect the UDP flow (IP:10.0.1.20) as a potential factor for causing a significant change of network condition. The network administrator can therefore easily find the factors causing significant changes in network condition and take appropriate measures for handling them by using the following information produced by IMF.

Detecting Flow by IMF: [Time] 15:55:20 [Source IP] 10.0.1.20 [Source Port] 56165 [Destination IP] 192.168.0.10 [Destination Port] 5001[Average ThroughPut][Mbit/s] 517 [AllPacketSize][bytes] 80136000 [Protocol Num] 17(UDP)

B. *Experiment on intranet*

The purpose of this experiment (see concept shown in Figure 12) is to verify whether the proposed method can detect a significant change of network-traffic condition on a real intranet. In the experiment, real intranet traffic was measured for one day by AFM. The bandwidth of the link used for the experiment was 100 Mbit/s. Four parameters (throughput, average packet size, link throughput, and data volume) were used as parameters in this evaluation. The results of the evaluation from AM0:00 to AM9:00 are shown in Figure 13 to Figure 21.

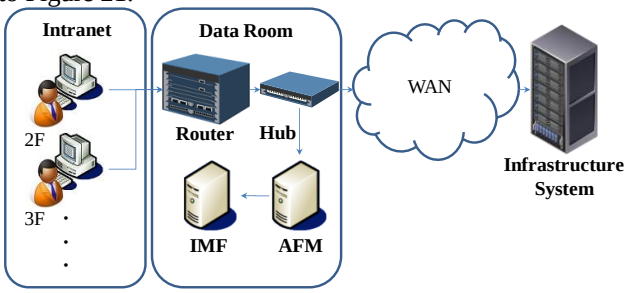


Figure 12.Concept of experiment on intranet

As shown in Figure 13, many flows have high throughput. It is therefore difficult to pinpoint in the figure whether a significant change of network-traffic condition was generated or not. On the contrary, in Figure 14, with the proposed method, significant changes of network-traffic condition are judged to occur two or more times. In the time zone when the Mahalanobis distance is over 2, namely, from area A to area C in the figure, traffic rapidly decreases and stays low for a short time. After that time, the traffic rapidly increases again. It is concluded that the proposed scheme could detect such significant changes of network-traffic condition.

The average packet size is shown in Figure 15, and calculated Mahalanobis distance that corresponds to average packet size is shown in Figure 16. As shown in Figure 15, average packet size does significantly not deviate from an average of 500 bytes. However, in Figure 16, two points (shown in area D and area E in the figure) are detected as significant changes of average packet size. The significant change of network-traffic condition is therefore detected according to average packet size. The times at which packet-size changes are detected are equivalent to the times at which throughput are detected in Figure 14. (Area-A,Area-B)

 ISBN: 978-1-61208-227-1

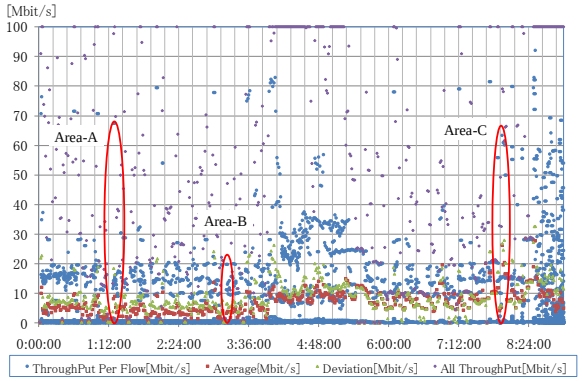


Figure 13. Throughput per flow

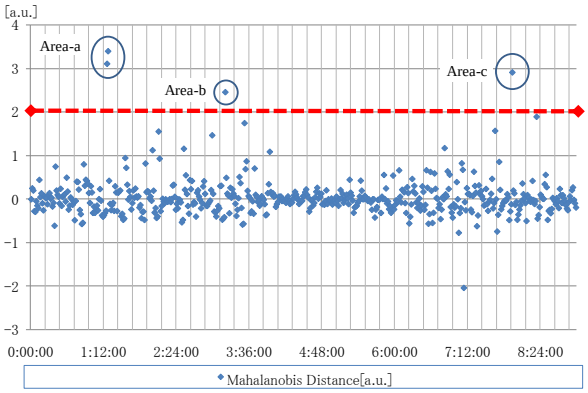


Figure 14. Mahalanobis distance with throughput per flow

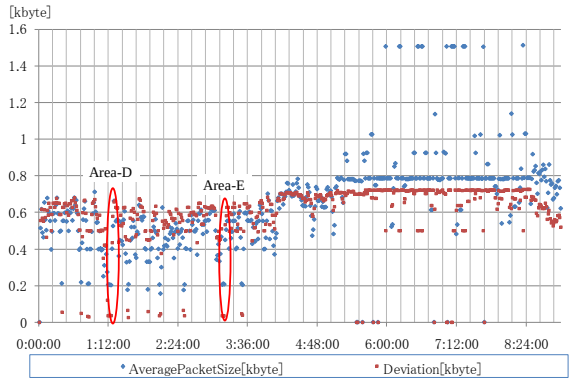


Figure 15. Packet size

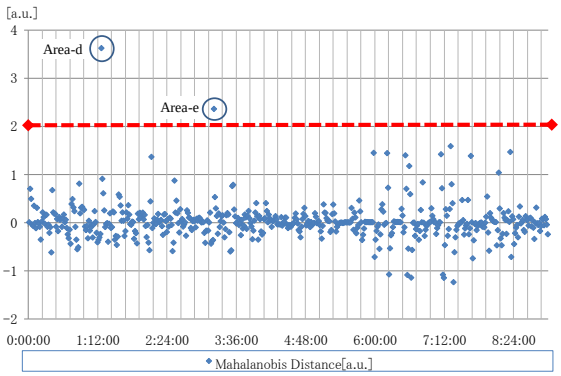


Figure 16. Mahalanobis distance with packet size

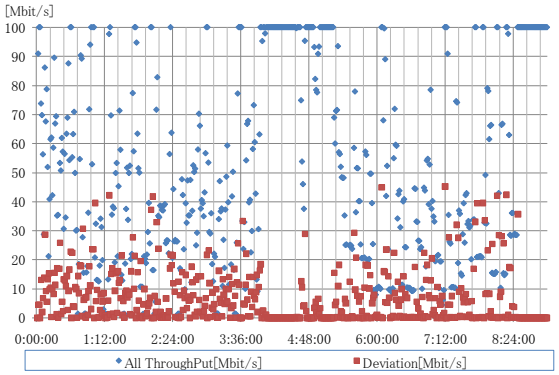


Figure 17. Link throughput

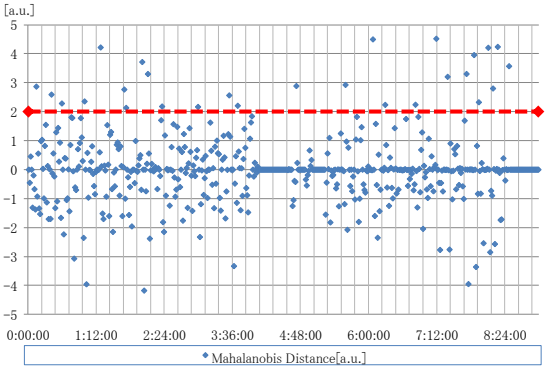


Figure 18. Mahalanobis distance with link throughput

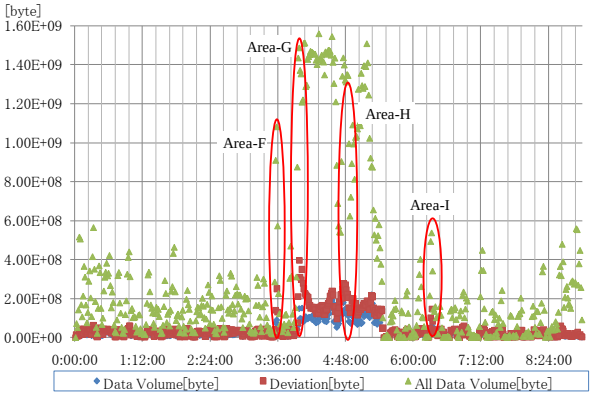


Figure 19. Data volume

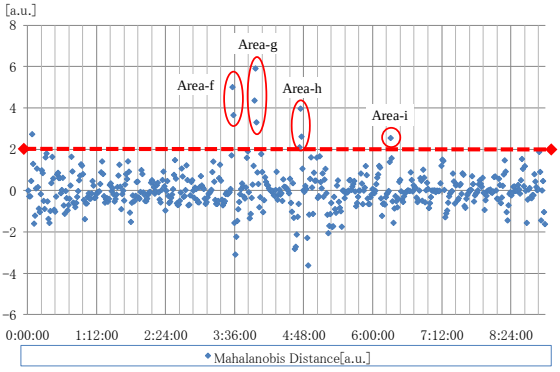


Figure 20. Mahalanobis distance with data volume

 ISBN: 978-1-61208-227-1

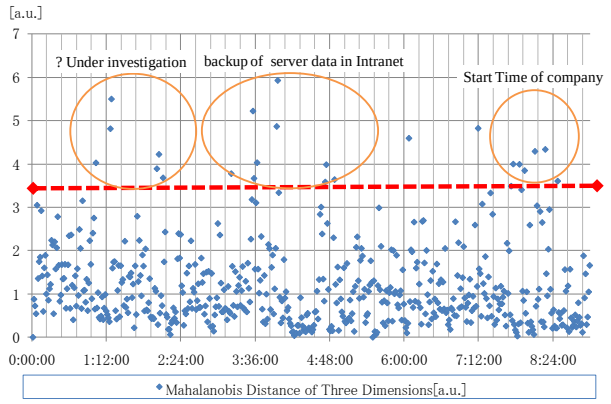


Figure 21. Mahalanobis distance with three dimensions

The measured link throughput is shown in Figure 17, and the calculated Mahalanobis distance with the link throughput is shown in Figure 18. It is difficult to pinpoint from Figure 17 the points at which a significant change of network-traffic condition was generated. On the other hand, significant changes are detected two or more times in Figure 18. However, there is no commonality between the detected changes in Figure 17 and the previously detected changes in Figures 14 and 16.

The measured data volume is shown in Figure 19. As shown from area F to area I in the figure, data volume rises rapidly in a certain time zone. Moreover, significant changes of network-traffic condition (i.e., data volume) were detected two or more times. The times of the change in data volume shown in Figure 19 is almost equivalent to those of the change in Mahalanobis distance shown in Figure 20. It is concluded that server data on the intranet should be backed-up at these times.

The calculated Mahalanobis distance converted into three dimensions is shown in Figure 21. In the calculation, three-dimensional data are selected from four types of monitored data (Throughput per flow, Packet size, Link throughput, data volume). This result bases on formula (2a) and (2b). Each of α,β,γ value used in this experiment is 1. α,β,γ can't be decided from theory because reasons of network fault are not always same and abnormal condition depends on system condition. So α,β,γ must be decided from past data and system condition, administrator experience. Currently how to decide α,β,γ is an issue in the future. Self-learning method could be used to decide those parameters.

The following focuses on from AM2:24 to AM4:48. As a result of setting the threshold to 80 Mbit/s and analyzing intranet traffic, the number of alarms exceeding the threshold value was detected 54 times in Figure 13. On the other hand, as shown in Figure 14, by proposed method, the number of alarms detected the significant changes of network- traffic condition is four times. Consequently, it is possible to reduce the number of alerts by 92%. Moreover, the one-dimensional judgment method was used to the four types of monitored data and the number of alarms exceeding the threshold value with Mahalanobis distance greater than the threshold of two is 14 times as sum of alarms with the four types of monitored data. When my proposed method was judged by the three-dimensional Mahalanobis distance, it was seven times, and the alarm decreases by half compared to one-dimensional judgment method. As a whole, the proposed method can reduce the number of alarms by 96% compared to one-dimensional judgment method.

VI. CONCLUDING REMARKS

A new network-monitoring system based on self-learning and multi-dimensional analysis (SLMDA) using the Mahalanobis distance was proposed. This system detects a significant change of network traffic. It uses Mahalanobis distance converted to multiple dimensions between standard distribution and measured distribution in real time. If the distance is larger than two standard deviations of standard distribution, the system judges that the monitored condition is abnormal. The system can therefore detect a rapid change of network traffic condition. A prototype network-monitoring system was developed and evaluated, When user datagram protocol (UDP) traffic is intentionally generated under the condition that only transmission control protocol (TCP) traffic is present even if the whole consumed bandwidth is not changed. As a result, the system could detect a significant change of network traffic. In addition to using real traffic, the system can reduce the number of unnecessary alerts by about 96% when throughput is fluctuating at normal rate of bandwidth consumption near a predefined threshold. As for future work, the proposed monitoring system will be extended to large-scale networks, and its performance will be evaluated.

ACKNOWLEDGMENT

Part of this research was supported by MIC (Ministry of Internal Affairs and Communications) as part of the "Research and development on network control technologies supporting cloud services for high availability and power saving" project

REFERENCES

[1] T. Suzuki et al., "Power-Saving ICT Platform That Guarantees Network Bandwidth for Cloud-Service Systems," World Telecommunications Congress (WTC), 2012

[2] Xiao Wei et al., "A Network Monitor System Model with Performance Feedback Function" E-Business and Information System Security, 2009. EBISS '09. International Conference on Digital Object Identifier: 10.1109/EBISS.2009.5137879 Publication Year: 2009 , Page(s): 1 - 5

[3] I. Shimokawa et al., "Examination of network fault detection method by use of AFM," IEICE CPSY, computer system 110(473), 31-38, 2011-03-11.

[4] Y. Shomura et al., "Analyzing the Number of Varieties in Frequently Found Flow," IEICE Trans. Commun., vol. E91-B, no. 6, pp. 1896-1905, Jun. 2008.

[5] Mahalanobis, Prasanta Chandra (1936), "On the generalised distance in statistics," Proceedings of the National Institute of Sciences of India **2** (1): 49–55.

[6] http://code.google.com/p/coreemu

 ISBN: 978-1-61208-227-1

Structure Learning of Bayesian Networks Using a New Unrestricted Dependency Algorithm

Sona Taheri
Centre for Informatics and Applied Optimization
School of Science, Information Technology and Engineering
University of Ballarat, VIC 3353, Australia
Email: sonataheri@students.ballarat.edu.au

Musa Mammadov
University of Ballarat, VIC 3353, Australia
Email: m.mammadov@ballarat.edu.au
National ICT Australia, VRL, VIC 3010, Australia
Email: musa.mammadov@nicta.com.au

***Abstract*—Bayesian Networks have deserved extensive attentions in data mining due to their efficiencies, and reasonable predictive accuracy. A Bayesian Network is a directed acyclic graph in which each node represents a variable and each arc a probabilistic dependency between two variables. Constructing a Bayesian Network from data is the learning process that is divided in two steps: learning structure and learning parameter. In many domains, the structure is not known a priori and must be inferred from data. This paper presents an iterative unrestricted dependency algorithm for learning structure of Bayesian Networks for binary classification problems. Numerical experiments are conducted on several real world data sets, where continuous features are discretized by applying two different methods. The performance of the proposed algorithm is compared with the Naive Bayes, the Tree Augmented Naive Bayes, and the $k-$Dependency Bayesian Networks. The results obtained demonstrate that the proposed algorithm performs efficiently and reliably in practice.**

Keywords-Data Mining; Bayesian Networks; Naive Bayes; Tree Augmented Naive Bayes; $k-$Dependency Bayesian Networks; Topological Traversal Algorithm.

I. INTRODUCTION

Data Mining is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. The whole process of data mining consists of several steps. Firstly, the problem domain is analyzed to determine the objectives. Secondly, data is collected and an initial exploration is conducted to understand and verify the quality of the data. Thirdly, data preparation is made to extract relevant data sets from the database. A suitable data mining algorithm is then employed on the prepared data to discover knowledge represented in different representations such as decision trees, neural networks, support vector machine and Bayesian Networks. Finally, the result of data mining is interpreted and evaluated. If the discovered knowledge is not satisfactory, these steps will be iterated. The discovered knowledge is then applied in decision making. Recently, there is an increasing interest in discovering knowledge represented in Bayesian Networks [13], [14], [17], [15], [19] and [28]. Bayesian networks (BNs), introduced by Pearl [21], can encode dependencies among all variables; therefore, they readily handle situations where some data entries are missing. BNs are also used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Moreover, since BNs in conjunction with Bayesian statistical techniques have both causal and probabilistic semantics, they are an ideal representation for combining prior knowledge and data [10]. In addition, BNs in conjunction with Bayesian statistical methods offer an efficient and principal approach for avoiding the over fitting of data [20]. BNs have been applied widely for data mining, causal modeling and reliability analysis [29].

This paper presents a novel unrestricted dependency algorithm to learn knowledge represented in BNs from data. A BN is a graphical representation of probability distributions over a set of variables that are used for building a structure of the problem domain. The BN defines a network structure and a set of parameters, class probabilities and conditional probabilities. Once the network structure is constructed, the probabilistic inferences are readily calculated, and can be performed to predict the outcome of some variables based on the observations of others.

The main task of learning BNs from data is finding directed arcs between variables, or, in other words, the structure discovery, which is the more challenging, and thus, more interesting phase. Two rather distinct approaches have been used widely to structure discovery in BNs: the constraint-based approach [22], [27] and the score-based approach [1], [5], [26]. In the the constraint-based approach, structure learning cares about whether one arc in the graph should be existed or not. This approach relies on the conditional independence test to determine the importance of arcs [4]. In the score-based approach, several candidate graph structures are known, and we need choosing the best one out. In order to avoid over fitting, investigators often use model selection methods, such as Bayesian scoring function [5] and entropy-based method [12]. Several exact algorithms based on dynamic programming have recently been developed to learn an optimal BN [16], [24], [25] and [31]. The main idea in these algorithms is to solve small subproblems first and

 ISBN: 978-1-61208-227-1

use the results to find solutions to larger problems until a global learning problem is solved. However, they might be inefficient due to their need to fully evaluate an exponential solution space.

It has been proved that learning an optimal BN is NP-hard [11]. In order to avoid the intractable complexity for learning BNs, the Naive Bayes [18] has been used. The Naive Bayes (NB) is the simplest among BNs. In the NB, features are conditionally independent given the class. It has shown to be very efficient on a variety of data mining problems. However, the strong assumption that all features are conditionally independent given the class is often violated on many real world applications. In order to relax this assumption of the NB while at the same time retaining its simplicity and efficiency, researchers have proposed many effective methods [7], [23] and [28]. Sahami [23] proposed the $k-$dependence BNs to construct the feature dependence with a given number, value of k. In this algorithm, each feature could have a maximum of k feature variables as parents, and these parents are obtained by using mutual information. The value of k in this algorithm is initially chosen before applying it, $k = 0, 1, 2, ...$ Friedman et al. [7] introduced the Tree Augment Naive Bayes (TAN) based on the tree structure. It approximates the interactions between features by using a tree structure imposed on the NB structure. In the TAN, each feature has the class and at most one other feature as parents.

Although the mentioned methods were shown to be efficient, the features in these methods depend on the class and a priori given number of features; $k = 0$ dependence for the NB, $k = 1$ dependence for the TAN, and an initially chosen k for the k-dependence BNs. In fact, by setting k, i.e., the maximum number of parent nodes that any feature may have, we can construct the structure of BNs. Since k is the same for all nodes, it is not possible to model cases where some nodes have a large number of dependencies, whereas others just have a few. In this paper, we propose a new algorithm to identify the limitations of each of these methods while also capturing much of the computational efficiency of the NB. In the proposed algorithm, the number k is defined by the algorithm internally, and it is an unrestricted dependency algorithm.

The rest of the paper is organized as follows. In the next section, we provide a brief description of BNs. In Section III, we introduce a new algorithm for structure learning of BNs from binary classification data. Section IV presents a brief review of the Topological Traversal algorithm. The results of numerical experiments are given in Section V. Section VI concludes the paper.

II. Representation of Bayesian Networks

A BN consists of a directed acyclic graph connecting each variables into a network structure and a collection of conditional probability tables, where each variable in the graph is denoted by a conditional probability distribution given its parent variables. The nodes in the graph correspond to the variables in the domain, and the arcs (edges) between nodes represent causal relationships among the corresponding variables. The direction of the arc indicates the direction of causality. When two nodes are joined by an arc, the causal node is called the parent of the other node, and another one is called the child. How one node influences another is defined by conditional probabilities for each node given its parents [21]. Suppose a set of variables $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, where X_i denotes both the variable and its corresponding node. Let $Pa(X_i)$ denotes a set of parents of the node X_i in $\mathbf{X}$. When there is an edge from X_i to X_j, then X_j is called the child variable for a parent variable X_i. A conditional dependency connects a child variable with a set of parent variables. The lack of possible edges in the structure encodes conditional independencies.

In particular, given a structure, the joint probability distribution for $\mathbf{X}$ is given by

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|Pa(X_i)), \tag{1}$$

here, $P(X_i|Pa(X_i))$ is the conditional probability of X_i given its parents $Pa(X_i)$, where

$$P(X_i|Pa(X_i)) = \frac{P(X_i, Pa(X_i))}{P(Pa(X_i))} = \frac{n_{X_i,Pa(X_i)}}{n_{Pa(X_i)}},$$

where $n_{Pa(X_i)}$ denotes the number of items in the set $Pa(X_i)$, and $n_{X_i,Pa(X_i)}$ is the number of items in $X_i \cap Pa(X_i)$.

However, accurate estimation of $P(X_i|Pa(X_i))$ is non trivial. Finding such an estimation requires searching the space of all possible network structures for one that best describes the data. Traditionally, this is done by employing some search mechanism along with an information criterion to measure goodness and differentiate between candidate structures met while traversing the search space. The idea would be to try and maximize this information measure or score by moving from one structure to another. The associated structure is then chosen to represent and explain the data. Finding an optimal structure for a given set of training data is a computationally intractable problem. Structure learning algorithms determine for every possible edge in the network whether to include the edge in the final network and which direction to orient the edge. The number of possible graph structures grows exponentially as every possible subset of edges could represent the final model. Due to this exponential growth in graph structure, learning an optimal BNs has been proven to be NP-hard [11].

During the last decades a good number of algorithms whose aim is to induce the structure of the BN that better represents the conditional dependence and independence

 ISBN: 978-1-61208-227-1

relationships underlying have been developed [4], [5], [7], [12], [16], [24] and [25]. In our opinion, the main reason for continuing the research in the structure learning problem is that mendelizing the expert knowledge has become an expensive, unreliable and time consuming job. We introduce a new algorithm for structure learning of BNs in the following section.

III. The Proposed Algorithm for Bayesian Networks

In this section, we propose a new algorithm to learn the structure of BNs for binary classification problems. Since the learning process in BNs is based on the correlations of children and parent nodes, we propose a combinatorial optimization model to find the dependencies between features. However, some features could be independent which is considered by intruding a threshold K. Let us consider an optimization model (2):

$$\begin{array}{c} \max \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (K_{ij} - K) w_{ij}, \\ \text{subject to} \quad w_{ij} + w_{ji} \leq 1, \end{array} \tag{2}$$

where $1 \leq i, j \leq n,\ i < j$ and $w_{ij} \in \{0, 1\}$. w_{ij} is the association weight (to be found), given by

$$w_{ij} = \begin{cases} 1 & \text{if feature } X_i \text{ is the parent of feature } X_j, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and for $1 \leq i, j \leq n,\ \ i \neq j$,

$$K_{ij} = \sum_{q_2=1}^{|X_j|} \sum_{q_1=1}^{|X_i|} \max\{P(X_{q_2 j} | C_1, X_{q_1 i}), P(X_{q_2 j} | C_2, X_{q_1 i})\}. \tag{4}$$

Here, $|X_j|$ and $|X_i|$ are the number of values of features X_j and X_i, respectively, and X_{ql} shows the qth value of the feature $X_l,\ 1 \leq l \leq n$. We assume binary classification; $C_1 = 1$ and $C_2 = -1$ are class labels. K is a threshold such that $K \geq 0$.

From the formula (2), $w_{ij} = 1$ if $K_{ij} > K_{ji}$ and $K_{ij} > K$, and therefore $w_{ji} = 0$ due to the constraint $w_{ij} + w_{ji} \leq 1$. It is clear that $w_{ii} = 0,\ 1 \leq i \leq n$. Thus problem (2) can be solved easily. Let us denote the solution of the problem (2) by $W(K) = [w_{ij}(K)]_{n \times n}$, where

$$w_{ij}(K) = \begin{cases} 1 & \text{if } K_{ij} > K_{ji} \text{ and } K_{ij} > K, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

and the set of arcs presented by

$$A(W) = \{(i, j) :\ if\ w_{ij} = 1,\ 1 \leq i, j \leq n,\ i \neq j\}, \tag{6}$$

(i, j) shows the arc from X_i to X_j. If we have set of arcs $A(W)$, then we have the corresponding matrix W that satisfies (6). It is clear that $A(W) \subset I$, where $I = \{(i, j),\ 1 \leq i, j \leq n\}$ is the set of all possible couples (i, j).

The best value for K will be found based on the maximum training accuracy for different values of $w_{ij}(K)$, where $0 \leq K \leq K^{max}$, and

$$K^{max} = max\{K_{ij},\ 1 \leq i, j \leq n,\ i \neq j\}. \tag{7}$$

More precisely, we find the values of $w_{ij}(K_r)$ for different $K_r = K^{max} - \varepsilon r,\ r = 0, 1, ...$ until $K_r < 0$, and we set $W(K_r) = [w_{ij}(K_r)]_{n \times n}$. With the matrix $W(K_r)$, the set of arcs $A(W(K_r))$ and, therefore, a network will be learnt. Based on the obtained network, the conditional probabilities will be found:

$$P(C|\mathbf{X}) \equiv \prod_{i=1}^{n} P(X_i | C, Pa(X_i)) P(C), \tag{8}$$

where $Pa(X_i)$ denotes the set of parents of the variable X_i to be found with $W(K_r)$. Now, based on these conditional probabilities, we calculate:

$$C(\mathbf{X}) = \begin{cases} 1 & \text{if } P(C_1 = 1|\mathbf{X}) > P(C_2 = -1|\mathbf{X}), \\ -1 & \text{otherwise,} \end{cases}$$

and then the maximum training accuracy will be found using the following formula:

$$accuracy(A(W(K_r))) = \frac{100}{ntr} \sum_{i=1}^{ntr} \delta(C(\mathbf{X}_i), C_i), r = 0, 1, ... \tag{9}$$

where

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{otherwise.} \end{cases}$$

We will choose that value of r corresponding to the highest training accuracy. Here, ntr stands for the number of instances in the training set.

Since BNs are directed acyclic graphs, we should not have any cycle in the structure obtained by $A(W(K_r))$. Therefore, the maximum training accuracy subject to no cycles will give the best value of K_r, denoted by K^*, and consequently, the best structure $A(W(K^*))$. Here, we apply the topological traversal algorithm to test if the corresponding graph to the obtained network is acyclic.

According to explanations above, the proposed algorithm constructs unrestricted dependencies between features based

 ISBN: 978-1-61208-227-1

on the structure of the NB. The proposed algorithm eliminates the strong assumptions of independencies between features in the NB, yet at the same time maintains its robustness. It is clear that $r = 0$ in the proposed algorithm gives the structure of the NB. In our algorithm, some features could have a large number of dependencies, whereas others just have a few. The number of these dependencies will be defined by the algorithm internally. The steps of our algorithm is presented in the following:

Step 1. Compute $\{K_{ij},\ 1 \leq i, j \leq n,\ i \neq j\}$ using (4).

Step 2. Determine K^{max} using (7). Set $r = 0$, and $p_0 = 0$.

Step 3. **while** $K^{max} - \varepsilon r \geq 0$ **do**

3.1. Calculate $K_r = K^{max} - \varepsilon r$.

3.2. Compute $w_{ij}(K_r),\ 1 \leq i, j \leq n,\ (i \neq j)$ using (5), and let $W(K_r) = [w(K_r)_{ij}]_{n \times n}$.

3.3. Find dependencies between features by a set of arcs $A(W(K_r))$ using (6).

3.4. Apply the topological traversal algorithm to test the network obtained by $A(W(K_r))$ for possible cycles. If any cycle is founds, then go to Step 4.

3.5. Compute the training accuracy, $p = accuracy(A(W(K_r))$, using (9). If $p > p_0$ then set $p_0 = p$, $K^* = K_r$, $r = r + 1$.

end

Step 4. Construct the optimal structure based on the basic structure of the NB and applying the set of arcs $A(W(K^*))$ between features.

Step 5. Compute the conditional probability tables inferred by the new structure.

Algorithm 1: Unrestricted Dependency BNs Algorithm

In this paper, we limit ourselves to binary classification, though a brief discussion on multiple class classification is warranted. The most straightforward approach in these classification problems is finding maximum of m conditional probabilities in the formula (4), where m is the number of classes. Moreover, the one-versus-all classification paradigm will be used to find either in the training accuracy, (9), or the test accuracy in the experiments.

IV. Topological Traversal Algorithm

The topological traversal algorithm [8] is applied for testing a directed graph if there exists any cycle. The degree of a node in a graph is the number of connections or edges the node has with other nodes. If a graph is directed, meaning that edges point in one direction from one node to another node. Then nodes have two different degrees, the in-degree, which is the number of incoming edges to this node, and the out-degree, which is the number of outgoing edges from this edge.

The topological traversal algorithm begins by computing the in-degrees of the nodes. At each step of the traversal, a node with in-degree of zero is visited. After a node is visited, the node and all the edges emanating from that node are removed from the graph, reducing the in-degree of adjacent nodes. This is done until the graph is empty, or no node without incoming edges exists. The presence of the cycle prevents the topological order traversal from completing. Therefore, the simple way to test whether a directed graph is cyclic is to attempt a topological traversal of its nodes. If all nodes are not visited, the graph must be cyclic.

V. Experiments

We have employed 12 well-known binary classification data sets. A brief description of the data sets is given in Table I. The detailed description of the data sets used in this experiments are downloadable in the UCI repository of machine learning databases [2] and the tools page of the LIBSVM [3]. The reason that we have chosen these data sets is: they are the most frequently binary classification data sets considered in the literature.

All continue features in data sets are discretized using two different methods. In the first one, we apply a mean value of each feature to discretize values to binary, $\{0, 1\}$. In the second one, we use the discretization algorithm using sub-optimal agglomerative clustering (SOAC) [30] to get more than two values for discretized features.

We conduct an empirical comparison for the Naive Bayes (NB), the Tree Augmented Naive Bayes (TAN), the $k-$Dependency Bayesian Networks ($k-$DBN), and the proposed algorithm (UDBN) in terms of test set accuracy. We have compared our algorithm with the mentioned algorithms because the basic structure of all, the TAN, the $k-$DBN and the UDBN, is based on the the structure of the NB. In all the cases we have used $10-$fold cross validation. We report the averaged accuracy over the ten test folds.

Table II presents the averaged test set accuracy obtained by the NB, the TAN, the $k-$DBN and the UDBN on 12 data sets, where continuous features are discretized using mean values for discretization. The results presented in this table demonstrate that the accuracy of the proposed algorithm (UDBN) is much better than that of the NB, and the TAN in all data sets. The UDBN also works better than the $k-$DBN in most of data sets. In 10 cases out of 12, the UDBN has higher accuracy than the $k-$DBN. The accuracy of this method almost ties with the $k-$DBN in data sets Phoneme CR and German.numer.

The averaged test set accuracy obtained by the NB, the TAN, the $k-$DBN and the UDBN on 12 data sets using discretization algorithm SOAC summarized in Table III. The

 ISBN: 978-1-61208-227-1

results from this table show that the accuracy obtained by the proposed algorithm in all data sets are higher than those obtained by the NB, the TAN, and the $k-$DBN.

According to the results explained, the proposed algorithm, UDBN, works well. It yields good classification compared to the NB, the TAN and the $k-$DBN. In addition, our algorithm is more general than the $k-$DBN. In the $k-$DBN, the number k is a priori chosen. In fact, by setting k, i.e., the maximum number of parent nodes that any feature may have, the structure of BNs could be constructed. Since k is the same for all nodes, it is not possible to model cases where some nodes have a large number of dependencies, whereas others just have a few. However, in the proposed algorithm, the number k is defined by the algorithm internally, and it is an unrestricted dependency algorithm. It might be various for different data sets, and even for each fold in the calculations. The computational times are not presented in Tables II and III. It is clear that the proposed algorithm needs more computational time than the others, since for example, the NB appears as a special case of UDBN when $r = 0$.

Table I
A BRIEF DESCRIPTION OF DATA SETS

Data sets	# Instances	# Features
Breast Cancer	699	10
Congressional Voting Records	435	16
Credit Approval	690	15
Diabetes	768	8
Haberman's Survival	306	3
Ionosphere	351	34
Phoneme CR	5404	5
Spambase	4601	57
Fourclass	862	2
German.numer	1000	24
Svmguide1	7089	4
Svmguide3	1284	21

Table II
TEST SET ACCURACY AVERAGED OVER $10-$FOLD CROSS VALIDATION FOR DATA SETS USING MEAN VALUES FOR DISCRETIZATION. NB STANDS FOR NAIVE BAYES, TAN FOR TREE AUGMENTED NAIVE BAYES, $k-$DBN FOR $k-$DEPENDENCY BAYESIAN NETWORKS, $k = 2$, AND UDBN FOR THE PROPOSED ALGORITHM

Data sets	NB	TAN	$k-$DBN	UDBN
Breast Cancer	97.18	96.52	97.31	97.66
Congressional Voting Records	90.11	93.21	94.62	95.48
Credit Approval	86.10	84.78	86.87	87.46
Diabetes	74.56	75.14	75.03	75.98
Haberman's Survival	75.09	74.41	76.43	77.86
Ionosphere	88.62	89.77	88.35	89.98
Phoneme CR	77.56	78.31	80.58	80.16
Spambase	90.41	89.78	89.27	92.37
Fourclass	77.46	77.61	77.94	79.06
German.numer	74.50	73.13	76.35	76.27
Svmguide1	92.39	91.61	92.98	94.17
Svmguide3	81.23	82.47	83.64	85.41

Table III
TEST SET ACCURACY AVERAGED OVER $10-$FOLD CROSS VALIDATION FOR DATA SETS USING DISCRETIZATION ALGORITHM SOAC. NB STANDS FOR NAIVE BAYES, TAN FOR TREE AUGMENTED NAIVE BAYES, $k-$DBN FOR $k-$DEPENDENCY BAYESIAN NETWORKS, $k = 2$, AND UDBN FOR THE PROPOSED ALGORITHM

Data Sets	NB	TAN	$k-$DBN	UDBN
Breast Cancer	96.12	95.60	96.76	97.65
Congressional Voting Records	90.11	91.42	92.61	94.16
Credit Approval	85.85	84.98	86.53	87.17
Diabetes	75.78	75.90	75.82	76.22
Haberman's Survival	74.66	73.78	75.64	77.31
Ionosphere	85.92	86.18	85.94	88.62
Phoneme CR	77.01	78.53	80.41	81.01
Spambase	89.30	89.04	90.69	92.54
Fourclass	78.58	79.52	78.97	79.96
German.Numer	74.61	74.01	75.31	76.15
Svmguide1	95.61	94.91	96.32	97.54
Svmguide3	77.25	79.99	80.75	82.92

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new algorithm for learning of the structure in Bayesian Networks. An important property of this algorithm is adding some numbers of arcs between features that captures unrestricted dependency among them. The number of arcs has been defined by the proposed algorithm internally. We have carried out a number of experiments on some binary classification data sets from the UCI machine learning repository and LIBSVM. The values of features in data sets are discritized by using mean value of each feature and applying discretization algorithm using sub-optimal agglomerative clustering. We have presented results of numerical experiments. These results clearly demonstrate that the proposed algorithm achieves comparable or better performance in comparison with traditional Bayesian Networks.

Our future work is applying the proposed algorithm to more complicated problems for learning BNs, e.g., problems with incomplete data, hidden variables, and multi class data sets.

REFERENCES

[1] H. Akaike, *Analysis of Cross Classified Data by AIC*. Ann. Inst. Statist. pp. 185-197, 1978.

[2] A. Asuncion and D. Newman, *UCI machine learning repository*. School of Information and Computer Science, University of California, 2007.

 ISBN: 978-1-61208-227-1

http://www.ics.uci.edu/mlearn/MLRepository.html, accessed May 2012.

[3] C. Chang and C. Lin, *LIBSVM: A library for support vector machines*. http://www.csie.ntu.edu.tw/cjlin/libsvm, accessed May 2012.

[4] J. Cheng, D. Bell, and W. Liu, *Learning Belief Networks from Data*. An Information Theory Based Approach. Artificial Intelligence, 137. pp. 43-90, 2002.

[5] G. F. Cooper and E. Herskovits, *A Bayesian Method for Constructing Bayesian Belief Networks from Databases*. Conference on Uncertainty in AI. pp. 86-94, 1990.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery: An overview*. In Advances in Knowledge Discovery in Data Mining. AAAI Press, Menlo Park, CA. pp. 1-34, 1996.

[7] N. Friedman, D. Geiger, and M. Goldszmidt, *Bayesian network classifiers*. Machine Learning 29. pp. 131-163, 1997.

[8] B. Haeupler, T. Kavitha, R. Mathew, S. Sen, and R. E. Tarjan, *Incremental Cycle Detection, Topological Ordering, and Strong Component Maintenance*. 35th ACM Transactions on Algorithms (TALG). pp. 1-33, 2012.

[9] D. Heckerman, A. Mamdani, and W. Michael, *Real-World Applications of Bayesian Networks*. Communications of the ACM. pp. 38-68, 1995.

[10] D. Heckerman, *Bayesian Networks for Data Mining*. Data Mining and Knowledge Discovery1. pp. 79-119, 1997.

[11] D. Heckerman, D. Chickering, and C. Meek, *Large-Sample Learning of Bayesian Networks is NP-Hard*. Journal of Machine Learning Research. pp. 1287-1330, 2004.

[12] E. Herskovits and G. F. Cooper, *An Entropy-Driven System for Construction of Probabilistic Expert Systems from Databases*. 6th Internatial Conference on Uncertainty in Artificial Intelligence (UAI90), Cambridge, MA, USA, Elsevier Science, New York. pp. 54-62, 1991.

[13] Y. Jing, V. Pavlovic, and J. Rehg, *Efficient discriminative learning of Bayesian network classifier via Boosted Augmented Naive Bayes*. The 22 nd International Conference on Machine Learning, Bonn, Germany. pp. 369 - 376, 2005.

[14] A. Jonsson and A. Barto, *Active Learning of Dynamic Bayesian Networks in Markov Decision Processes*. Springer-Verlag Berlin Heidelberg. pp. 273284, 2007.

[15] D. Kitakoshi, H. Shioya, and R. Nakano, *Empirical analysis of an on-line adaptive system using a mixture of Bayesian networks*. Information Sciences, Elsevier. pp. 2856-2874, 2010.

[16] M. Koivisto and K. Sood, *Exact Bayesian structure discovery in Bayesian networks*. Journal of Machine Learning 5. pp. 549573, 2004.

[17] P. Kontkanen, T. Silander, T. Roos, and P. Myllymki, *Bayesian Network Structure Learning Using Factorized NML Universal Models*. Information Theory and Applications Workshop (ITA-08), IEEE Press. pp. 272 - 276, 2008.

[18] P. Langley, W. Iba, and K. Thompson, *An Analysis of Bayesian Classifiers*. In 10th International Conference Artificial Intelligence, AAAI Press and MIT Press. pp. 223-228, 1992.

[19] W. Liaoa and Q. Ji, *Learning Bayesian network parameters under incomplete data with domain knowledge*. Pattern Recognition, Elsevier. pp. 3046-3056, 2009.

[20] P. Myllymaki, *Advantages of Bayesian networks in data mining and knowledge discovery*. http://www.bayesit.com/docs/advantages.html, 2005, acessed April 2012.

[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Networks of Plausible Inference, Morgan Kaufmann, 1988.

[22] J. Pearl, *Causality: Models, Reasonings and Inference*. Cambridge University Press. pp. 675685, 2003.

[23] M. Sahami, *Learning Limited Dependence Bayesian Classifiers*. In the 2nd International Conference. Knowledge Discovery and Data mining (KKD). pp. 335-338, 1996.

[24] T. Silander and P. Myllymaki, *A simple approach for finding the globally optimal Bayesian network structure*. In Proceedings of UAI-06. pp. 445-452, 2006.

[25] A. Singh and A. Moore, *Finding optimal Bayesian networks by dynamic programming*. Technical Report CMU-CALD-05-106, Carnegie Mellon University, 2005.

[26] G. Schwarz, *Estimating the Dimension of a Model*. Annals of Stastics. pp. 461-464, 1978.

[27] P. Spirtes, C. Glymour, and R. Sheines, *Causation, Prediction, and Search*. 1993.

[28] S. Taheri, M. Mammadov, and A. M. Bagirov, *Improving Naive Bayes Classifier Using Conditional Probabilities*. In the proceedings of Ninth Australasian Data Mining Conference (AusDM), Ballarat, Australia. Vol. 125. pp. 63-68, 2011.

[29] P. Weber, G. Medina-Oliva, C. Simon, and B. Iung, *Overview on Bayesian networks applications for dependability*. Risk-analysis and maintenance areas, Engineering Applications of Artificial Intelligence. pp. 671-682, 2010.

[30] A. Yatsko, A. M. Bagirov, and A. Stranieri, *On the Discretization of Continuous Features for Classification*. In the proceedings of Ninth Australasian Data Mining Conference (AusDM 2011), Ballarat, Australia. Vol. 125, 2011.

[31] C. Yuan, B. Malone, and X. Wu, *Learning Optimal Bayesian Networks Using A* Search*. In the proceedings of twenty second international joint conferenceon Artificial intelligence. pp. 2186-2191, 2011.

 ISBN: 978-1-61208-227-1

Comparison of Different Calculations of the Density-Based Local Outlier Factor

Vanda Vintrová*, Tomas Vintr†, Hana Řezanková‡
Department of Statistics and Probability
University of Economics, Prague
130 67 Prague, Czech Republic
*Email: *vanda.vintrova@vse.cz, †tomas.vintr@vse.cz, ‡hana.rezankova@vse.cz*

***Abstract*—In the paper, we propose several new density-based algorithms for outlier detection. We present the detailed synoptic theoretical analysis of the algorithms that compute the local outlier factor as a function of the densities of the neighborhood of the objects in a set of objects. Based on this analysis we propose a new calculation of the radius of the neighborhood and we create 66 algorithms to compute the outlier factor. All the algorithms are tested in the complex experiments to describe their basic and also specific characteristics. The results are presented and discussed. Intuitively it seems that the way how the radius of the neighborhood is calculated is important. This idea led to numerous modifications of this part of the algorithms, but on the basis of the experiments we demonstrate that these modifications have only little influence, and we describe which part of the algorithms influence the outlier score the most and we recommend three generally applicable algorithms with specific characteristics.**

Keywords-local outlier factor; density-based algorithm; outlier detection.

I. INTRODUCTION

In real data files, outliers, as objects considerably inconsistent with other objects from the same dataset, are often present. Some statistical and data mining methods regard these outliers as a noise that should be identified and eliminated as they falsify the analysis. However, outliers can also contain useful information and therefore it is important to investigate outliers in detail.

The outlier detection can be used in clustering methods that applied in the segmentation and typology of students [1], in the text mining and consequently in the information retrieval [2], in the database merge in medicine [3], or in the prediction of inflation [4].

There are different approaches to outlier detection. We focus on local density-based algorithms that can capture not only global outliers, but also local outliers. The original concept to score the local outliers compares a local density of an object with local densities of its k-nearest neighbors [5]. There are several modifications of this approach. In general, we can say that local density-based algorithms for outlier detection assign to every object of a set of objects a value that quantifies the density of the neighborhood of the object, and by comparing the value with the values of the objects in its neighborhood they assign to every object a score representing a degree of being an outlier. The score is mostly computed as a ratio of a density of a neighborhood of an object to an average density of neighborhoods of the objects in the object's neighborhood.

The important subject of the papers concerning about the local density-based algorithms has been the attempt to determine a meaningful neighborhood of an object that should be compared [5], [6], [7]. It is startling that even though these methods have been widely developed, the fundamentals are not synoptically formalized. That leads to the existence of many confusing structures whose idea is difficult to understand.

There exist two basic approaches for determining a density of a neighborhood. The first one is to determine a radius and then to compute how many objects lie in the neighborhood $\mathbf{O}(\mathbf{x}_p)$ of the object $\mathbf{x}_p$, where the radius of the neighborhood R_p is the parameter of the algorithm [8], [7]. As the analysis is usually performed on all the N objects of the set of objects, $p = 1, \ldots, N$. The second approach is determining the number k of the nearest neighbors of the object $\mathbf{x}_p$ and then to find out the radius as a function of the distance between the object $\mathbf{x}_p$ and the k nearest neighbors [5], [6], [9], [10], [11]. In both cases, it is necessary to have a priori knowledge of the set of objects. In the first case we need to know at least a range of clusters in the set of objects and in the second case we need to know at least a minimal number of objects in clusters in the set of objects. Usually, it is more difficult to determine the minimal range correctly; therefore, generally, it is more proper to determine the minimal number of objects in a cluster, as it can also represent a border between clusters that will be considered for the analysis and clusters that are too small for the intended analysis and will be considered as a noise.

In the paper, the computation of the outlier factor is synoptically explained. First, there is a discussion about how to calculate the average radius, then how to calculate the radius of the object's neighborhood. On the basis of these mentioned discussions, we propose 66 different combinations of the averages and radii to calculate the outlier factor. Some of these combinations represent the original algorithms, but most of them are newly proposed by us. In the fourth section, the algorithms are applied on synthetic datasets and compared in the complex experiments, the results are presented and discussed and in the last section

 ISBN: 978-1-61208-227-1

we propose the recommendation which algorithm to apply for what purpose or dataset type.

II. Related Work

Breunig et al. [5] is the first to introduce the concept of local density outliers and a local outlier factor (LOF). It compares a local density of an object with local densities of its k-nearest neighbors. The local density is estimated by a specific distance at which a point can be reached from its neighbors, called reachability distance, what produces stable results within clusters. LOF value of approximately 1 indicates that the point is located in a region of homogenous density. Higher LOF values signify an outlier, as it is a degree of being an outlier, but the scaling is different for different datasets.

An advantage of the LOF algorithm is that it can detect outliers even if the clusters of a dataset have different density and different size. This algorithm depends only on one parameter k; however this parameter strongly affects the outlierness of an object. If the parameter k is set too low, LOF does not detect outliers which are close to a dense cluster if the parameter is set too high, small clusters are regarded as outliers.

The LOF algorithm was modified several times especially with an aim to speed up the algorithm. An improvement of LOF known as Connectivity Outlier Factor (COF) [11] was proposed to overcome an ineffectiveness of LOF in detecting outliers in sparse datasets. Another modification of the LOF algorithm is LOF', LOF" and GridLOF [6]. LOF' simplifies the formula of LOF for ease of understanding, LOF" distinguishes between a neighborhood for computing the density of an object and a neighborhood for comparing the densities of the neighbors of an object. The GridLOF utilizes grid-based method to prune objects that are not outliers and then compute LOF score.

Another density-based algorithm was proposed by Papadimitriou et al. [7] named Local Correlation Integral (LOCI) based on the idea of a multi-granularity deviation factor (MDEF). The difference between LOF and LOCI is that LOCI uses neighborhood instead of k nearest neighbors. LOCI is less sensitive to input parameters than LOF.

The outlier scores provided by various outlier algorithms differ widely in their scale, range and meaning. For most methods the outlier scores are not comparable from dataset to dataset, for many methods the outlier scores are not comparable even within one dataset. The same outlier score for one object means that this object is an outlier and for another object (even within the same dataset, but from a different cluster) that this object is not extraordinary.

To overcome this problem, a new method has been proposed in [12] to unify outlier scores provided by different outlier algorithms. They propose two types of operations, regularization and normalization. Regularization means that the score is transformed into a range $[0; \infty)$, score equals approximately 0 for inliers, higher values signify outliers. The outlier factor can be regularized only if there exist an unambiguous numerous border between inliers and outliers. Such a border is not common for the existing algorithms. Normalization transforms scores into a range $[0; 1]$. These transformations do not change the ordering obtained by the original score. The contribution of this approach is not only unification of outlier scores, but also the fact that these operations can increase a contrast between outlier and inlier scores. A transformed outlier score is a rough probability, if an object is an outlier. Transformed outlier scores are therefore easier to understand and to interpret.

III. Computation of Outlier Factor

The outlier factor (degree of outlierness) is defined as the ratio of the radius of the neighborhood of the object to the average radius of the neighborhoods of the objects in the neighborhood of the object $\mathbf{x}_p$, i. e.,

$$OF = R_p / R_{avg} \ . \tag{1}$$

It means that the more is the object $\mathbf{x}_p$ suspected of being an outlier the higher is the outlier factor.

A. Discussion about the Average Radius

The density of the neighborhood of the selected object $\mathbf{x}_p$ can be defined as

$$d = k / C_n R^n \ , \tag{2}$$

where k is the number of objects in the neighborhood of the object $\mathbf{x}_p$ and $C_n R^n$ is the volume of the neighborhood (n-dimensional hypersphere), $C_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$.

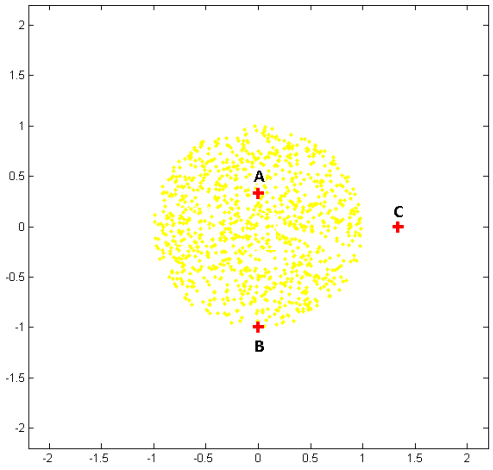
The outlier factor is calculated according to the formula

$$OF = \frac{\sqrt[n]{\frac{\sum_{i=1}^{k_p} \frac{k_i}{C_n R_i^n}}{k_p}}}{\sqrt[n]{\frac{k_p}{C_n R_p^n}}} \ , \tag{3}$$

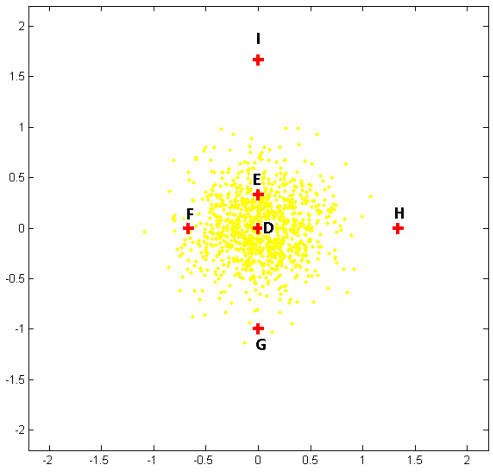
where R_i is the individual radius of the neighborhood $\mathbf{O}(\mathbf{x}_i)$ of the object $\mathbf{x}_i \in \mathbf{O}(\mathbf{x}_p)$, k_i is the number of objects in the individual neighborhoods $\mathbf{O}(\mathbf{x}_i)$, k_p is the number of objects in the neighborhood $\mathbf{O}(\mathbf{x}_p)$ and R_p is the radius of its neighborhood $\mathbf{O}(\mathbf{x}_p)$.

If the radius of the neighborhood of every object contains exactly k objects, the formula can be easily modified as follows:

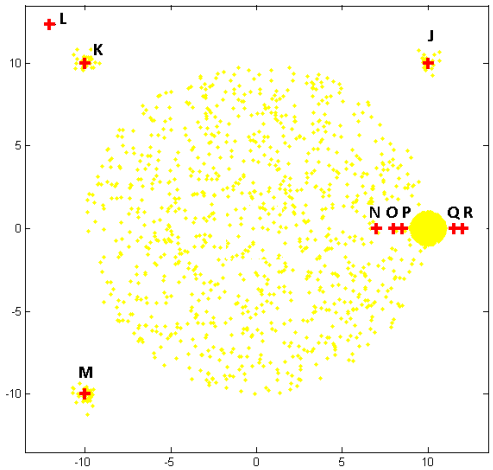
$$OF = \frac{\sqrt[n]{\frac{\sum_{i=1}^{k} \frac{k}{C_n R_i^n}}{k}}}{\sqrt[n]{\frac{k}{C_n R_p^n}}} = \frac{\sqrt[n]{\frac{\sum_{i=1}^{k} \frac{k}{R_i^n}}{k}}}{\sqrt[n]{\frac{k}{R_p^n}}} = \frac{\sqrt[n]{\frac{\sum_{i=1}^{k} \frac{1}{R_i^n}}{k}}}{\sqrt[n]{\frac{1}{R_p^n}}} = \\ = \frac{R_p}{\sqrt[-n]{\frac{\sum_{i=1}^{k} R_i^{-n}}{k}}} = \frac{R_p}{R_{avg}} \ , \tag{4}$$



(a) The depiction of the dataset generated by the uniform distribution.



(b) The depiction of the dataset generated by the normal distribution.



(c) The depiction of the dataset generated for the third experiment.

Figure 1. The illustrative pictures of the objects (·) in the datasets and added objects (+) for the outlier factor tests.

where

$$R_{avg} = \sqrt[-n]{\frac{\sum_{i=1}^{k} R_i^{-n}}{k}} \quad (5)$$

is the radius of the average dense neighborhood.

It is important to say that we define the score differently than the authors of LOF [5] and LOF' [6], who define the outlier factor as

$$OF = \frac{R_p}{\sqrt[-1]{\frac{\sum_{i=1}^{k} R_i^{-1}}{k}}} \ . \quad (6)$$

From this formula, it is evident that the authors replace the volume by the radius what we do not consider a felicitous solution.

B. Discussion about the Radius

The radius of the neighborhood of the object $\mathbf{x}_p$ is in the algorithms LOF' and DSNOF [9] an average of a set of distances $\mathbf{D}_p = \{d(\mathbf{x}_p, \mathbf{x}_i)\}_{i=1}^{k}$, where $d(\mathbf{a}, \mathbf{b})$ is the Euclidean distance of an object $\mathbf{a}$ from $\mathbf{b}$. The LOF' algorithm finds the maximum distance and the DSNOF algorithm finds the median distance. Using median in the DSNOF algorithm is a similar idea as using k_1 and k_2, where $k_2 < k_1$, in the LOF" [6] algorithm. The LOF algorithm uses more difficult calculation, the radius is computed as the arithmetic mean of the values that are either the distances $d(\mathbf{x}_p, \mathbf{x}_i)$ or distances $d(\mathbf{x}_i, \mathbf{x}_{ik})$ between an object $\mathbf{x}_i$ and its k-th nearest neighbor $\mathbf{x}_{ik}$. The set $\mathbf{Q}_p = \{q_i\}_{i=1}^{k}$, which is used to calculate the radius R_p in LOF, contains values that fulfill the condition $q_i = max(\{d(\mathbf{x}_p, \mathbf{x}_i,), d(\mathbf{x}_i, \mathbf{x}_{ik})\})$. It is expected that this operation decreases the radius of the neighborhood of the object on the border of the cluster and therefore objects on the border of clusters obtain lower outlier factor OF then in case of the LOF' algorithm. It seems that to calculate R_p as the mean of distances similar to (5) is geometrically meaningful alternative to the computation of the arithmetic mean.

We assume, in the case of one very numerous cluster generated by the uniform or normal distribution (without a noise) and simultaneously if quantile or mean of a set of distances is used as a radius determination, that the radius of the neighborhood of the object on a border of this cluster has to be approximately twice greater than the radius of its k-th most distant neighbor. From this reflection, we propose to determine the radius not only as the average of distances, but also to use some kind of a measure of dispersion. Inspired by the Box's M test that uses the determinants of the sample covariance matrices to test the equality of covariance matrices we propose to determine the radius using the determinants of the sample covariance matrices. As the number of objects in the sample is the same, we do not have to adjust the determinants and we can compare them

 ISBN: 978-1-61208-227-1

directly. We compute R_p as a determinant of a covariance matrix $\mathbf{C}_p$ of a set of objects $\mathbf{K}_p = \{\{\mathbf{x}_i\}_{i=1}^k, \mathbf{x}_p\}$,

$$R_p = \det \mathbf{C}_p \ . \tag{7}$$

C. Determining the Outlier Factor

On the basis of the above mentioned reflections, we can compute the outlier factor OF in several ways. The radius R_p of the object $\mathbf{x}_p$ can be calculated from the set $\mathbf{D}_p$ or $\mathbf{Q}_p$ using the following characteristics: first decile, maximum, median, arithmetic mean and the mean similar to (5). Minimum is not an appropriate characteristic because of its low information value about the neighborhood of $\mathbf{x}_p$ and consequent computational instability. The eleventh possibility to calculate the radius R_p is to compute the determinant of a covariance matrix $\mathbf{C}_p$ of a set of objects $\mathbf{K}_p$ proposed by us.

The calculation of R_p can be combined with the calculation of R_{avg}, which can be calculated according to the original algorithms as a harmonic mean of a set of radii R_i of neighborhoods of objects $\mathbf{x}_i$

$$R_{avg} = \sqrt[-1]{\frac{\sum_{i=1}^k R_i^{-1}}{k}} \ , \tag{8}$$

or as an average (5) or as a maximum, minimum, median or ninth decile of a set of a set $\{R_i\}$. We have to mention that a maximum, resp. a minimum of $\{R_i\}$ is equivalent to minimum, resp. maximum of densities, no matter whether the density is computed as a function R^{-1} or R^{-n}. By combining the different averages, radii and sets we create 66 $(11 \cdot 6)$ algorithms to calculate the outlier score.

Because there are many combinations, it is necessary to distinguish individual computations, systematically. For the purpose of this paper we have decided to create the names of individual combinations by combining the shortcuts of the functions included. The first characters represent a shortcut of the chosen average used for the calculation of the radius R_p, the following capital letter represents the set used for the computation of R_p and the characters on the third place represent a shortcut of the average used for the calculation of R_{avg}. Minimum is denoted as *min*, maximum as *max*, median as *med*, arithmetical mean as *mean*, harmonic mean as *hean*, the mean defined in (5) as *nean*, the first decile as 0.1 and the ninth decile as 0.9. The calculation of the radius R_p as a determinant of the covariance matrix will be denoted with the prefix *det*.

For example, the original LOF algorithm will be denoted as *meanQhean*, because it computes R_p as an arithmetic mean of the set $\mathbf{Q}_p$ and R_{avg} is computed as a harmonic mean. LOF' algorithm will be denoted as *maxDhean*, because it computes R_p as a maximum of the set $\mathbf{D}_p$ and R_{avg} is computed as a harmonic mean. An algorithm that will use the determinant of the covariance matrix of the set $\mathbf{K}_p$ to compute R_p and that will compute R_{avg} according to (5) will be denoted as *detKnean*, and so on.

IV. Experiments

We compared the algorithms in three experiments. The first two experiments are very simple, just to show the basic characteristics of the algorithms. We generated two datasets consisting of 1000 vectors. The first dataset was generated by the two–dimensional uniform distribution within the borders of the sphere with radius 1 and center $(0,0)^T$. We added 3 vectors with the coordinates $\mathbf{v}_A = (0, \frac{1}{3})^T$, $\mathbf{v}_B = (0,-1)^T$ and $\mathbf{v}_C = (\frac{4}{3}, 0)^T$ (see Fig. 1 (a)). The second dataset was created by the two–dimensional normal distribution with the mean $(0,0)^T$ and the standard deviation of every variable $\sigma = \frac{1}{3}$. We added 6 vectors to the datasets with the coordinates $\mathbf{v}_D = (0,0)^T$, $\mathbf{v}_E = (0, \frac{1}{3})^T$, $\mathbf{v}_F = (-\frac{2}{3}, 0)^T$, $\mathbf{v}_G = (0,-1)^T$, $\mathbf{v}_H = (\frac{4}{3}, 0)^T$ and $\mathbf{v}_I = (0, \frac{5}{3})^T$ (see Fig. 1 (b)). For both experiments we set $k = 40$. We performed both experiments ten times. The sample mean and the sample standard deviation of outlier factors of the added vectors for the tested algorithms are presented in the Table I, where the double vertical line separates these two experiments and the simple vertical lines highlight the hypothetical boarder of the clusters. We suppose that the mean values of the computed outlier factors should be strongly higher for the vectors to the right from the line to highlight the outliers. The horizontal lines separate the different groups of algorithms.

The third experiment is more complex to show further characteristics of the algorithms. There are 5 clusters in the dataset and we add 9 vectors denoted $\mathbf{v}_J$ to $\mathbf{v}_R$ on which we will demonstrate the behavior of the algorithms. There is one big cluster consisting of 1000 vectors created by the two–dimensional uniform distribution with the center $(0,0)^T$ and radius 10 units. On the border of this cluster is the center $(10,0)^T$ of another cluster with the radius 1 unit consisting also of 1000 vectors created by the two–dimensional uniform distribution. Around the big cluster there are three other clusters consisting of 21, 40 and 40 vectors generated by the two–dimensional normal distribution with mean vectors $\mathbf{v}_J = (10,10)^T$, $\mathbf{v}_K = (-10,10)^T$ and $\mathbf{v}_M = (-10,-10)^T$ respectively. Next to the cluster with the mean vector $\mathbf{v}_K$ is an outlying vector $\mathbf{v}_L = (-12,12)^T$. The vectors $\mathbf{v}_K$ and $\mathbf{v}_L$ have a similar set of k neighbors, but the vector $\mathbf{v}_L$ is an obvious outlier while the vector $\mathbf{v}_K$ is an obvious inlier. Each of the vectors $\mathbf{v}_K$ and $\mathbf{v}_M$ has one significantly distant vector in its neighborhood, for the vector $\mathbf{v}_K$ it is the outlier $\mathbf{v}_L$ and for the vector $\mathbf{v}_M$ it is an inlier from the big sparse cluster in the middle of the dataset, but the vector $\mathbf{v}_M$ does not belong to the set of k neighbors of this vector. The vectors $\mathbf{v}_J$, $\mathbf{v}_Q$ and $\mathbf{v}_M$ are placed in the centers of the small clusters and therefore they are inliers. The vector $\mathbf{v}_N = (7,0)^T$ belongs to the cluster generated by the two–dimensional uniform distribution, within its neighborhood are few or none vectors

 ISBN: 978-1-61208-227-1

Table I
OUTLIER FACTORS ($\bar{x} \pm s'$) OF ADDED VECTORS (1. AND 2. EXPERIMENT).

	A	B	C	D	E	F	G	H	I
0.1Dmin	1.7 ± 0.39	2.6 ± 0.81	9.6 ± 1.25	1.5 ± 0.38	1.7 ± 0.69	2.2 ± 0.77	4.9 ± 1.03	13 ± 2.77	18 ± 4.24
0.1Dnean	1.1 ± 0.18	1.5 ± 0.37	6.4 ± 0.57	1.0 ± 0.15	1.0 ± 0.31	1.3 ± 0.22	2.7 ± 0.38	6.9 ± 1.08	9.7 ± 1.19
0.1Dhean	1.1 ± 0.17	1.5 ± 0.34	6.2 ± 0.54	1.0 ± 0.15	1.0 ± 0.30	1.2 ± 0.21	2.4 ± 0.32	6.3 ± 0.96	8.9 ± 0.95
0.1Dmed	1.1 ± 0.18	1.4 ± 0.31	6.2 ± 0.64	1.0 ± 0.14	0.9 ± 0.26	1.2 ± 0.24	2.3 ± 0.40	6.0 ± 0.91	8.3 ± 1.19
0.1D0.9	0.8 ± 0.13	1.1 ± 0.22	4.5 ± 0.31	0.7 ± 0.11	0.7 ± 0.21	0.7 ± 0.13	1.1 ± 0.26	2.7 ± 0.44	4.2 ± 0.88
0.1Dmax	0.7 ± 0.13	0.9 ± 0.22	3.8 ± 0.36	0.6 ± 0.12	0.6 ± 0.17	0.5 ± 0.15	0.7 ± 0.24	1.6 ± 0.37	2.6 ± 0.65
neanDmin	5.7 ± 8.32	5.7 ± 2.68	16 ± 8.77	3.6 ± 2.82	2.1 ± 0.59	3.0 ± 1.10	7.5 ± 1.94	22 ± 10.9	23 ± 8.85
neanDnean	1.8 ± 1.71	2.2 ± 0.51	7.0 ± 1.71	1.5 ± 0.76	1.2 ± 0.27	1.3 ± 0.38	3.1 ± 0.29	8.2 ± 2.20	10 ± 1.58
neanDhean	1.2 ± 0.49	1.8 ± 0.36	6.2 ± 0.89	1.3 ± 0.46	1.1 ± 0.27	1.2 ± 0.33	2.7 ± 0.28	6.7 ± 1.30	8.7 ± 0.74
neanDmed	0.9 ± 0.26	1.5 ± 0.34	5.3 ± 0.51	1.0 ± 0.21	1.0 ± 0.30	1.1 ± 0.34	2.3 ± 0.42	5.6 ± 0.65	7.6 ± 0.80
neanD0.9	0.7 ± 0.22	1.1 ± 0.23	3.9 ± 0.25	0.8 ± 0.13	0.8 ± 0.19	0.7 ± 0.19	1.2 ± 0.27	2.7 ± 0.41	3.8 ± 0.57
neanDmax	0.7 ± 0.21	0.9 ± 0.23	3.2 ± 0.28	0.7 ± 0.14	0.7 ± 0.16	0.5 ± 0.17	0.8 ± 0.17	1.6 ± 0.38	2.4 ± 0.45
medDmin	1.2 ± 0.15	1.7 ± 0.22	4.1 ± 0.42	1.3 ± 0.15	1.2 ± 0.10	1.8 ± 0.25	3.1 ± 0.29	5.8 ± 0.76	8.1 ± 1.11
medDnean	1.0 ± 0.10	1.4 ± 0.13	3.2 ± 0.30	1.0 ± 0.07	1.0 ± 0.06	1.2 ± 0.14	2.1 ± 0.13	4.0 ± 0.38	5.2 ± 0.39
medDhean	1.0 ± 0.10	1.4 ± 0.12	3.1 ± 0.29	1.0 ± 0.07	1.0 ± 0.06	1.2 ± 0.13	2.0 ± 0.11	3.8 ± 0.34	5.0 ± 0.35
medDmed	1.0 ± 0.09	1.4 ± 0.13	3.1 ± 0.26	1.0 ± 0.07	1.0 ± 0.06	1.2 ± 0.13	2.1 ± 0.17	3.9 ± 0.35	5.0 ± 0.44
medD0.9	0.9 ± 0.10	1.1 ± 0.08	2.5 ± 0.15	0.9 ± 0.07	0.8 ± 0.07	0.8 ± 0.11	1.2 ± 0.13	2.1 ± 0.28	2.8 ± 0.39
medDmax	0.8 ± 0.10	1.0 ± 0.08	2.2 ± 0.17	0.9 ± 0.07	0.8 ± 0.06	0.6 ± 0.08	0.8 ± 0.14	1.4 ± 0.24	2.0 ± 0.28
meanDmin	1.2 ± 0.09	1.6 ± 0.17	4.0 ± 0.51	1.1 ± 0.10	1.1 ± 0.11	1.7 ± 0.23	3.0 ± 0.24	5.8 ± 0.70	8.2 ± 0.99
meanDnean	1.0 ± 0.06	1.4 ± 0.09	3.3 ± 0.30	1.0 ± 0.06	1.0 ± 0.06	1.2 ± 0.12	2.1 ± 0.14	4.0 ± 0.42	5.4 ± 0.36
meanDhean	1.0 ± 0.06	1.4 ± 0.08	3.2 ± 0.29	1.0 ± 0.06	1.0 ± 0.06	1.2 ± 0.11	2.0 ± 0.12	3.9 ± 0.38	5.1 ± 0.33
meanDmed	1.0 ± 0.05	1.4 ± 0.07	3.3 ± 0.31	1.0 ± 0.06	1.0 ± 0.05	1.2 ± 0.12	2.0 ± 0.19	4.0 ± 0.38	5.2 ± 0.49
meanD0.9	0.9 ± 0.07	1.1 ± 0.05	2.6 ± 0.15	0.9 ± 0.06	0.8 ± 0.06	0.8 ± 0.08	1.1 ± 0.12	2.2 ± 0.29	2.9 ± 0.41
meanDmax	0.9 ± 0.07	1.0 ± 0.05	2.4 ± 0.14	0.9 ± 0.07	0.8 ± 0.07	0.6 ± 0.05	0.8 ± 0.14	1.5 ± 0.26	2.1 ± 0.31
maxDmin	1.1 ± 0.05	1.6 ± 0.13	3.0 ± 0.34	1.1 ± 0.07	1.2 ± 0.09	1.7 ± 0.19	2.7 ± 0.16	4.6 ± 0.55	6.3 ± 0.67
maxDnean	1.0 ± 0.03	1.3 ± 0.07	2.5 ± 0.24	1.0 ± 0.04	1.0 ± 0.05	1.2 ± 0.11	1.9 ± 0.11	3.3 ± 0.32	4.2 ± 0.26
maxDhean	1.0 ± 0.03	1.3 ± 0.07	2.4 ± 0.24	1.0 ± 0.04	1.0 ± 0.05	1.2 ± 0.10	1.8 ± 0.10	3.1 ± 0.30	4.0 ± 0.23
maxDmed	1.0 ± 0.03	1.3 ± 0.08	2.4 ± 0.27	1.0 ± 0.04	1.0 ± 0.04	1.2 ± 0.10	1.9 ± 0.14	3.2 ± 0.31	4.1 ± 0.29
maxD0.9	0.9 ± 0.05	1.0 ± 0.04	2.0 ± 0.17	0.9 ± 0.04	0.9 ± 0.08	0.8 ± 0.07	1.1 ± 0.12	1.9 ± 0.23	2.4 ± 0.26
maxDmax	0.9 ± 0.05	1.0 ± 0.04	1.9 ± 0.12	0.9 ± 0.04	0.8 ± 0.08	0.7 ± 0.04	0.8 ± 0.13	1.4 ± 0.21	1.8 ± 0.22
0.1Qmin	1.1 ± 0.05	1.2 ± 0.11	2.2 ± 0.28	1.0 ± 0.03	1.1 ± 0.05	1.5 ± 0.15	2.2 ± 0.17	3.7 ± 0.41	5.9 ± 0.68
0.1Qnean	1.0 ± 0.02	1.1 ± 0.06	2.1 ± 0.25	1.0 ± 0.02	1.0 ± 0.02	1.2 ± 0.08	1.7 ± 0.11	2.9 ± 0.34	4.2 ± 0.32
0.1Qhean	1.0 ± 0.02	1.1 ± 0.05	2.1 ± 0.25	1.0 ± 0.02	1.0 ± 0.02	1.1 ± 0.08	1.6 ± 0.10	2.8 ± 0.33	4.1 ± 0.30
0.1Qmed	1.0 ± 0.02	1.1 ± 0.06	2.1 ± 0.26	1.0 ± 0.01	1.0 ± 0.01	1.1 ± 0.08	1.6 ± 0.15	3.0 ± 0.36	4.2 ± 0.34
0.1Q0.9	1.0 ± 0.04	1.0 ± 0.03	1.9 ± 0.20	1.0 ± 0.03	0.9 ± 0.03	0.9 ± 0.06	1.1 ± 0.09	1.9 ± 0.30	2.8 ± 0.39
0.1Qmax	0.9 ± 0.05	1.0 ± 0.04	1.8 ± 0.17	0.9 ± 0.05	0.9 ± 0.05	0.7 ± 0.05	0.9 ± 0.12	1.4 ± 0.30	2.1 ± 0.38
neanQmin	1.1 ± 0.03	1.3 ± 0.10	2.3 ± 0.25	1.0 ± 0.03	1.1 ± 0.04	1.6 ± 0.15	2.3 ± 0.15	3.8 ± 0.41	5.7 ± 0.64
neanQnean	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.21	1.0 ± 0.01	1.0 ± 0.01	1.2 ± 0.08	1.7 ± 0.10	2.9 ± 0.31	4.1 ± 0.26
neanQhean	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.21	1.0 ± 0.01	1.0 ± 0.01	1.2 ± 0.07	1.7 ± 0.09	2.9 ± 0.29	4.0 ± 0.24
neanQmed	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.22	1.0 ± 0.01	1.0 ± 0.02	1.2 ± 0.06	1.7 ± 0.14	3.0 ± 0.31	4.0 ± 0.28
neanQ0.9	1.0 ± 0.04	1.0 ± 0.02	2.0 ± 0.16	1.0 ± 0.02	0.9 ± 0.03	0.9 ± 0.05	1.1 ± 0.09	1.9 ± 0.25	2.6 ± 0.31
neanQmax	0.9 ± 0.04	1.0 ± 0.03	1.9 ± 0.12	0.9 ± 0.03	0.9 ± 0.03	0.7 ± 0.05	0.9 ± 0.12	1.4 ± 0.24	2.0 ± 0.28
medQmin	1.1 ± 0.04	1.3 ± 0.11	2.4 ± 0.27	1.0 ± 0.04	1.1 ± 0.06	1.6 ± 0.19	2.4 ± 0.15	4.1 ± 0.48	5.9 ± 0.66
medQnean	1.0 ± 0.02	1.2 ± 0.05	2.2 ± 0.23	1.0 ± 0.02	1.0 ± 0.02	1.2 ± 0.11	1.8 ± 0.10	3.1 ± 0.32	4.2 ± 0.26
medQhean	1.0 ± 0.02	1.1 ± 0.05	2.2 ± 0.23	1.0 ± 0.02	1.0 ± 0.02	1.2 ± 0.10	1.7 ± 0.09	3.0 ± 0.30	4.1 ± 0.23
medQmed	1.0 ± 0.02	1.2 ± 0.05	2.2 ± 0.26	1.0 ± 0.01	1.0 ± 0.02	1.2 ± 0.09	1.7 ± 0.15	3.1 ± 0.32	4.1 ± 0.27
medQ0.9	1.0 ± 0.04	1.0 ± 0.04	2.0 ± 0.18	1.0 ± 0.03	0.9 ± 0.04	0.9 ± 0.07	1.1 ± 0.10	2.0 ± 0.25	2.6 ± 0.31
medQmax	0.9 ± 0.05	1.0 ± 0.05	1.9 ± 0.14	0.9 ± 0.04	0.9 ± 0.05	0.7 ± 0.04	0.9 ± 0.13	1.4 ± 0.23	2.0 ± 0.27
meanQmin	1.1 ± 0.03	1.3 ± 0.10	2.3 ± 0.25	1.0 ± 0.03	1.2 ± 0.04	1.6 ± 0.15	2.3 ± 0.14	3.8 ± 0.42	5.6 ± 0.64
meanQnean	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.21	1.0 ± 0.01	1.0 ± 0.01	1.2 ± 0.07	1.7 ± 0.11	2.9 ± 0.30	4.0 ± 0.25
meanQhean	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.21	1.0 ± 0.01	1.0 ± 0.01	1.2 ± 0.07	1.6 ± 0.09	2.8 ± 0.29	3.9 ± 0.23
meanQmed	1.0 ± 0.01	1.1 ± 0.05	2.1 ± 0.22	1.0 ± 0.01	1.0 ± 0.02	1.2 ± 0.06	1.7 ± 0.14	3.0 ± 0.29	3.9 ± 0.28
meanQ0.9	1.0 ± 0.03	1.0 ± 0.02	2.0 ± 0.15	1.0 ± 0.02	0.9 ± 0.02	0.9 ± 0.04	1.1 ± 0.09	1.9 ± 0.25	2.6 ± 0.29
meanQmax	0.9 ± 0.04	1.0 ± 0.03	1.9 ± 0.11	0.9 ± 0.03	0.8 ± 0.03	0.7 ± 0.05	0.9 ± 0.11	1.4 ± 0.24	2.0 ± 0.27
maxQmin	1.1 ± 0.06	1.3 ± 0.10	2.2 ± 0.28	1.1 ± 0.04	1.2 ± 0.10	2.0 ± 0.21	2.3 ± 0.41	3.2 ± 0.37	4.3 ± 0.48
maxQnean	1.0 ± 0.02	1.1 ± 0.03	1.9 ± 0.12	1.0 ± 0.04	1.0 ± 0.05	1.3 ± 0.12	1.6 ± 0.29	2.2 ± 0.21	2.9 ± 0.16
maxQhean	1.0 ± 0.02	1.1 ± 0.03	1.9 ± 0.12	1.0 ± 0.04	1.0 ± 0.05	1.3 ± 0.11	1.5 ± 0.26	2.1 ± 0.21	2.8 ± 0.15
maxQmed	1.0 ± 0.02	1.0 ± 0.03	1.9 ± 0.12	1.0 ± 0.04	1.0 ± 0.06	1.3 ± 0.13	1.5 ± 0.26	2.1 ± 0.23	2.7 ± 0.21
maxQ0.9	1.0 ± 0.02	1.0 ± 0.02	1.8 ± 0.10	0.9 ± 0.05	0.8 ± 0.04	0.8 ± 0.06	1.0 ± 0.05	1.4 ± 0.28	2.0 ± 0.17
maxQmax	1.0 ± 0.03	1.0 ± 0.02	1.8 ± 0.08	0.9 ± 0.05	0.8 ± 0.02	0.7 ± 0.11	0.9 ± 0.08	1.0 ± 0.10	1.7 ± 0.32
detKmin	1.7 ± 0.56	2.1 ± 1.07	3.9 ± 1.37	1.7 ± 0.49	1.9 ± 0.75	6.1 ± 2.64	14 ± 6.19	24 ± 10.4	51 ± 32.2
detKnean	1.2 ± 0.26	1.3 ± 0.33	2.7 ± 0.79	1.1 ± 0.25	1.1 ± 0.26	2.7 ± 0.87	6.7 ± 3.82	11 ± 4.41	21 ± 9.91
detKhean	1.1 ± 0.24	1.3 ± 0.27	2.6 ± 0.75	1.1 ± 0.21	1.0 ± 0.22	2.2 ± 0.61	5.3 ± 2.84	9.3 ± 3.34	17 ± 6.30
detKmed	1.1 ± 0.20	1.2 ± 0.14	2.5 ± 0.72	1.1 ± 0.21	1.0 ± 0.16	1.7 ± 0.39	3.9 ± 2.01	8.3 ± 2.72	13 ± 5.19
detK0.9	0.8 ± 0.22	0.8 ± 0.18	1.9 ± 0.58	0.7 ± 0.16	0.6 ± 0.14	0.6 ± 0.14	1.1 ± 0.13	2.4 ± 0.76	4.6 ± 1.10
detKmax	0.7 ± 0.20	0.8 ± 0.19	1.7 ± 0.52	0.6 ± 0.16	0.5 ± 0.11	0.4 ± 0.15	0.8 ± 0.16	1.2 ± 0.22	2.8 ± 0.86

 ISBN: 978-1-61208-227-1

Table II
OUTLIER FACTORS ($\bar{x} \pm s'$) OF ADDED VECTORS (3. EXPERIMENT).

	J	K	L	M	N	O	P	Q	R
0.1Dmin	1.3 ± 0.27	1.3 ± 0.36	22 ± 5.54	1.3 ± 0.41	4.3 ± 4.20	12 ± 3.83	12 ± 2.40	14 ± 3.61	25 ± 4.19
0.1Dnean	0.6 ± 0.14	0.7 ± 0.14	12 ± 1.67	0.8 ± 0.18	1.4 ± 0.62	6.5 ± 1.56	7.3 ± 1.02	8.7 ± 1.13	17 ± 1.94
0.1Dhean	0.6 ± 0.13	0.7 ± 0.13	11 ± 1.33	0.7 ± 0.15	1.2 ± 0.29	5.4 ± 1.28	6.8 ± 0.97	8.4 ± 0.97	16 ± 1.71
0.1Dmed	0.4 ± 0.11	0.6 ± 0.12	10 ± 1.13	0.7 ± 0.17	1.0 ± 0.20	5.9 ± 1.81	7.0 ± 0.73	8.5 ± 0.95	16 ± 1.82
0.1D0.9	0.3 ± 0.06	0.3 ± 0.08	5.5 ± 0.81	0.3 ± 0.09	0.8 ± 0.19	0.9 ± 0.10	2.0 ± 1.56	5.2 ± 0.50	10 ± 1.00
0.1Dmax	0.2 ± 0.05	0.1 ± 0.02	4.0 ± 1.01	0.2 ± 0.03	0.6 ± 0.13	0.7 ± 0.08	0.9 ± 0.07	0.5 ± 0.01	1.9 ± 0.03
neanDmin	1.9 ± 0.78	1.8 ± 0.74	27 ± 10.2	2.2 ± 0.79	5.2 ± 3.17	21 ± 16.5	17 ± 9.17	29 ± 21.8	48 ± 28.0
neanDnean	0.7 ± 0.23	0.8 ± 0.28	12 ± 3.01	0.9 ± 0.33	1.7 ± 0.54	7.9 ± 4.14	7.1 ± 1.93	11 ± 4.97	18 ± 5.99
neanDhean	0.6 ± 0.21	0.7 ± 0.24	11 ± 2.00	0.8 ± 0.29	1.4 ± 0.31	5.9 ± 2.62	6.0 ± 1.31	9.0 ± 2.40	15 ± 3.17
neanDmed	0.4 ± 0.19	0.6 ± 0.21	9.0 ± 0.85	0.7 ± 0.27	1.1 ± 0.24	5.5 ± 2.30	5.4 ± 1.00	7.1 ± 0.74	12 ± 1.28
neanD0.9	0.2 ± 0.11	0.3 ± 0.12	4.8 ± 0.53	0.3 ± 0.12	0.8 ± 0.19	0.8 ± 0.27	1.7 ± 1.13	4.9 ± 0.46	8.3 ± 0.78
neanDmax	0.2 ± 0.09	0.1 ± 0.03	3.4 ± 0.63	0.2 ± 0.07	0.7 ± 0.18	0.6 ± 0.20	0.7 ± 0.13	0.6 ± 0.01	1.7 ± 0.03
medDmin	2.1 ± 0.18	1.1 ± 0.05	7.6 ± 0.89	1.0 ± 0.07	3.8 ± 2.39	7.4 ± 0.70	4.7 ± 0.38	5.2 ± 0.31	9.1 ± 0.69
medDnean	1.3 ± 0.06	0.8 ± 0.05	5.5 ± 0.61	0.7 ± 0.05	1.4 ± 0.36	4.8 ± 0.65	3.5 ± 0.27	4.0 ± 0.23	7.0 ± 0.41
medDhean	1.3 ± 0.05	0.7 ± 0.05	5.3 ± 0.58	0.7 ± 0.05	1.2 ± 0.15	4.1 ± 0.65	3.3 ± 0.28	4.0 ± 0.22	6.9 ± 0.39
medDmed	1.1 ± 0.05	0.8 ± 0.05	5.4 ± 0.56	0.7 ± 0.04	1.0 ± 0.07	4.9 ± 1.27	3.6 ± 0.28	4.0 ± 0.25	7.0 ± 0.48
medD0.9	0.9 ± 0.02	0.5 ± 0.05	3.5 ± 0.30	0.5 ± 0.08	0.9 ± 0.04	0.8 ± 0.05	1.4 ± 0.78	3.1 ± 0.20	5.4 ± 0.32
medDmax	0.8 ± 0.03	0.1 ± 0.02	2.8 ± 0.38	0.2 ± 0.04	0.8 ± 0.05	0.8 ± 0.05	0.6 ± 0.02	0.6 ± 0.01	1.8 ± 0.03
meanDmin	2.1 ± 0.17	1.0 ± 0.01	6.0 ± 0.48	1.0 ± 0.01	3.7 ± 2.40	6.7 ± 0.68	4.5 ± 0.25	5.1 ± 0.32	8.9 ± 0.60
meanDnean	1.4 ± 0.07	0.8 ± 0.01	4.7 ± 0.38	0.8 ± 0.01	1.4 ± 0.36	4.5 ± 0.70	3.6 ± 0.26	4.1 ± 0.25	7.2 ± 0.44
meanDhean	1.3 ± 0.06	0.8 ± 0.01	4.6 ± 0.37	0.8 ± 0.01	1.2 ± 0.15	3.9 ± 0.69	3.4 ± 0.28	4.1 ± 0.24	7.1 ± 0.42
meanDmed	1.2 ± 0.04	0.8 ± 0.02	4.7 ± 0.32	0.8 ± 0.03	1.0 ± 0.07	4.8 ± 1.32	3.7 ± 0.25	4.2 ± 0.30	7.2 ± 0.51
meanD0.9	0.9 ± 0.02	0.5 ± 0.02	3.3 ± 0.24	0.5 ± 0.04	0.9 ± 0.04	0.8 ± 0.04	1.5 ± 0.83	3.2 ± 0.18	5.6 ± 0.25
meanDmax	0.9 ± 0.04	0.2 ± 0.01	2.8 ± 0.37	0.3 ± 0.03	0.9 ± 0.04	0.7 ± 0.03	0.6 ± 0.02	0.6 ± 0.01	1.8 ± 0.03
maxDmin	2.7 ± 0.15	1.3 ± 0.07	1.6 ± 0.11	1.6 ± 0.11	4.2 ± 2.13	5.2 ± 0.42	3.6 ± 0.12	3.8 ± 0.28	6.2 ± 0.58
maxDnean	1.7 ± 0.09	1.0 ± 0.01	1.3 ± 0.04	1.0 ± 0.01	1.5 ± 0.34	3.5 ± 0.38	2.7 ± 0.14	3.0 ± 0.15	5.0 ± 0.25
maxDhean	1.6 ± 0.07	1.0 ± 0.01	1.3 ± 0.04	1.0 ± 0.01	1.3 ± 0.16	3.1 ± 0.40	2.6 ± 0.16	2.9 ± 0.14	4.9 ± 0.25
maxDmed	1.4 ± 0.05	1.0 ± 0.02	1.2 ± 0.05	1.0 ± 0.02	1.0 ± 0.05	3.8 ± 0.79	2.8 ± 0.12	3.0 ± 0.19	5.0 ± 0.30
maxD0.9	1.0 ± 0.02	0.9 ± 0.03	1.1 ± 0.03	0.9 ± 0.01	0.9 ± 0.05	0.7 ± 0.05	1.3 ± 0.59	2.4 ± 0.11	4.0 ± 0.17
maxDmax	0.9 ± 0.03	0.8 ± 0.03	1.0 ± 0.03	0.9 ± 0.01	0.9 ± 0.05	0.6 ± 0.04	0.6 ± 0.03	0.6 ± 0.01	1.7 ± 0.03
0.1Qmin	2.5 ± 0.18	1.0 ± 0.00	1.0 ± 0.00	2.0 ± 0.14	4.1 ± 1.86	5.0 ± 0.38	2.9 ± 0.19	3.0 ± 0.22	5.7 ± 0.37
0.1Qnean	1.8 ± 0.10	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.01	1.4 ± 0.28	3.8 ± 0.49	2.5 ± 0.14	2.8 ± 0.18	5.2 ± 0.29
0.1Qhean	1.7 ± 0.08	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.2 ± 0.11	3.4 ± 0.50	2.5 ± 0.15	2.7 ± 0.17	5.2 ± 0.29
0.1Qmed	1.5 ± 0.05	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.08	4.3 ± 1.01	2.7 ± 0.13	2.8 ± 0.19	5.3 ± 0.35
0.1Q0.9	0.9 ± 0.02	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.00	0.8 ± 0.05	0.8 ± 0.06	1.4 ± 0.69	2.5 ± 0.16	4.8 ± 0.26
0.1Qmax	0.9 ± 0.04	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.00	0.7 ± 0.04	0.7 ± 0.05	0.5 ± 0.02	0.5 ± 0.01	1.9 ± 0.04
neanQmin	2.5 ± 0.16	1.0 ± 0.00	1.0 ± 0.00	1.9 ± 0.12	4.3 ± 2.05	5.0 ± 0.24	3.0 ± 0.14	3.1 ± 0.18	5.3 ± 0.32
neanQnean	1.8 ± 0.08	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.01	1.5 ± 0.33	3.7 ± 0.34	2.6 ± 0.11	2.8 ± 0.15	4.8 ± 0.25
neanQhean	1.6 ± 0.06	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.3 ± 0.13	3.3 ± 0.38	2.5 ± 0.11	2.8 ± 0.15	4.8 ± 0.24
neanQmed	1.4 ± 0.04	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.04	4.2 ± 0.90	2.7 ± 0.10	2.8 ± 0.17	4.9 ± 0.27
neanQ0.9	1.0 ± 0.01	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.02	0.7 ± 0.04	1.4 ± 0.66	2.5 ± 0.15	4.4 ± 0.23
neanQmax	0.9 ± 0.02	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.03	0.6 ± 0.03	0.6 ± 0.02	0.6 ± 0.01	1.7 ± 0.03
medQmin	2.6 ± 0.16	1.0 ± 0.01	1.0 ± 0.00	1.8 ± 0.12	4.5 ± 2.20	5.1 ± 0.25	3.1 ± 0.17	3.2 ± 0.13	5.5 ± 0.27
medQnean	1.8 ± 0.09	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.01	1.6 ± 0.37	3.7 ± 0.34	2.6 ± 0.13	2.8 ± 0.14	4.9 ± 0.24
medQhean	1.7 ± 0.07	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.3 ± 0.16	3.3 ± 0.38	2.5 ± 0.13	2.8 ± 0.14	4.9 ± 0.24
medQmed	1.4 ± 0.03	1.0 ± 0.01	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.02	4.1 ± 0.85	2.7 ± 0.13	2.9 ± 0.16	5.0 ± 0.28
medQ0.9	1.0 ± 0.01	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.03	0.7 ± 0.05	1.4 ± 0.64	2.5 ± 0.15	4.3 ± 0.25
medQmax	1.0 ± 0.02	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.04	0.6 ± 0.04	0.6 ± 0.02	0.6 ± 0.01	1.7 ± 0.02
meanQmin	2.5 ± 0.16	1.0 ± 0.00	1.0 ± 0.00	1.8 ± 0.11	4.3 ± 2.14	5.2 ± 0.26	3.1 ± 0.16	3.1 ± 0.19	5.3 ± 0.33
meanQnean	1.8 ± 0.08	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.5 ± 0.35	3.9 ± 0.36	2.6 ± 0.11	2.8 ± 0.15	4.8 ± 0.25
meanQhean	1.6 ± 0.06	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.3 ± 0.14	3.4 ± 0.39	2.5 ± 0.11	2.8 ± 0.15	4.7 ± 0.24
meanQmed	1.4 ± 0.03	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.03	4.4 ± 0.96	2.7 ± 0.12	2.8 ± 0.17	4.8 ± 0.27
meanQ0.9	1.0 ± 0.01	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.02	0.7 ± 0.03	1.4 ± 0.67	2.5 ± 0.15	4.3 ± 0.22
meanQmax	0.9 ± 0.01	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.03	0.7 ± 0.03	0.5 ± 0.02	0.6 ± 0.01	1.7 ± 0.03
maxQmin	2.2 ± 0.13	1.0 ± 0.03	1.0 ± 0.03	1.7 ± 0.08	3.8 ± 2.57	6.6 ± 0.63	4.2 ± 0.18	4.6 ± 0.38	4.5 ± 0.36
maxQnean	1.6 ± 0.07	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.4 ± 0.36	5.0 ± 0.88	3.7 ± 0.29	4.0 ± 0.20	4.0 ± 0.20
maxQhean	1.5 ± 0.05	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.2 ± 0.13	4.3 ± 0.85	3.5 ± 0.31	4.0 ± 0.19	4.0 ± 0.19
maxQmed	1.3 ± 0.06	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.04	5.6 ± 1.84	3.9 ± 0.31	4.0 ± 0.25	4.0 ± 0.23
maxQ0.9	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.03	0.9 ± 0.03	1.7 ± 1.14	3.8 ± 0.16	3.7 ± 0.18
maxQmax	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	0.9 ± 0.04	0.9 ± 0.04	0.6 ± 0.06	1.0 ± 0.00	1.0 ± 0.00
detKmin	9.0 ± 2.39	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.03	3080 ± 4071	679 ± 227	51 ± 32.2	23 ± 7.36	26 ± 8.95
detKnean	4.8 ± 1.03	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.02	592 ± 814	389 ± 104	32 ± 15.5	15 ± 4.13	18 ± 5.37
detKhean	3.7 ± 0.69	1.0 ± 0.01	1.0 ± 0.01	1.0 ± 0.02	137 ± 197	302 ± 82	29 ± 12.5	14 ± 3.80	17 ± 4.96
detKmed	1.7 ± 0.14	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.2 ± 0.27	365 ± 174	30 ± 11.1	15 ± 3.64	17 ± 4.66
detK0.9	1.0 ± 0.05	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.04	0.7 ± 0.10	0.1 ± 0.04	4.8 ± 6.28	8.9 ± 3.05	10 ± 3.83
detKmax	0.9 ± 0.15	0.9 ± 0.22	0.9 ± 0.22	0.1 ± 0.03	0.6 ± 0.10	0.1 ± 0.05	0.1 ± 0.02	0.9 ± 0.05	1.2 ± 0.07

 ISBN: 978-1-61208-227-1

from the much denser cluster. The vectors $\mathbf{v}_O = (8, 0)^T$, $\mathbf{v}_R = (11, 0)^T$ and $\mathbf{v}_P = (8.5, 0)^T$, $\mathbf{v}_Q = (10.5, 0)^T$ are in the same distance from the center of the dense cluster, so we can compare how can the algorithms detect the outliers in the dataset with noise ($\mathbf{v}_O$, $\mathbf{v}_P$) and without noise ($\mathbf{v}_Q$, $\mathbf{v}_R$). The vectors $\mathbf{v}_Q$ and $\mathbf{v}_R$ are obvious outliers, each of them lies within the set of k neighbors of the other vector, but $\mathbf{v}_R$ is the most distant vector in the neighborhood of the vector $\mathbf{v}_Q$ and also in its own neighborhood. Any other vector, except for the vector $\mathbf{v}_Q$ and $\mathbf{v}_R$, does not have the vector $\mathbf{v}_Q$ or $\mathbf{v}_R$ within its set of k neighbors (see Fig. 1 (c)). The parameter k is set on 40 in all the algorithms. The experiment was performed ten times. One can see the sample mean and the sample standard deviation of outlier factors of the added vectors for the tested algorithm in the Table II, where the vertical lines highlights the different groups of selected vectors. The horizontal lines separate the different groups of algorithms.

On the basis of the two first experiments, we can state that the usage of the set **Q** considerably decreases the deviation of the OF values from 1 compared to the set **D**. The shortcoming of the usage of the set **Q** is clearly demonstrated on the vector $\mathbf{v}_L$, which is in all the cases labeled as an inlier. It is possible to say that if there is k considerably detached vectors and the set **Q** is applied, then they will be all labeled as inliers, no matter what is their mutual position. The set **Q** considerably suppress the effect of the combination of low quantiles for the calculation of both R_p and R_{avg} simultaneously in the case that the cluster consists of less than k vectors. The algorithm *maxQmax* is the only algorithm that labeled the vector $\mathbf{v}_R$ as an inlier.

The results of the algorithms applying *detK* are very similar to the algorithms applying **Q**, both label the vector $\mathbf{v}_L$ as an inlier, but the OF value of outliers strongly increases. The vector $\mathbf{v}_L$ is labeled as an inlier, because the identical set **K** with the identical distribution was used for the calculation of the OF value of all the vectors within the cluster around the vector $\mathbf{v}_K$, and also of the vector $\mathbf{v}_L$. Unlike the algorithm *...Qmin* the algorithm *detKmin* labeled the vector $\mathbf{v}_M$ as an inlier.

Low quantiles applied for the calculation of R_p increase the deviation of the OF values from 1, high quantiles decrease the deviation of the OF values from 1. The average *nean* is similar to a very low quantile, whereas the average *mean* is similar to the median or a little higher quantile. The average *nean* is computationally unstable when applied on the set **D**, especially when combined with a low quantile for the calculation of R_{avg}. Vectors of the clusters consisting of less than k vectors, but simultaneously denser than their neighborhood, can be labeled as inliers, if a quantile low enough is applied for the calculation of R_p.

The OF values are influenced the most by the way how the R_{avg} is calculated. There are two extremes. When the minimum is applied, only the vector with the smallest R_p from all the vectors within its neighborhood is labeled as an inlier. In other words, only the vector with the densest neighborhood from all the vectors in its neighborhood is labeled as an inlier. On the other extreme, when the maximum is applied, only the vector with the biggest R_p from all the vectors within its neighborhood is labeled as an outlier. In other words, the vector is labeled as an outlier only when it is the most outlying vector within its neighborhood. The averages *nean* and *hean* in most cases generate similar results as the median, but in the case demonstrated by the vector $\mathbf{v}_N$ they generate the results similar to lower quantiles. It means that they are influenced by the presence of a small number of vectors with strongly higher density of their neighborhood, in the neighborhood of the examined vector. The average *nean* is influenced more.

V. Conclusion and Future Work

The original algorithms *meanQhean* (LOF) and *maxDhean* (LOF') are comparable, *maxDhean* is little bit faster and *meanQhean* has better results for the vectors on the border of clusters generated by the uniform distribution. We are convinced that LOF should be defined as *neanQnean* not only because it is geometrically much more elegant, but also because *neanQnean* increases the OF values for outliers and therefore highlights them, what is described as convenient in [12].

As demonstrated by the results of the experiments, the OF values are only very little influenced by the way how the R_p is calculated. Therefore we recommend that the researchers apply the individual quantiles of the set **D**, which is easier to calculate, according to whether they want to detect even denser regions smaller than k. The parameter k can be set relatively high, it means much more than generally recommended $k = 20$.

Especially, if we suppose that the dataset contains a lot of noise and relatively sparse clusters, it is essential to set the parameter k high and to apply a low quantile of the set **D** what cannot be replaced by the original LOF algorithm. The similar idea is proposed by the LOF" algorithm.

It is much more important how the R_{avg} is calculated. If a researcher wants to find only strong outliers with a low probability to label an inlier wrongly as an outlier, then it is important to compute R_{avg} as a high quantile of the set of all R_i in the neighborhood of the examined vector. In the extreme case, it is possible to apply max R_i.

If a researcher wants to be sure that only the vectors with considerably denser neighborhood will be labeled as inliers, or if a researcher wants to minimize the probability to label an outlier wrongly as an inlier, then it is important to compute R_{avg} as a low quantile of the set of all R_i in the neighborhood of the examined vector.

In general, the following algorithms are recommended: the algorithm $0.1D0.1$ with high parameter k for the detection of the centers of the clusters or in case of a dataset with

 ISBN: 978-1-61208-227-1

a lot of noise, the algorithm $maxD0.9$ for the detection of the most distant outliers or the clusters smaller than k in the dataset with relatively low portion of noise, and the algorithm $medDmed$ if a researcher wants to eliminate as many vectors as possible without the loss of the information.

In the future work, we would like to focus on preparing the datasets for the clustering, where we would like to use outlier factors as applicable weights for clustering algorithms. We would like to extend the method experimental evaluation using large real world datasets such as the one described in [13]. Thus, we would be able to evaluate impact of the outlier detection on robustness of AI algorithms used in the mobile robotics domain [14], [15], especially for an UAV [16].

ACKNOWLEDGMENTS.

This work was supported by the project IGA F4/6/2012.

REFERENCES

[1] M. Žambochová, "Typology of foreign students interested in studying at czech universities," *E+M Ekonomika a Management*, no. 2, pp. 141–154, 2012.

[2] A. Frolov, D. Husek, and P. Polyakov, "Recurrent-neural-network-based boolean factor analysis and its application to word clustering," *Neural Networks, IEEE Transactions on*, vol. 20, no. 7, pp. 1073–1086, 2009.

[3] K. Wegrzyn-Wolska, G. Dziczkowski, and L. Bougueroua, "Linking the drugs and pharmaceutical databases," in *Next Generation Web Services Practices, 2009. NWESP'09. Fifth International Conference on*. IEEE, 2009, pp. 3–8.

[4] E. Zimková and V. Úradníček, "Inflačné cielenie a možnosti predikovania inflácie v podmienkach slovenska," *Ekonomický časopis*, no. 06, p. 658, 2004.

[5] M. Breunig, H. Kriegel, R. Ng, J. Sander *et al.*, "Lof: identifying density-based local outliers," *Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.

[6] A. Chiu and A. Fu, "Enhancements on local outlier detection," in *Seventh International Database Engineering and Applications Symposium, 2003. Proceedings.* IEEE, 2003, pp. 298–307.

[7] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in *19th International Conference on Data Engineering, 2003. Proceedings.* IEEE, 2003, pp. 315–326.

[8] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the International Conference on Very Large Data Bases.* Citeseer, 1998, pp. 392–403.

[9] H. Cao, G. Si, W. Zhu, and Y. Zhang, "Enhancing effectiveness of density-based outlier mining," in *International Symposiums on Information Processing (ISIP), 2008*. IEEE, 2008, pp. 149–154.

[10] W. Jin, A. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," *Advances in Knowledge Discovery and Data Mining*, pp. 577–593, 2006.

[11] J. Tang, Z. Chen, A. Fu, and D. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," *Advances in Knowledge Discovery and Data Mining*, pp. 535–548, 2002.

[12] H. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[13] T. Vintr, L. Pastorek, and H. Rezankova, "Autonomous robot navigation based on clustering across images," *Research and Education in Robotics-EUROBOT 2011*, pp. 310–320, 2011.

[14] T. Krajník and L. Přeučil, "A Simple Visual Navigation System with Convergence Property," in *European Robotics Symposium 2008*. Heidelberg: Springer, 2008, pp. 283–292.

[15] T. Krajník, J. Faigl, M. Vonásek, V. Kulich, K. Košnar, and L. Přeučil, "Simple yet stable bearing-only navigation," *J. Field Robot.*, 2010.

[16] T. Krajník, M. Nitsche, S. Pedre, L. Přeučil, and M. Mejail, "A Simple Visual Navigation System for an UAV," in *International Multi-Conference on Systems, Signals and Devices*. Piscataway: IEEE, 2012, p. 34.

 ISBN: 978-1-61208-227-1

Data Mining Application for Anti-Crisis Management

Nafisa Yusupova
Department of Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
yussupova@ugatu.ac.ru

Gyuzel Shakhmametova
Department of Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
shakhgouzel@mail.ru

***Abstract*—The paper is devoted to data mining as applied to anti-crisis management, in particular to bankruptcy monitoring. The decision support system for bankruptcy monitoring, based on the intelligent information technologies (data mining, expert systems) is considered. The main stages of data mining technology applied to anti-crisis management are described in detail. The results of data mining implementation are estimated.**

Keywords-data mining; forecasting; decision support system; anti-crisis management; bankruptcy monitoring

I. INTRODUCTION

Anti-crisis management is a process of preventing or dealing with the crisis the enterprise (company) is in. This definition incorporates two components of the anti-crisis management: prevention of the crisis that has not yet come and overcoming of the crisis that has already come [1]. Interpretation of the anti-crisis management may be different depending on the state of the enterprise. In case the enterprise is in a stable position, the anti-crisis management consists in monitoring which aims at the data retrieval and processing and forecasting the enterprise performance. In case the enterprise is in an unstable position, i.e. there is a bankruptcy risk, the anti-crisis management assumes a form of regulation which is a set of measures to protect the enterprise in crisis situations and prevent bankruptcy. When an enterprise is in crisis it is necessary to resort directly to anti-crisis management as an instrument to pull the enterprise out of the crisis or to prepare it to be wound up or reorganized.

The carried-out analysis of methods of enterprises bankruptcy predicting and the analysis of possibilities of well known IT-decisions in this field showed that the development of a decision support system for bankruptcies monitoring is needed [7]. The data required for anti-crisis management are semi-structured data in the majority of cases and therefore the application of intelligent information technologies is necessary [4][6].

The researches into anti-crisis management in the field of bankruptcy monitoring have been carried on for a long time and can be found in the papers of many scientists, as well as in the IT-decisions. These problems are considered in detail in [7]. The distinguishing feature of this research is the possibility of fraudulent bankruptcy indications forecasting at its early stages when it is possible to take preventive measures.

There are two main approaches to enterprise bankruptcy forecasting in modern business and financial performance practice [1]. Quantitative methods are based on financial data and include the following coefficients: Altman Z-coefficient (the USA); Taffler coefficient (Great Britain); two-factor model (the USA); Beaver metrics system and the others. A qualitative approach to enterprise bankruptcy forecasting relies on the comparison of the financial data of the enterprise under review with the data of the bankrupt business (Argenti A-account, Scone method). Integrated points-based system used for the comprehensive evaluation of business solvency includes the characteristics of both quantitative and qualitative approaches. An apparent advantage of the methods consists in their system and complex approach to the forecasting of signs of crisis development, their weaknesses lie in the fact that the models are quite complicated in making decisions in case of a multi-criteria problem, it is also worth mentioning that the taken forecasting decision is more subjective. Modern information technologies, Data Mining in particular, provide ample opportunities for solving the problems of anti-crisis management [10].

Financial and economic application software that is available on the market nowadays is quite varied and heterogeneous. The necessity to develop such software products is dictated by the need of enterprises to promptly receive management data in due time and to forecast the signs of crisis development. To one extent or another, tools for anti-crisis management are available in a number of ready-made IT-decisions [7]. But the data analysis in many software products actually consists in providing the necessary strategic materials, while software products should meet the increasing needs such as analysis and forecasting enterprise financial performance in the next period of report.

The authors of the study aim to develop models and algorithms based on intelligent technologies for the detection of the crisis state of the enterprise while still in its early stages for the timely changes of the development strategy of the enterprise, which will increase stability and economic independence of the enterprise, as well as reduce the impact of the human (subjective) factor on making important management decisions. Data mining application in the decision support system for anti-crisis management is

 ISBN: 978-1-61208-227-1

discussed in this article on the example of monitoring bankruptcies.

The second section deals with the general scheme of the proposed decision support system. The main modules of this system are the data mining module and the expert system. The third section considers in detail the application of data mining technology to forecast the financial performance of the enterprise (company). Computer-based processing of the data on the basis of the proposed system is presented in the fourth section. Forecasting of the financial performance of the enterprise is made by the analytical platform. The stages of the implementation processes are also described. Section 5 of the article presents the results of the practical application of the system which allows assessing the effectiveness of the system.

II. DECISION SUPPORT SYSTEM FOR MONITORING BANKRUPTCY

The major aspect of the bankruptcy monitoring problem is the analysis and identification in good time of the signs of fraudulent bankruptcies [1].

The basis of the whole complex of techniques for the decision support system (DSS) is legally approved methodical instructions on accounting and analysis of enterprise' financial position and solvency so as to group the enterprises depending on the level of risk of bankruptcy, as well as techniques for the identification of the signs of fictitious and deliberate bankruptcy. These techniques are currently used by auditors and arbitration managers.

To develop the decision support system for monitoring bankruptcy the authors propose the following general scheme of DSS (Figure 1) and used knowledge engineering, expert system (ES) technology [4] and data mining (DM) technology [2][3].

The expert system technology underlies two modules of DSS in bankruptcy monitoring [8]:

- module for grouping companies depending on the level of risk of bankruptcy (module1);
- module for the identification of the signs of illegal bankruptcy (module 2).

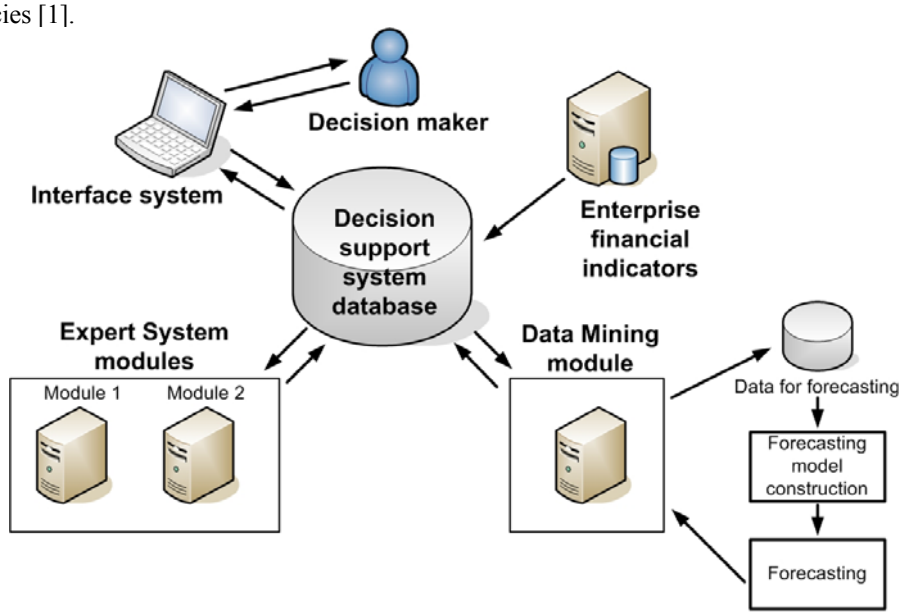


Figure 1. The general scheme of the DSS in monitoring bankruptcies

Another module of the DSS is a data mining module. This module helps to solve problems which include cleaning the data for qualitative forecasting and predicting financial indicators of the enterprise with the use of several prognostic model-building mechanisms, including self-teaching algorithms. The objective of this module is to identify negative trends in changing financial indicators as well as possible signs of fraudulent bankruptcy based on the comparative analysis of the current financial indicators and financial indicators forecasted by the data mining module.

Primary, intermediate and resulting data are stored in the main decision support database organized according to the relational model. To keep the decision support system operating the primary data on the company is imported in the system either automatically or manually. Interaction between the DSS and the user is carried out by means of an interface subsystem.

In the first phase of the DSS the enterprise is classified according to the degree of the threat of bankruptcy by means of module 1 of the expert system (Figure 2).

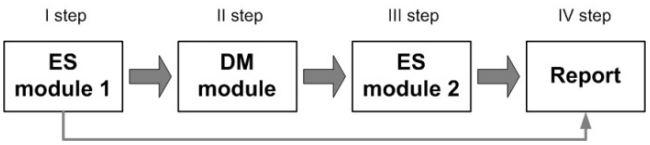


Figure 2. The steps of the DSS modules using

 ISBN: 978-1-61208-227-1

Depending on the results, the enterprise is either checked for signs of fraudulent bankruptcy (I step, module 1 of the expert system), or financial performance is forecasted using the data mining technology (II step, DM module). In the third phase on the basis of the forecasted values the signs of the deliberate bankruptcy are identified (III step, module 2 of expert system). On the IV step a report is made for the decision maker.

III. DATA MINING MODULE

The problem which is solved by the data mining module in DSS in monitoring enterprise bankruptcy is the problem of forecasting financial indicators of the enterprise (company). This problem can be seen as a problem of forecasting the time series, as the data for the prediction of financial indicators are presented in the form of measurement sequences, collated at non-random moments of time. In contrast to the analysis of random sampling the analysis of time series is based on the assumption that successive values are observed in equal periods of time. Like many other kinds of analysis the analysis of time series implies that the data contain a systematic component (generally including several components) and a random noise (error) which makes it hard to detect regular components.

A time series may be presented as decomposition of four constituents [11]:

$$X_t = f(S_t, T_t, C_t, R_t)$$

where S_t – seasonality effect; T_t – trend, or systematic movement; C_t – fluctuations around the trend with more or less regularity (cyclicality); R_t – random (unsystematic) residual component. The action of these constituents may be interdependent. The models in which the time series is presented as a sum of the given components is called additive if multiplicative models are in the form of the product of numbers. The additive model takes the form: $X_t=S_t+T_t+C_t+R_t$, the multiplicative one: $X_t=S_t*T_t*C_t*R_t$. There exist also mixed models. The main task in investigating the time series consists in detecting and determining the quantification of each component (S_t, T_t, C_t, R_t) for forecasting the future values of the series.

The dynamics of lots of financial and economic indicators has a stable fluctuation constituent. In order to obtain accurate predictive estimates it is necessary to represent correctly not only the trend but the seasonal components as well. The use of data mining methods in time series forecasting makes the solution of the given task possible. These methods have a number of benefits:

- possibility to process large volumes of data;
- possibility to discover hidden patterns;
- use of neural networks in forecasting allows obtaining the result of the required accuracy without determining the precise mathematical dependence.

There are a lot of other benefits of data mining such as basic data pre-processing, their storage and transformation, batch processing, importing and exporting of large volumes of data, availability of data pre-processing units as well as ample opportunities for data analysis and forecasting.

The algorithm for forecasting the companies' financial indicators has been developed. It works as follows (Figure 3). Let us assume that as a result of transformation by the "sliding window" method we obtain a sequence of time counts [6]**:**

$$X_{-n},\ldots..,X_{-2},X_{-1},X$$

where X is a current value. Forecasting for X_{+1} is carried out on the basis of the built model. In order to forecast the value of X_{+2} it is necessary to shift the whole of the sequence one count to the left so that the forecast of X_{+1} carried out earlier could be included in the initial values. Then once again the algorithm for computing the predicted value will be started. X_{+2} is calculated with regard for X_{+1} and further in a similar way, according to the defined forecasting horizon. To debug the prediction algorithm it is necessary to define the forecasting horizon as well as the table fields which must be filled in to carry out a forecast (to calculate the output field of the model).

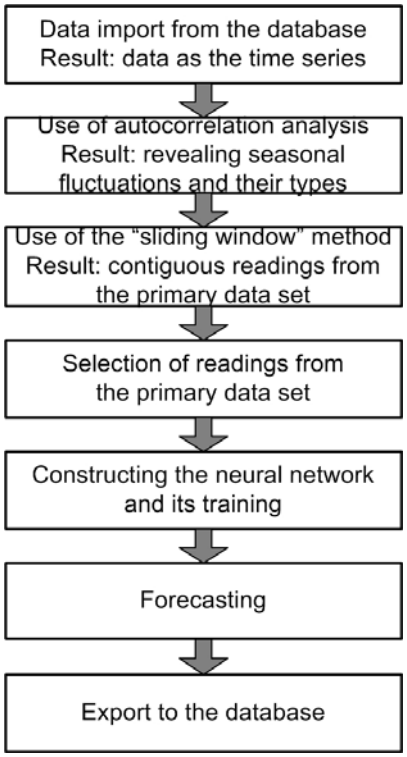


Figure 3. Main stages of the data mining module

Forecasting of enterprise financial indicators in the DSS can be performed by means of a number of DM techniques such as partial and complex data preprocessing, autocorrelation analysis, the method of "sliding window" and neural networks.

In solving the problem of forecasting the time series with the aid of a neural net it is required to input the values of several adjacent counts from the initial set of data into the analyzer. This method of data sampling is called "sliding

 ISBN: 978-1-61208-227-1

window" (window – because only a certain area of data is highlighted, sliding – because this window "moves" across the whole data set). The efficiency of implementation considerably increases, if we do not sample the data out of a number of consecutive records, but successively locate the data related to the specific position of the window in one record. The values in one of the writing fields will be related to the current count and in other ones they will be shifted from the current count to the "future" or the "past". Thus, transformation of the sliding window has two parameters: "depth of plunging" – the number of the "past" counts in the window and "forecasting horizon" – the number of "future" counts. It should be mentioned that for the boundary positions of the window (relative to the beginning and the end of the whole sampling) incomplete records will be formed, i.e. records containing empty values for the missing past and future counts. The transformation algorithm allows either excluding such records from the sampling (in that case for several boundary counts there will be no records) or including them (in the latter case records will be made for all the counts available, but some of them will be incomplete).

With the use of the neural network the forecasting problem can be set in the following way: to find the best approach to the function defined by the final set of input values (teaching examples). In our case the neural networks help to solve the problem of recovery of the missing values as well as the forecasting of financial indicators of the enterprise being analyzed.

IV. SOFTWARE IMPLEMENTATION OF THE DATA MINING MODULE

The software implementation of the data mining module to forecast the financial indicators of the enterprise is performed by means of the analytical platform [9]. As it was mentioned in Section 3, the data mining module is realized by the following main steps:

1) primary data input (Figure 4);
2) using of "sliding window" (Figure 6);
3) neural network programming – constructing and teaching (Figures 7, 8);
4) forecasting (Figure 9).

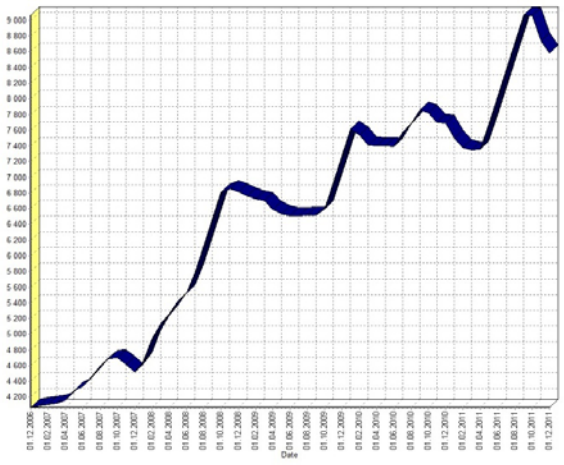


Figure 4. Primary data input (graphic representation)

Let us consider the whole of the data mining process on the example of the enterprise's financial indicator "fixed assets".

After the input of the initial information preprocessing of the raw data is required. Figure 5 shows a possible sequence of steps in preprocessing the raw data before using data mining models.

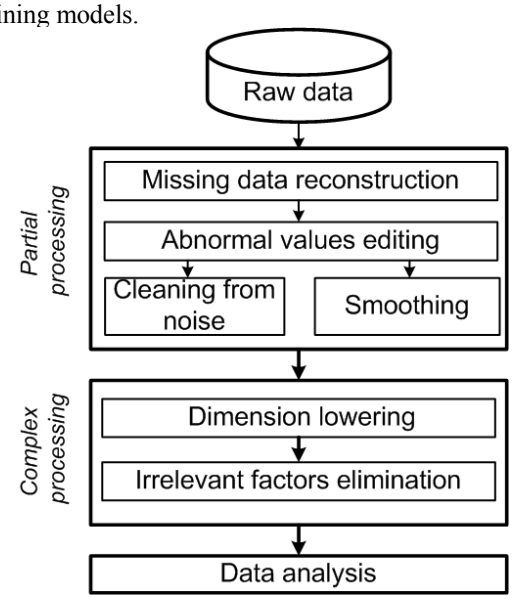


Figure 5. Data preprocessing

Partial and complex processing are distinguished in preprocessing. In partial processing the missing data are restored. Abnormal values are edited, noise is subtracted, and smoothing is carried out. For these purposes correlation analysis, factor analysis, main component method, regression analysis and other methods are used.

In complex processing the lowering of the dimension of the input data and/or elimination of irrelevant factors take place. Robust filtering, spectral and wavelet analysis are used. In practice, partial and complex preprocessing of the raw data can be performed in any sequence with any parameters at each step, any number of times, that is, the preprocessing script can be quite complex [5].

Analysis of the primary ("raw") data is made step-by-step. Cleaning, transformation and forecasting of the data is done individually with each time series of the enterprise's financial indicators. In forecasting the time series by means of the neural networks it is required to input the values of several adjacent counts from the source data set – "sliding window" (Figure 6).

The efficiency of implementation considerably increases, if we do not sample the data every time from a number of consecutive records, but successively locate the data related to the specific position of the window in one record. The experiments show that for "sliding window" the optimal value of the depth of plunging is 5, because 5 inputs is enough for neural network to forecast the fixed assets.

 ISBN: 978-1-61208-227-1

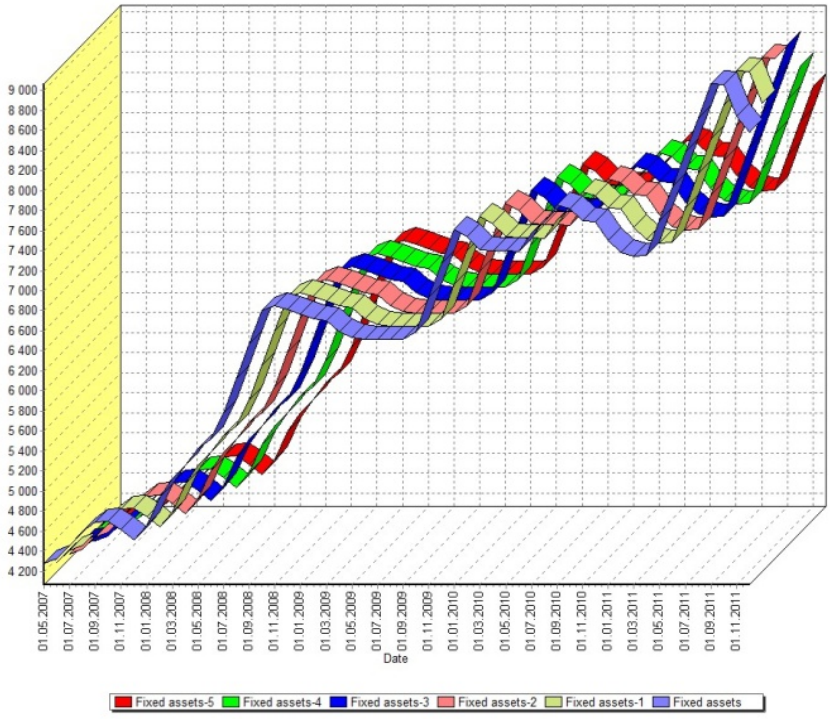


Figure 6. "Sliding window" for fixed assets

The neural network structure for forecasting the enterprise's indicator "fixed assets" has the form 5-4-3-2-1. The graph of the neural network and dispersing diagram for the neural network quality estimation are shown in Figures 7 and 8.

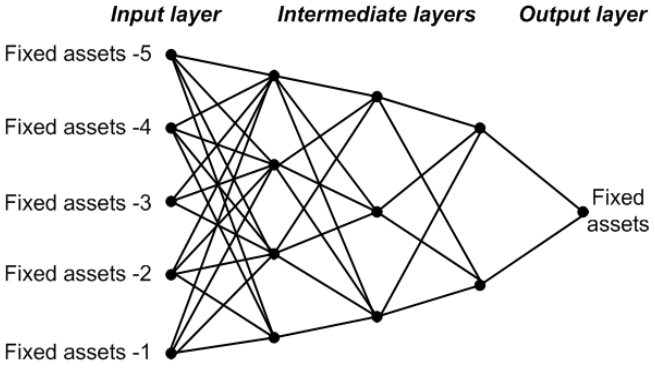


Figure 7. The neural network structure for indicator "fixed assets"

The dispersing diagram allows estimating the quality of the neural network constructing and teaching – the points are closer to the central axis the accuracy of the neural network model is higher.

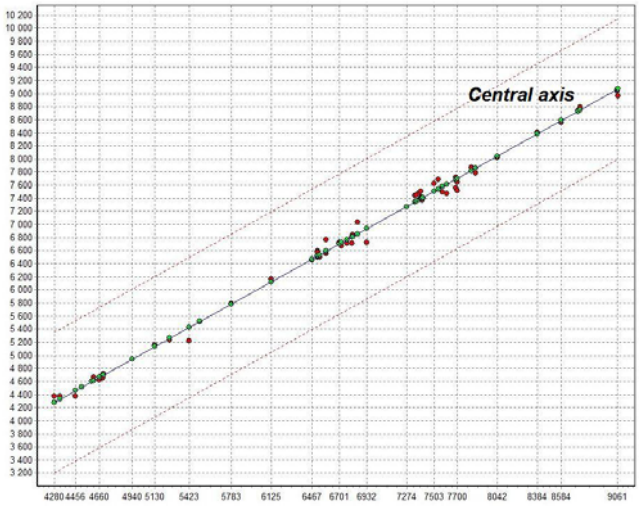


Figure 8. The dispersing diagram

The neural network performance may be evaluated also by the errors diagram (Figure 9), that shows the neural network error for each of the indices and the average error (in percentage).

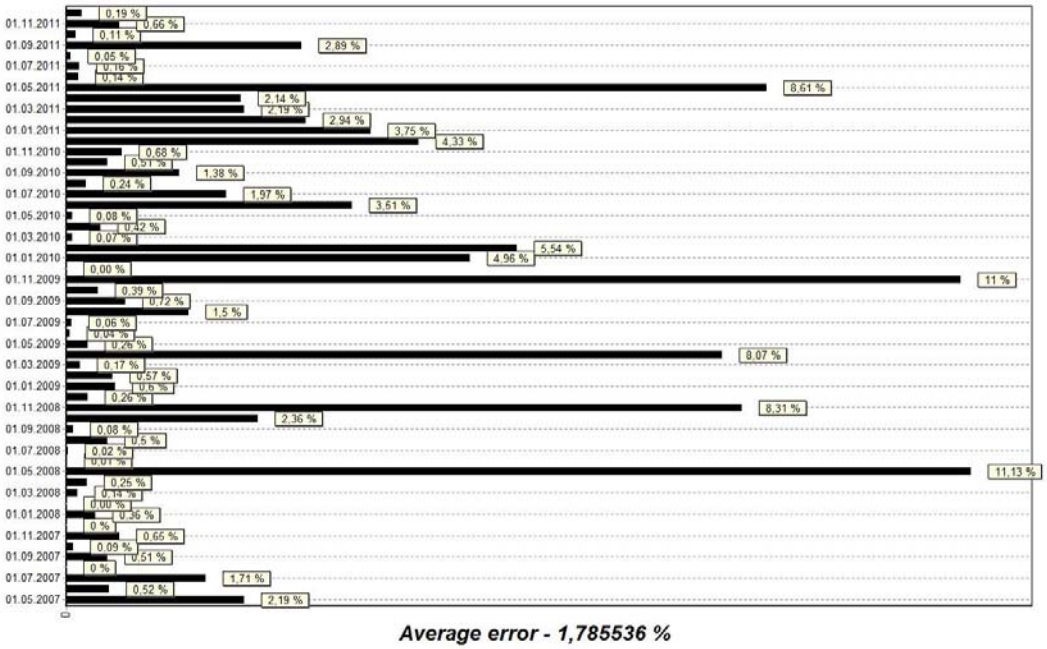


Figure 9. The errors of the neural network performance for indicator "fixed assets"(graphic representation)

The final stage in the program implementation of data mining is forecasting (Figure 10).

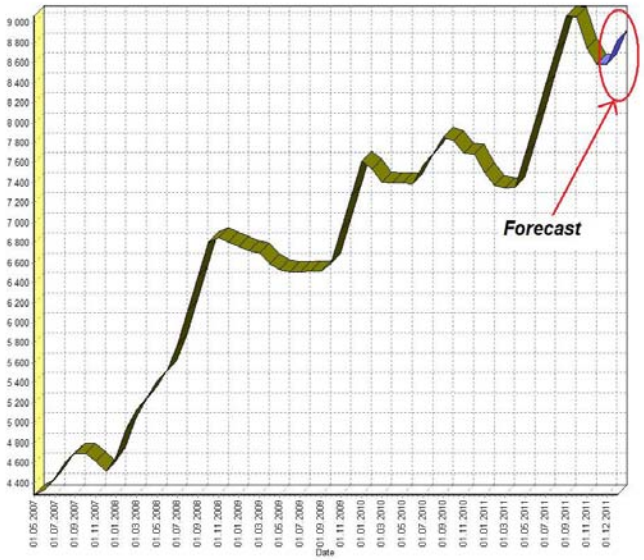


Figure 10. Results of forecasting of the fixed assets (graphic representation)

Each of the financial indicators has its own prediction algorithm that includes the size of the step of the sliding window, neural network structure, the form of the activation function and its value (Table I). These parameters are defined for each enterprise individually.

TABLE I. ALGORITHMS OF DATA MINING APPLICATION TO FORECAST ENTERPRISE'S FINANCIAL INDICATORS

Financial indicators of the enterprise	Depth of plunging of "sliding window"	Form of the activation function	Value of the activation function	Neural network structure
Fictitious assets, i.e. patents, licenses, trade marks	5	Sigma form	0.80	5-3-1
Fixed assets	***5***	***Sigma form***	***1.30***	***5-4-3-2-1***
Long-term financial investments	3	Sigma form	0.95	3-2-1
Total of non-working assets	5	Hyper tangent form	1.05	5-3-2-1
…	…	…	…	…
Reserves of forth-coming expenses and payments	5	Arctangent form	1.25	5-2-1
Total of short-term liabilities	4	Sigma form	1.10	4-2-1
Liabilities balance	5	Hyper tangent form	0.85	5-2-1

 ISBN: 978-1-61208-227-1

V. DATA MINING IMPLEMENTATION EFFICIENCY ANALYSIS

The analysis of the effectiveness of the data mining module is based on the comparative analysis of the financial indicators for the same period of time, obtained directly from the enterprise and forecasted through data mining.

This approach has been used in state monitoring of a number of industrial and agro-industrial enterprises of Republic Bashkortostan (Russia).

The fragment of the analysis of the effectiveness of data mining with deviation of the forecasted values of the financial indicators from the actual data is presented in Table II.

TABLE II. DEVIATIONS OF FORECASTED VALUES FROM ACTUAL DATA, IN PERCENTAGE

Financial indicators of the enterprises	**Enterprise 1**	**Enterprise 2**	**Enterprise 3**	**Enterprise 4**
Fixed assets	3.57	4.28	4.86	5.05
Long-term investments	4.62	1.45	5.61	6.63
Total non-current assets	4.89	4.22	3.77	3.25
Total current assets	6.32	4.65	5.90	7.56
All long-term liabilities	5.41	8.42	4.27	7.34
Loans and credits	5.75	2.05	7.82	8.74
Creditor indebtedness	2.56	3.41	7.68	1.35
…	…	…	…	…
Tax liabilities	7.32	1.91	6.82	5.89
Result of short-term liabilities	6.83	5.88	6.24	8.67
Balance of liabilities	2.42	4.47	8.45	3.57
Revenue for the period under review	5.87	3.55	5.62	1.52

Analysis of the effectiveness of data mining for values forecasting showed that the deviations of the forecasted values from the real data are in the range from 1.35% to 8.74%. The average deviation is about 6.5 %, which is quite a good result for forecasting.

CONCLUSION AND FUTURE WORK

The data mining application in anti-crisis management for bankruptcy monitoring is investigated. The decision support system for bankruptcy monitoring including the data mining module is developed. The decision maker using the DSS may be the top manager or supervisory authority. It is possible for users of the system to monitor the major trends in the economic processes of the enterprise. With the help of data mining means, neural networks in particular, the enterprise financial indicators can be forecasted for the definite period of time (for example for 3 months). The aim of the neural network at this stage is to catch the regularities of the financial indicators changes and detect them. Then with the help of the expert system the enterprise is classified on the basis of the forecasted indicators according to the degree of the bankruptcy threat. In other words the condition of the enterprise is defined not for present moment but for the definite time period (for example for 3 months). It gives an opportunity to take measures preventing the enterprise from fraudulent bankruptcy. The description of financial indicators forecasting based on data mining technology is given. For each of the forecasted financial indicators the separate forecasting algorithm has been developed. The efficiency analysis reveals good results of data mining implementation for forecasting financial indicators in the decision support system for bankruptcy monitoring. These results confirm the possibility of using data mining technology in the developed decision support system for bankruptcy monitoring. Future work will be connected with the increasing accuracy of financial performance forecasts.

ACKNOWLEDGMENT

This research has been supported by grants № 11-07-00687-a, № 12-07-00377-a of the Russian Foundation for Basic Research and grant “The development of tools to support decision making in different types of management activities in industry with semi-structured data based on the technology of distributed artificial intelligence” of the Ministry of Education of the Russian Federation.

REFERENCES

[1] A. Belyaev and E. Korotkov, Anti-Crisis Management. UNITI-DANA, Moscow, 2011.

[2] A. Syväjärvi and J. Stenvall, Data Mining in Public and Private Sectors: Organizational and Government Application. Information Science Reference, Hershey, New York, 2010.

[3] A.V. Senthil Kumar, Knowledge Discovery Practices and Emerging Application of Data Mining: Trends and New Domains. Information Science Reference, Hershey, New York, 2011.

[4] Peter Jackson. Introduction to Expert Systems. Williams, Moscow, 2001.

[5] A. Barsegyan, M. Kupriyanov, V. Stepanenko, and I. Kholod, Technologies of Data Analysis: Data Mining, Visual Mining, Text Mining, OLAP. Sec. ed. BHV-Petersburg, Saint-Petersburg, 2007.

[6] V. Duk and A. Samoilenko, Data Mining. Peter, Saint-Petersburg, 2001.

[7] N. Yusupova and G. Shakhmametova, Intelligent Information Technologies in the Decision Support System for Enterprises Bankruptcy Monitoring. UNC RAN, Ufa, 2010.

[8] G. Shakhmametova, D. Amineva, and V. Dolzhenko, “Expert System for Decision Support in Anti-Crisis Monitoring,” Proc. of the 13-th International Workshop on Computer Science and Information Technologies, Germany, Garmisch-Partenkirchen, 2011. Vol. 1. pp. 151-155.

[9] http://www.basegroup.ru 20.06.2012.

[10] http://EconPapers.repec.org/RePEc:ovi:oviste:v:xi:y:2011:i:9:p:233-236, 20.08.2012.

[11] V. Afanasyev, The Time Series Analysis and Forecasting, Finance and Statistics, Moscow, 2001.

 ISBN: 978-1-61208-227-1

Oracle NoSQL Database – Scalable, Transactional Key-value Store

Ashok Joshi, Sam Haradhvala, Charles Lamb
Oracle: Database Development
Burlington, MA 01803
ashok.joshi@oracle.com, sam.haradhvala@oracle.com, charles.lamb@oracle.com

Abstract **- Oracle NoSQL Database is a highly scalable, highly available key-value database, designed to address the low latency data access needs of the "big data" space. Among its unique features, Oracle NoSQL Database provides major and minor keys, flexible durability and consistency policies, and integration with both MapReduce infrastructure and conventional relational data. Major and minor keys enable transactional guarantees across multiple data records. Flexible durability and consistency polices allow applications to trade off durability, consistency, availability, and performance on a per-operation basis. Integration with alternate data processing and management systems provides a cohesive environment for accommodating today's big data management needs.**

Keywords-NoSQL; Big Data; Database systems.

I. INTRODUCTION

Simply put, big data is an informal term that encompasses all sorts of data such as web logs, sensor data, tweets, blogs, user reviews, SMS messages etc. It is characterized by high volume (hundreds of terabytes or more), high variety (e.g., no inherent structure, one row "looks" very different from another) and high velocity (hundreds of thousands of operations per second). It is possible to derive valuable information (e.g., sentiment analysis) by aggregating big data in some way, though, quite often, an individual data row may or may not provide much value or insight. The conventional wisdom is that it is cost-prohibitive to manage and process big data using only traditional relational database technologies.

Companies such as Google [1], [2], Amazon [3], LinkedIn [4] and Facebook [5] have demonstrated that there is significant business benefit in harnessing big data. They have also made major contributions to the database community in terms of algorithms and frameworks for massive-scale distributed processing using commodity processors.

This has resulted in a huge surge of interest in big data in the information technology industry and commercial community. The allure of increasing profitability, reducing costs, being better "connected" with customers and improving the business is too hard to ignore. However, for a variety of reasons described below, only a small number of organizations have been able to successfully leverage the business value of big data.

The pioneers of big data processing employ armies of super-smart developers to work on big data problems. Commercial organizations have neither the budget nor the ability to invest significantly in "deep" software development projects. Big data are inherently unstructured or semi-structured; this makes it difficult to process big data. Finding the "nuggets of gold" in the vast amounts of unstructured data requires specialized technologies and expertise. Finally, in order to obtain the maximum benefit, big data need to be combined with traditional structured data (SQL data repositories).

In October 2011, Oracle announced a suite of products and technologies aimed at providing a complete and comprehensive solution to address the big data needs of the market. A key piece of this technology stack is Oracle's entrant into the NoSQL space, the Oracle NoSQL Database.

We begin with an overview of Oracle NoSQL Database. This is followed by a description of important features including major and minor keys, sharding and replication, and consistency and durability policies that can be selected on a per-operation basis. The next section discusses the role and benefits of interactive big data processing. We present some performance results that clearly demonstrate the excellent throughput, response time and scalability of Oracle NoSQL Database. Finally, we conclude the paper with our view of the big data processing landscape and a short description of the comprehensive hardware and software technologies and solutions from Oracle, designed and optimized to address the big data processing needs of the market.

II. ORACLE NOSQL DATABASE

Big data processing falls into two major categories – interactive data management and batch processing. The Oracle NoSQL Database [6] is a scalable, highly available, key-value store that can be used to acquire and manage vast amounts of interactive information.

Oracle NoSQL Database uses Oracle Berkeley DB Java Edition [7] as the underlying data storage engine, thus leveraging all the performance and availability benefits that it has to offer. Berkeley DB Java Edition is a mature product that also provides many of the features and characteristics that are necessary for a building a distributed key-value store such as Oracle NoSQL Database.

Figure 1 illustrates the architecture of an Oracle NoSQL Database deployment; in this example, there are two client nodes and multiple server nodes for managing key-value data. The system is designed to handle large numbers of client nodes as well as servers.

 ISBN: 978-1-61208-227-1

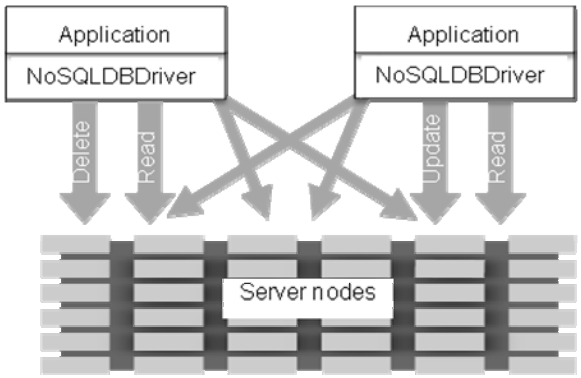


Figure 1: Oracle NoSQL Database architecture.

A. *Major Keys, Minor keys and values*

Oracle NoSQL Database provides a key-value paradigm to the application developer. Every entity (record) is a set of key-value pairs. A key has multiple components, specified as an ordered list. The major key identifies the entity and consists of the leading components of the key. The subsequent components are called minor keys. This organization is similar to a directory path specification in a file system (e.g., /Major/minor1/minor2/). The "value" part of the key-value pair is simply an uninterpreted string of bytes of arbitrary length.

This concept is best explained using an example. Consider storing information about a person, John Smith, who works at Oracle Headquarters, start date Jan 1, 2012 and has a telephone number +1650-555-9999. The user's Id might be a logical choice for major key for the person entity (for example, 123456789). Further, the 'person' entity might contain personal information (such as the person's telephone number) and employment information (such as work location and hire date). The application designer can associate a minor key (e.g. personal_info) with the personal information (+1-650-555-9999) and another minor key (e.g. employment_info) with the employment information (Oracle Headquarters, start date Jan 1, 2012).

Specifying the major key "123456789" would return "John Smith". Specifying "/123456789/personal_info" as the key would access John Smith's personal information; similarly, "/123456789/employment_info" would be the key to access the employment information. Leading components of the key are always required. NoSQL Database internally stores these as three separate key-value pairs; one for the user_id, a second for user_id/personal_info and the third for user_id/employment_info.

The API for manipulating key-value pairs is simple. The user can insert a single key-value pair into the database using a put operation. Given a key, the user can retrieve the key-value pair using a get operation or delete it using a delete operation. The get, put, and delete operations operate on only a single (multi-component) key. NoSQL Database provides additional APIs that allow the application to operate on multiple key-value pairs within an entity (same major key) in a single transaction.

The major key determines which shard the record will belong to. All key-value pairs associated with the same entity (same major key) are always stored on the same shard. This implementation enables efficient, "single shard" access to logically related subsets of the record. Figure 2 illustrates the concept of major and minor keys.

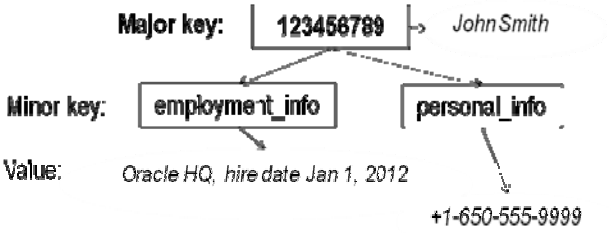


Figure 2: Major and minor keys.

Oracle NoSQL Database also provides an unordered scan API that can be used to iterate through all the records in the database; unordered scans do not have transaction semantics, though only committed data will be returned to the application.

B. *Shards and Replicas*

Oracle NoSQL Database is a client-server, sharded, shared-nothing system. The data in each shard are replicated on each of the nodes which comprise the shard. As discussed earlier, Oracle NoSQL Database provides a simple key-value paradigm to the application developer. The major key for a record is hashed to identify the shard that the record belongs to. Each key-value pair is always stored and managed within a single shard. Oracle NoSQL Database is designed to support changing the number of shards dynamically in response to availability of additional hardware. If the number of shards changes, key-value pairs are redistributed across the new set of shards dynamically, without requiring a system shutdown and restart.

Each shard is highly available. A shard is made up of a single master node which can serve read and write requests, and several replicas (usually two or more) which can serve read requests. Replicas are kept up to date using streaming replication. Each change on the master node is committed locally to disk and also propagated to the replicas. If the master node should fail, one of the surviving replicas is automatically elected as a master and processing continues uninterrupted. As soon as the failed node is repaired, it rejoins the shard, is brought up to date and then becomes available for processing read requests. Thus, the Oracle NoSQL Database server can tolerate failures of nodes within a shard and also multiple failures of nodes in distinct shards. By proper placement of masters and replicas on server hardware (racks and interconnect switches), Oracle NoSQL Database achieves very high levels of availability on commodity servers.

C. *Consistency and Durability*

Distributed systems need to address the notion of consistency, since there is a lag between making a change on one node and propagating the same change to another replica. On the other hand, distributed systems can take

 ISBN: 978-1-61208-227-1

advantage of the multiple copies of data to alleviate the "commit to disk" bottleneck. In particular, these systems can consider a transaction as being committed after receiving acknowledgements for the changed record from the replicas, without waiting for the disk I/O to complete. NoSQL Database supports the notions of variable consistency for read operations and varying degrees of durability for update operations. Further, NoSQL Database exposes these options at the API level so that the application designer can make the appropriate tradeoffs between performance and consistency/durability on a per operation basis. Many other systems with a similar architecture provide only a coarse level of control over consistency and durability (e.g., system-wide choice configured at system start-up); per-operation consistency and durability enables a broad class of applications by giving developers more control over the data.

There are several choices for read consistency. The application can specify absolute consistency if it needs the most recent version or can also specify time or LSN-based (log sequence number) consistency for read operations. Note that LSNs are not visible to the application directly; rather, they manifest themselves as point-in-time version handles. For example, an application might be willing to tolerate reading data that is no more than one second out-of-date. LSN-based consistency is useful in scenarios where the application modifies a record at a certain LSN x and wants to ensure that a subsequent read operation will read a version of that same record that is at least as current as the change identified by LSN x (it is okay to read a more recent version). Finally, the application can also specify that it doesn't care how consistent the data are, for a particular read request. Oracle NoSQL Database routes the request to the appropriate node (master or one of the replicas) in the shard based on the desired consistency.

In a shared everything architecture [8], making a transaction durable requires that the data management system commit the changes to disk (log) before acknowledging the completion of the transaction. In a distributed, master-replica architecture such as Oracle NoSQL Database, the transaction must be made durable on the master and also propagated to the replicas. This presents some interesting opportunities and tradeoffs. For example, the master node may choose to issue a lazy log write and concurrently send commit messages to the replicas. This strategy makes the transaction durable by "committing to the network", which is desirable if network latency is lower than disk latency. The transaction can be considered as committed if one or more replicas have received the changes associated with a transaction. The lowest latency choice is to issue a lazy log write at the master, concurrently send non-blocking commit messages to the replicas, and acknowledge the completion of the transaction. For additional assurance of durability, the system may choose to wait for acknowledgement messages from the replicas. Several other variants of this strategy are possible.

Transaction durability is thus determined by a combination of log write at the master node, log writes at the replicas, sending transaction commit messages to the replicas and receiving commit message acknowledgements from the replicas. Further, the system can decide whether to wait for acknowledgements from a majority of the replicas or all replicas. Of course, each legitimate combination of these options also influences performance and availability. For example, "write to local disk, write to replica disk, wait for acknowledgements from every replica" provides the highest level of durability but is also expensive in terms of latency and throughput.

Figure 3 illustrates the durability and consistency options that are available in Oracle NoSQL Database. NoSQL Database allows the user to choose the durability policy on a per operation basis. NoSQL Database uses this information during commit processing in order to achieve the best performance while honoring the durability requirements of the operation.

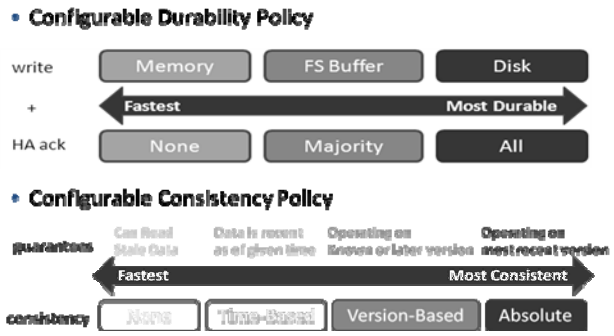


Figure 3: Durability and Consistency.

D. *Interactive Big Data Processing*

Oracle NoSQL Database has been designed for applications that need fast, predictable, low latency access to vast amounts of data. Let us examine how Oracle NoSQL Database benefits such applications by considering a typical E-commerce environment. Such systems manage vast numbers of user profiles and have stringent response time requirements. Whenever a user visits the site, the retailer provides a personalized experience based on the user's profile. If no such profile exists, the site must create one. These user profiles will change over time as the retailer learns more about the users through interactions. Different user profiles may contain radically different information and the retailer may decide to collect new information at any time. Oracle NoSQL Database addresses this use case by virtue of its flexible key-value paradigm and scales to meet increasing customer demand. Figure 4 illustrates some performance data [9] for Oracle NoSQL Database using Yahoo Cloud Serving Benchmark [10]. The graph on the left illustrates scalable write performance while the graph on the right illustrates scalable performance for a mixed read and update workload.

 ISBN: 978-1-61208-227-1

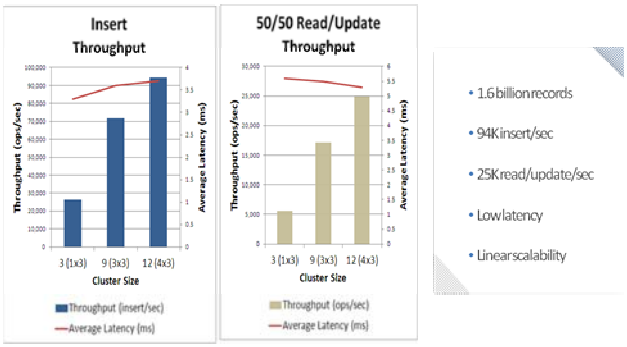


Figure 4: Performance on YCSB.

E. Big Data: The Big Picture

Oracle recognized early on that in order to derive maximum benefit from big data, it is necessary to combine and process unstructured and semi-structured content together with structured data. Business-critical information is primarily stored in relational database repositories. This information can be augmented with information from unstructured content in order to obtain business insights that can be gained only by combining structured and unstructured data. For example, sales forecast information is typically stored in relational repositories. By combining sales forecast information with big data content such as political trends, weather predictions, etc., it is possible to improve the accuracy of the sales forecasts.

Relational database systems have sophisticated algorithms for data warehousing, analysis and reporting. These algorithms typically work well with structured (row and column) data. There is also a wide array of products, tools and services available to analyze relational data, generate business intelligence and drive decisions. These tools and processes form the cornerstone of business data processing and analysis today.

In comparison, the tools available for processing big data (unstructured data) are relatively scarce and immature. However, it is possible to transform unstructured content into structured (row and column) format so that it becomes available for processing with data warehousing and business intelligence technologies. During this transformation process, one can also aggregate and cleanse the data to reduce the volume and increase information density of the content.

Oracle recently introduced a suite of complementary technologies to manage and process big data and also combine big data with traditional data warehousing and data analysis technologies for maximum business benefit [11]. Oracle NoSQL Database provides the interactive big data management component of this integrated solution. Oracle has adopted Cloudera's distribution of Apache Hadoop [12] in order to provide MapReduce capabilities and the open source distribution of R [13] for advanced analytics. Oracle Big Data Connectors, a separately licensed product, provides a high-performance Hadoop to Oracle Database integration solution. Oracle database and Oracle Business Intelligence tools provide the data warehousing, mining and analysis capabilities.

Oracle's approach to big data is unique for three reasons: it leverages the existing investments in data management and processing, seamlessly brings the benefits of big data to the enterprise, and finally, provides a commercial-grade, comprehensive solution to process and leverage all the data in the enterprise.

Oracle has gone a step further and also delivered the Big Data Appliance [14] that delivers software and hardware packaged together into an optimized platform that simplifies the management, analysis and mining of all the data in an enterprise.

ACKNOWLEDGMENT

Margo Seltzer, Alan Bram, Dave Segleau and Marie-Anne Neimat provided valuable feedback on earlier drafts of this paper.

We are very grateful to the Oracle Labs team and Jeffrey Alexander, in particular, for their help and advice on the architecture and design of Oracle NoSQL Database.

We are very grateful to Cisco and Raghu Nambiar in particular, for partnering with us to run the benchmarks on Cisco UCS.

REFERENCES

[1] http://research.google.com/archive/bigtable.html [retrieved: Oct, 2012]

[2] http://research.google.com/archive/mapreduce.html [retrieved: Oct, 2012]

[3] http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf [retrieved: Oct, 2012]

[4] http://project-voldemort.com/ [retrieved: Oct, 2012]

[5] http://en.wikipedia.org/wiki/Apache_Cassandra [retrieved: Oct, 2012]

[6]Oracle NoSQL Database Documentation: http://www.oracle.com/technetwork/products/nosqldb/overview/index.html [retrieved: Oct, 2012]

[7] Oracle Berkeley DB Java Edition documentation: http://www.oracle.com/technetwork/products/berkeleydb/documentation/index.html [retrieved: Oct, 2012]

[8] Anupam Bhide: An Analysis of Three Transaction Processing Architectures. VLDB 1988: 339-350

[9] Ashok Joshi, Raghunath Nambiar: Cisco UCS Ecosystem for Oracle: Extend Support to Big Data and Oracle NoSQL Database: http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/le_34301_wp.PDF [retrieved: Oct, 2012]

[10] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan and Russell Sears. http://research.yahoo.com/node/3202 [retrieved: Oct, 2012]

[11] http://www.oracle.com/us/technologies/big-data/index.html [retrieved: Oct, 2012]

[12] http://www.cloudera.com/ [retrieved: Oct, 2012]

[13] http://www.r-project.org/ [retrieved: Oct, 2012]

[14] http://www.oracle.com/us/products/database/big-data-appliance/overview/index.htm [retrieved: Oct, 2012

Efficient Extraction of Motion Flow Data From a Repository of Three-Dimensional Trajectories Using Bi-Dimensional Indexes

Antonio d'Acierno*, Marco Leone†, Alessia Saggese† and Mario Vento†
*Institute of Food Science, National Research Council, Avellino, Italy
† Department of Electronic and Information Engineering (DIEII), University of Salerno, Italy
Email: dacierno.a@isa.cnr.it, {mleone,asaggese,mvento}@unisa.it

***Abstract*—Motivated by the growing presence of acquisition peripherals throughout the world, we propose a novel method for storing and querying moving objects' trajectories extracted from surveillance cameras. Once moving objects have been detected and tracked, we suggest to store and index the related spatio-temporal data by using an innovative scheme based on widely available bidimensional indexes; moreover, a segmentation stage is performed to increase the overall efficiency. Thus, starting from the limitations of most of the clustering and similarity-based approaches, which restrict the choice of the query parameters, we present a trajectory storing system which efficiently supports Dynamic Spatio-Temporal (DST) queries, which are unrestricted time interval queries over moving objects. For the statistical description of the motion flow in the scene, we use a novel query typology, namely the Flow-DST, that is formulated as a sequence of DST. The experimental results, conducted over real and synthetic data, show the efficiency of the approach.**

Keywords-information retrieval; spatio-temporal queries; indexing; segmentation.

I. INTRODUCTION

The presence of monitoring cameras in public and private areas has grown significantly in the last decades. In fact, besides the increasing need for security, more and more public exercises are also interested in using the information extracted from these video sequences for commercial purposes. Think, as an example, to a hypermarket which would like to improve the marketing posters arrangement on the basis of the customers' preferences: a system able to analyze the customers' trajectories and infer their commercial preferences would serve the purpose.

This topic has been recently addressed by a lot of authors. Some examples of real applications able to analyze moving objects' trajectories for commercial purposes are proposed, for instance, in [1] and [2]: in the former, a real hypermarket case study is presented which investigates the relation between daily necessity products and higher flow pattern; in the latter, the authors use laser range finder and cameras to analyze pedestrian behavior in a large shopping mall.

Similar systems, able to perform the analysis of moving objects' trajectories, are far from being simple. In fact, all the sub-systems in which the problem can be decomposed are characterized by its intrinsic challenging issues. In general, the architecture can be summarized as composed by (at least) three main components:

- a *Detection and Tracking Module* that, starting from the acquired video sequence, detects the objects moving in the scene and extracts their trajectories;
- a *Storage Module*, which is in charge of storing the extracted data by means of suited indexing strategies;
- a *Retrieval Module*, which allows to retrieve salient data for visualization and statistical purposes on the basis of the specific queries submitted by the user.

As for the first of the above mentioned phases, that is the extraction of the motion trajectories of moving objects, different tracking algorithms have been proposed [3][4], providing reasonably usable solutions.

On the other hand, only a modest attention has been devoted to systems for storing and retrieving information from very large scale Moving Object Databases (MODs) [5]. In fact, the major part of the works dealing with the analysis of motion information has mainly focused on clustering [6] or anomalous detection [7]; even some of these approaches are particularly efficient, they suffer from a common limitation: these methods only allow pre-determined queries, which involves the use of devised and optimized system architecture for supporting a bunch of queries referring to a given spatial area. It also means that the user is not allowed to choose the query parameters at query time, but he can only fix the retrieval parameters in the pre-processing phase.

While it is simple to imagine how much the flexibility degree of such a type of system can increase, it is not likewise to design a system architecture having these potentiality.

The most significant contributions to design an architecture able to cope with large amount of trajectory data can be obtained by browsing the literature coming from the database field. A widely adopted solution for spatial indexing founds on R-trees [8], which are tree data structures that hierarchically organize geometric bidimensional data by representing each object through its Minimum Bounding Rectangle (MBR). Starting from Guttman's pioneering paper, a lot of spatio-temporal indexing scheme have been proposed for many applications contexts, most of which are optimizations of R-trees [9][10].

 ISBN: 978-1-61208-227-1

In order to overcome the above mentioned limitations of the most common trajectories analysis systems by taking advantage from the solution designed in the database field, we propose a general-purpose system for the analysis of moving objects' trajectories which allows the user to choose at query time the query parameters. The generic scenario can be briefly described as follows: a camera records the moving objects in the entire scene for a given, generally long, period of time, but no area of interest is a priori defined. The system finds and efficiently stores the objects' trajectories and allows to solve Dynamic Spatio-Temporal (*DST*) queries, e.g. queries finding all the trajectories passing through an area defined directly by the user within the query (i.e. at query time). Moreover, the system also supports Flow-DST queries, which provide complex statistical analysis in a fully configurable format. Our system extends an off-the-shelf solution for storing the collected data; in particular, it handles with the problem that the actual spatial indexes for three-dimensional data are not widely available and extends a redundant solution we proposed recently [11].

II. The Proposed Method

A generic trajectory T^i can be expressed as a sequence of spatio-temporal points $T^i =< P_1^i, P_2^i, ..., P_N^i >$, where the generic point $P_k^i = (x_k^i, y_k^i, t_k)$ represents the position (x_k^i, y_k^i) of the i-th object at time t_k. We choose to represent each trajectory T^i according to the line segments model, so that a trajectory is obtained by interpolating consecutive points of the sequence. Furthermore, we associate to each trajectory its relative *Minimum Bounding Rectangle (MBR)*, corresponding to its maximum extents in the three-dimensional space.

Once proper indexing strategies have been performed, the problem becomes to find fast and effective means for information extraction from the stored data; for this purpose, we introduce two novel query types, which have been made available in the system; the former, namely the Dynamic Spatio-Temporal (*DST*) query, is the basis for the formulation of more complex queries and allows to answer requests from the user searching for the objects' trajectories passing through a given spatial area in a given time interval.

More formally, a typical *DST* query would appear as: "find all the people passing through a given area A_1 in a given time interval $[t_s, t_e]$"; according to this formulation, we can think to the area as a rectangle with coordinates (x_s, y_s) and (x_e, y_e) while $[t_s, t_e]$ are the starting and final time instants. This leads to the definition of a query box B, which can be associated to each *DST* query:

$$B = \{(x_s, y_s, t_s), (x_e, y_e, t_e)\},$$

The spatio-temporal volume is geometrically defined by a volume delimited by its bottom-left-back point (x_s, y_s, t_s) and its top-right-front point (x_e, y_e, t_e); this volume, in turn, is composed by a spatial interval with top-left point (x_s, y_s) and bottom-right point (x_e, y_e) in a bidimensional space, and a temporal interval (t_s, t_e) representing the third coordinate. An instance of a *DST* query is shown in Figure 1a.

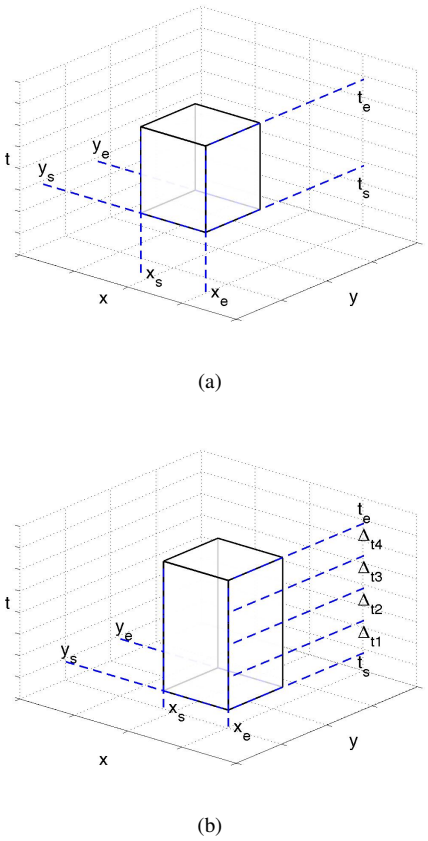


Figure 1: Geometric interpretation of different types of query: *DST* (a) and *F-DST* (b).

Starting from the formalization of the *DST* query, we propose a specialization of it, namely the Flow-DST *(Flow-DST)*, useful for real commercial applications. The Flow-DST *(F-DST)* query is introduced for analyzing the objects' motion flow in the observed scene; typical examples of this query typology are *"find the total number of vehicles passing by a given street (the spatial area can be dynamically defined at query time) each week-end of the last three months"*, or *"find the number of pedestrians passing by a given access gate each couple of hours during the last two days"*. From a geometric perspective, a *F-DST* can be seen as the application of many DST queries so as to obtain results at fixed time intervals, as shown in Figure 1b.

A. Indexing and Query Answering

In this subsection, we will describe the proposed indexing scheme with reference to the *DST* query since, as it has been clearly described in the query formalization section, the *F-DST* can be derived from it.

 ISBN: 978-1-61208-227-1

A simple algorithm for retrieving the trajectories satisfying a *DST* query is based on processing, for each trajectory, all its segments, starting from the first one: as soon as the intersection occurs, it can be concluded that the trajectory intersects the query box. In order to determine if a trajectory segment lies inside or outside a query box, a *clipping* algorithm [12] is needed. Unfortunately, despite of its simplicity, the use of a clipping algorithm is not suited for handling large datasets, so demanding for more efficient approaches. This is the main motivation why we propose to use spatial indexing strategies to reduce both the time needed to extract the trajectories from the database and the number of trajectories to be clipped; this leads, as a consequence, to the real-time processing of a *DST* query. Before going into detail about the use of these indexes, we here recall some basic aspects of spatial indexing.

Spatial indexes are usually aimed to improve the efficiency when handling with geometric data types like points, lines and polygons and querying spatial relationships among them. Although many commercial databases provide efficient three-dimensional indexes, these usually restrict the intersection operation to the bidimensional case; for this reason, we propose to represent the 3D problem in terms of one or more 2D sub-problems. While this choice allows to take advantage of off-the-shelf 2D solutions, it must be noticed that, in the bidimensional space, the intersection between the trajectory and the corresponding query box is a necessary but not sufficient condition; in fact, when the trajectory intersection with the query box holds in each of the three 2D planes, it will not necessarily hold in the 3D plane too, while the opposite is trivially true.

Starting from the above considerations, we represent the *i-th* trajectory T^i through the original sequence of points in the 3D space (x, y, t), together with the three different MBR projections (MBR_{xy}, MBR_{xt} and MBR_{yt}), as shown in Figure 2.a and 2.b.

We verify the intersection of the trajectories with the corresponding bidimensional query box in each of the three 2D planes, as depicted in Figure 2.c. Let I_{xy}, I_{xt} and I_{yt} be the resulting sets of trajectories intersecting the query box and defined as:

$$I_{xy} = \{T : MBR_{xy}(T) \cap B_{xy} \neq \emptyset\} \quad (1)$$

$$I_{xt} = \{T : MBR_{xt}(T) \cap B_{xt} \neq \emptyset\} \quad (2)$$

$$I_{yt} = \{T : MBR_{yt}(T) \cap B_{yt} \neq \emptyset\}, \quad (3)$$

where B_{xy}, B_{xt} and B_{yt} are the three projections of the 3D query box B. The set C of candidate trajectories to be clipped in the 3D space is therefore defined as $C = \{I_{xy} \cap I_{xt} \cap I_{yt}\}$. At last, we apply the clipping algorithm and obtain the final intersection result, as shown in Figure 2.d.

According to the indexing strategy above described, the capability to significantly reduce the number of trajectories to be clipped plays a crucial role, as the huge amount of trajectory data represents a key factor of complexity. As a consequence, we introduce a segmentation stage aimed at improving the *selectivity* of the indexes, which, in turn, only depends on the trajectory geometry.

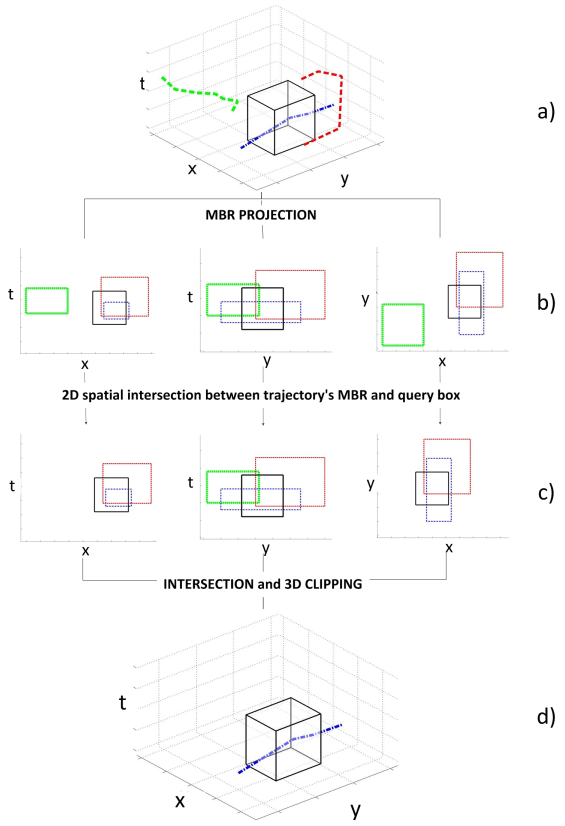


Figure 2: An overview of the method. (a) a query box and three trajectories; (b) the projections of the trajectory MBRs on the planes; (c) the MBR projections intersecting the query box (the three MBR projections of the red and blue trajectories intersect the query box); (d) the final result of our method, after the application of the clipping algorithm.

B. Segmentation

The *selectivity* of the indexes in each plane is related to the area of the corresponding MBR which, in turn, only depends on the trajectory geometry, so being (apparently) fixed. This is the reason why we decided to introduce a segmentation stage, aimed at increasing the selectivity of the indexes.

The proposed algorithm works recursively: initially (that is at iteration 0) it assumes that the trajectory T^k is composed by a single unit ${}^0U_1^k$, that is split into a set of m consecutive smaller units $\{{}^1U_1^k, \ldots, {}^1U_m^k\}$; each of the ${}^1U_i^k$ is in turn inspected and, if the stop criteria are not satisfied, it is further split.

Let us analyze how a generic unit ${}^{(i-1)}U = \{P_1, \ldots, P_m\}$ at iteration $i-1$ is split into $\{{}^iU_1, \ldots, {}^iU_n\}$ at iteration i; we first choose a *split-dimension* and a *split-value*. Assume,

as an example and without loss of generality, that x has been chosen as the *split-dimension* and let x^* be the *split-value*. In addition, assume that $x_1 < x^*$. According to these hypotheses, ${}^iU_1$ is the set of the consecutive points lying on the left of the *split-value* ${}^iU_1 = \{P_1, \ldots, P_k\}$, where P_k is the first point such that $x_k \geq x^*$. Then, the second unit will be formed by the sequence of consecutive points lying on the right of the *split-value*, where P_l is the first point such that $x_k \leq x^*$. The inspection of ${}^{(i-1)}U$ ends when the last point P_m is reached.

According to the above considerations, the criteria for the choice of the two parameters, *split-dimension* and *split-value*, play a crucial role. Since we aim at optimizing the indexing strategy, the proposed segmentation algorithm is based on the occupancy percentage on each 2D coordinate plane. Let V be the volume containing all the trajectories stored until this moment. First, we calculate the coordinate plane corresponding to the maximum among the three occupancy percentage values O_{xy}, O_{xt} and O_{yt} of the trajectory unit's MBRs, with respect to the projections of V on the coordinate planes (V_{xy}, V_{xt} and V_{yt}):

$$O_{xy} = \frac{MBR_{xy}(U)}{V_{xy}} \tag{4}$$

$$O_{xt} = \frac{MBR_{xt}(U)}{V_{xt}} \tag{5}$$

$$O_{yt} = \frac{MBR_{yt}(U)}{V_{yt}} \tag{6}$$

Without loss of generality, suppose that the maximum occupancy percentage value is O_{xy} and, consequently, the corresponding plane is xy; let $width$ and $height$ be the two dimensions of $MBR_{xy}(U)$, respectively along the coordinates x and y; the *split-dimension* sd is defined as:

$$sd = \begin{cases} x & \text{if width} > \text{height} \\ \\ y & \text{otherwise} \end{cases}$$

Given the *split-dimension* sd we choose, as the *split-value* sd^*, the MBR average point on the coordinate sd.

The regular termination of the algorithm is reached when all the trajectory points have been processed; anyway, an abnormal termination is also possible during each iteration step, on the basis of two stop criteria: PS^{min} attains the minimum number of points belonging to a trajectory unit, while PA^{min} is the minimum allowed size, in percentage value with respect to the entire scenario, of an MBR area.

III. Experimental Results

The system has been implemented by storing the trajectory data in Postgres using PostGIS [13], being the latter Postgres extension for storing spatial data like points, lines and polygons. Data are indexed using the standard bidimensional R-tree over GiST (Generalized Search Trees) indexes; the specialized literature highlights that this choice guarantees higher performance in case of spatial queries, if compared with the PostGIS implementation of R-trees. Once data have been indexed, PostGIS provides a very efficient function to perform intersections between boxes and MBRs in a 2D space.

We conducted our experiments on a PC equipped with an Intel quad core CPU running at 2.66 GHz, using the 32 bit version of the PostrgreSQL 9.1 server and the 1.5 version of PostGIS. We tested our retrieval system with real and synthetic data. The synthetic data have been generated as follows. Let W and H be the width and the height of our scene and S the temporal interval. Each trajectory T^i starting point is randomly chosen in our scene at a random time instant t_1^i; the trajectory length L^i is assumed to follow a Gaussian distribution, while the initial directions along the x axis and the y axis, respectively d_x^i and d_y^i, are randomly chosen. At each time step t, we first generate the new direction, assuming that both d_x^i and d_y^i can vary with probability PI_x and PI_y respectively; subsequently, we choose the velocity along x and y at random. The velocity is expressed in pixels/seconds and is assumed to be greater than 0 and less than two fixed maximum, V_x^{max} and V_y^{max}. Therefore, the new position of the object can be easily derived; if it does not belong to our scene, new values for d_x and/or d_y are generated. Table I reports the defined free parameters with the values used to generate our data.

Table I: THE PARAMETERS USED IN OUR EXPERIMENTS.

Scene width (pixels)	10^4
Scene height (pixels)	10^4
Time interval length (secs)	10^5
PI_x	5%
PI_y	5%
V_x^{max}	10 pixels/secs
V_y^{max}	10 pixels/secs

First, we decided to test our segmentation algorithm assuming $PA^{min} = 0.1\%$ and $PS^{min} = 100$; we generated and segmented 6000 trajectories with $L \in \{1000, 2000, 3000, 4000, 5000, 10000\}$; for each trajectory T^i we measured the number of obtained segments (N_{seg}^i). Last, the obtained N_{seg}^i are averaged over L, so obtaining $\overline{N_{seg}}$. Not surprisingly, we have, with very good approximation, that the number of segments $\overline{N_{seg}}$ linearly increases with the trajectory length L, as shown in Figure 3.

The use of the segmentation algorithm clearly makes the proposed system able to outperform the method presented in [11], as depicted in Figure 5. For the sake of readability, Figure 5 only highlights the improvement for $D_c \in \{10\%, 50\%\}$ and having $L = 5000$. In particular, the diamonds refer to the new method, while the circles refer to the previous one.

Furthermore, we investigate on the time needed to process a Flow-DST query (FT). We can note that FT is N times (N being the number of intervals we are interested in) the

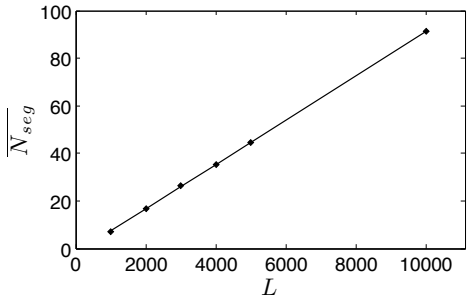


Figure 3: The performance of the segmentation algorithm.

Table II: NUMBER N OF TIMES EACH QUERY IS REPEATED AS D_c VARIES.

D_c	5%	10%	20%	30%	50%
N	40	20	10	7	4

time needed to perform a DST query (QT); QT, in turn, is a function of at least four parameters, namely the number of trajectories T, the average trajectories length L, the query cube dimension D_c, expressed as percentage of the entire scenario, and the position of the query box P_c:

$$FT = N * f(T, L, D_c, P_c). \quad (7)$$

Among the above parameters, P_c strongly influences the time needed to extract the trajectories as these are not uniformly distributed, especially when considering real world scenarios. In order to avoid the dependency on the query cube position, we decided to repeat the query a number of times which is inversely proportional to the query cube dimension, positioning the query cube in different points, as shown in row N of Table II; finally, the results have been averaged to obtain:

$$\overline{FT} = f(T, L, D_c, N). \quad (8)$$

Furthermore, we have experimentally verified that $\overline{N_{seg}}/L = k$, with k constant (Figure 3). It means that, on average, the length of each unit can be assumed to be constant. We can thus derive, with good approximation, that:

$$\overline{FT} = N * f(\overline{N_{seg}}, D_c). \quad (9)$$

Starting from these considerations and assuming $L = 1000$, we can measure the time needed to perform several Flow-DST queries with $D_c \in \{5\%, 10\%, 20\%, 30\%, 50\%\}$ as N_{seg} varies. Figure 4 shows the obtained results. It can be noticed that $\overline{N_{seg}}$ linearly increases with f. However, it is worth pointing out that the time needed to process each query is significantly influenced by the clipping procedure which, in turn, is strongly dependent on the time for the extraction of each trajectory: in fact, a segment unit need to be extract before to be clipped. It clearly means that an optimization of the segmentation parameters can still improve the performance of our method.

Table III: AVERAGED TIME (IN SECONDS) TO SOLVE A *DST* QUERY ON THE MIT TRAJECTORIES DATASET.

D_c	N	T^1	T^2	T^3	$\overline{QT}$
1%	200	0.003	0.010	0.009	0.022
5%	40	0.007	0.064	0.115	0.186
10%	20	0.013	0.154	0.320	0.487
20%	10	0.038	0.533	1.383	1.954
30%	7	0.097	1.566	4.014	5.673
50%	4	0.173	5.878	14.924	20.975

It must be noticed that the synthetic data really stressed the system, resulting in trajectories with tens of millions of points, which is over and above the average trajectory length of available datasets. To confirm this consideration, we also tested the performance of our indexing scheme on a well-known real dataset, the freely available MIT trajectory dataset [14], obtained from a parking lot scene within five days; the dataset is composed of approximately $4 * 10^4$ trajectories with 108.81 points in each trajectory (on average). At loading time, each trajectory has been segmented using $PA^{min} = 1$ and $PS^{min} = 100$, so obtaining approximately $1.92 * 10^6$ segments with 23.71 points in each segment (on average). Table III shows $\overline{QT}$ (in seconds) as D_c varies. The table also shows $\overline{QT}$ results from the sum of three terms: T^1 is the time needed to select the segments whose bounding boxes intersect the query box on each bi-dimensional plane, T^2 is the time to clip the segments while T^3 is the time needed to extract the whole trajectory. It is possible to note that the obtained results confirm the efficiency of the proposed method.

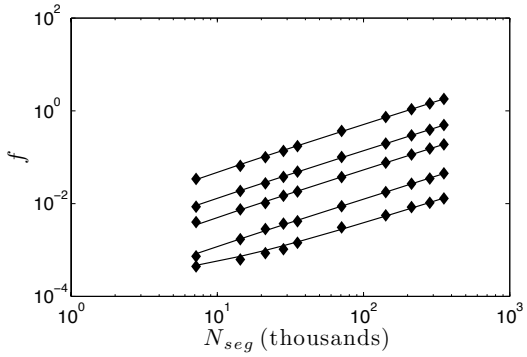


Figure 4: The performance of the system as N_{seg} vary for different values of D_c.

IV. CONCLUSION

We addressed the problem of efficiently storing, indexing and querying spatio-temporal data for motion trajectory analysis. Our main contributions lie in a significant improvement of the overall efficiency and the wide adaptability of the queries. The former contribution is achieved by proposing an

 ISBN: 978-1-61208-227-1

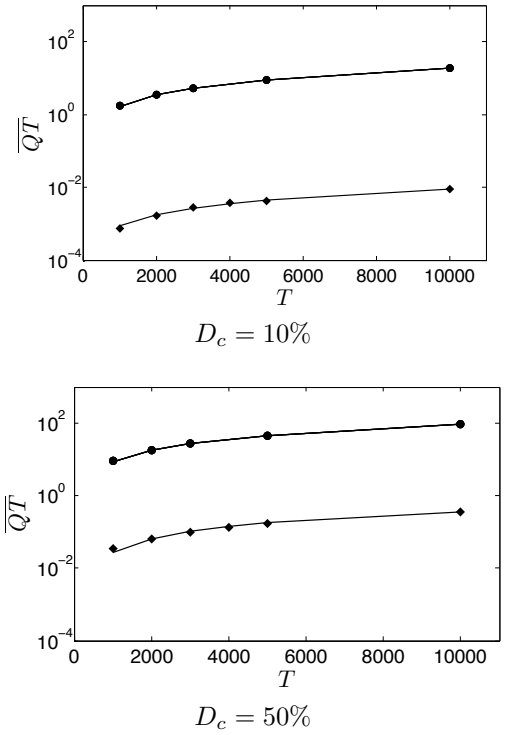


Figure 5: The results obtained with the solution proposed in [11] (circles) compared with the results obtained with the solution here as T varies (L=5000).

indexing scheme based on the use of off-the-shelf solutions; in addition, the introduction of *Dynamic Spatio-Temporal* and *Flow DST (F-DST)* queries has provided means for defining the query parameters at runtime and also for answering frequently occurring retrieval problems.

The preliminary tests, conducted over synthetic and real data, confirm the effectiveness of the approach in terms of query generality and computational efficiency. Anyway, some improvements are still possible. First, the application of multithreading for the clipping algorithm could significantly improve the performance of our method, since the system could take advantage from multi-core and multi-processors systems. In addition, a deeper analysis could be conducted on the optimal choice of the segmentation parameters.

ACKNOWLEDGMENT

This research has been partially supported by A.I.Tech s.r.l. (a spin-off company of the University of Salerno, www.aitech-solutions.eu) and by FLAGSHIP *InterOmics* project (PB.P05, funded and supported by the Italian MIUR and CNR organizations).

REFERENCES

[1] K. Teknomo and P. G. Gerilla, "Pedestrian Static Trajectory Analysis of a Hypermarket," in *Proceedings of the 8th International conference of the Eastern Asia Society for Transportation Studies*, vol. 7, Nov. 2009, pp. 1–12.

[2] K. Okamoto, A. Utsumi, T. Ikeda, and H. Yamazoe, "Classification of pedestrian behavior in a shopping mall based on lrf and camera observations," in *IAPR Conference on Machine Vision Applications*, vol. 110, no. 421, 2011, pp. 233–238.

[3] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, Dec. 2006.

[4] R. DiLascio, P. Foggia, A. Saggese, and M. Vento, "Tracking interacting objects in complex situations by using contextual reasoning," in *Computer Vision Theory and Applications (VISAPP), 2012 International Conference on*, 2012, pp. 104–113.

[5] A. d'Acierno, M. Leone, A. Saggese, and M. Vento, "A system for storing and retrieving huge amount of trajectory data, allowing spatio-temporal dynamic queries," in *Proceedings of the "IEEE Conference on Intelligent Transportation Systems (ITSC)"*, 2012, pp. 989–994.

[6] B. T. Morris and M. M. Trivedi, "Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 312–319.

[7] G. Acampora, P. Foggia, A. Saggese, and M. Vento, "Combining neural networks and fuzzy systems for human behavior understanding," in *Proceedings of the "IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)"*, 2012, pp. 88–93.

[8] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *Proceedings ACM SIGMOD Conference*. New York, NY, USA: ACM, 1984, pp. 47–57.

[9] Z. Song and N. Roussopoulos, "Seb-tree: An approach to index continuously moving objects," in *Proceedings of the 4th Conference on MDM*. London, UK: Springer-Verlag, 2003, pp. 340–344.

[10] J. Priyadarshini, P. AnandhaKumar, M. Aparna, J. Geetha, and N. Shobana, "Indexing and querying technique for dynamic location updates using r k-d trajectory trie tree," in *International Conference on Recent Trends in Information Technology (ICRTIT)*, 2011, pp. 1143 –1148.

[11] A. d'Acierno, A. Saggese, and M. Vento, "A redundant bi-dimensional indexing scheme for three-dimensional trajectories," in *Proceedings of the 1th Conference on Advances in Information Mining and Management*, 2011, pp. 73–78.

[12] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice in C (2nd Edition)*. Addison-Wesley, 2004.

[13] R. Obe and L. Hsu, *PostGIS in Action*. Greenwich, CT, USA: Manning Publications Co., 2011.

[14] X. Wang, K. T. Ma, G.-W. Ng, and W. E. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *Int. J. Comput. Vision*, vol. 95, pp. 287–312, December 2011.

 ISBN: 978-1-61208-227-1

On Biometric Verification of a User by Means of Eye Movement Data Mining

Youming Zhang and Martti Juhola
Computer Science, School of Information Sciences
33014 University of Tampere
Tampere, Finland
Youming.Zhang@uta.fi, Martti.Juhola@sis.uta.fi

***Abstract*—In biometric verification, a signal, image or other dataset is measured from a subject to detect him or her to be or not to be an authenticated subject such as the user of a computer. So far, biometric verification has mainly been on the basis of fingerprints or face images, infrequently other images, e.g., iris. We studied the idea to apply fast eye movements called saccades to verify an authenticated user from among other subjects. We recorded eye movement signals with eye movement cameras using a suitable visual stimulation for a subject. By means of machine learning methods, we classified a subject's eye movements to verify whether one was an authenticated user. We employed multilayer perceptron networks, radial basis function networks, support vector machines and logistic discriminant analysis for classification. The best accuracy results obtained were approximately 90% and showed that it is possible to verify a subject according to saccade eye movements.**

Keywords-biometric verification; eye movements; saccades; multilayer percetron neural networks; radial basis function networks; support vector machines; logistic discriminant analysis

I. INTRODUCTION

So far, various biometric data sources have been used to verify a subject. Mostly fingerprints [1, 2] and face images [3] are applied to this task. Other images measured from subjects such as iris images [2, 4] are also studied. In addition to these two-dimensional data sources, one-dimensional signals are also used, e.g., voice signals [5]. Usually, these datasets contain an abundance of data and several variables are computed from them to ground the verification procedure on variable values of different subjects. Data mining tasks needed here may be complicated because of complex data.

Eye movements are a new potential alternative for biometric verification. Eye movements have been researched for decades in medicine. During the past 15 years eye movements have become an important research objective for human-computer interfaces. Along with these applications efficient eye movement cameras have been developed. Since there is long-term experience in the signal analysis of eye movements, for example [6-8], for biomedical and physiological applications, it was a direct development to attempt to utilize them for biometric verification of a subject simulating a computer user. Note that verification corresponds to the binary classification between two classes: an authenticated user and other subjects.

There are a few different eye movement types such as saccade, nystagmus, smooth pursuit and vestibulo-ocular reflex eye movements [7]. Probably the most frequent of all are saccades which are made while looking at surroundings or reading a text. In addition, they are very fast, in fact the fastest movements of man. They are easy to visually stimulate and their recording does not require more time than a few minutes for our tests. Those other eye movement types would require longer recordings or more complicated stimulation arrangements [7]. For these reasons, we chose saccades to be our data sources here, particularly after observing differences between saccades of individuals [7].

Up to now, a couple of attempts only have been published about this idea to use eye movements for biometric verification. In one research [8] they recorded saccade eye movement signals to compute cepstrum from these and classified signal analysis outcomes by using naïve Bayesian method, nearest neighbour searching, decision trees as well as support vector machines. In another research [9] they used a computational oculomotor model on the parameters of which verification was based using nearest neighbour searching and decision trees. Our approach differs from those since we use physiological variables computed from eye movement signals. Most of these variables have been employed for long in biomedical investigations [6,7].

II. EYE MOVEMENT DATA

We recorded saccade eye movements with a two-camera system (Visual Eyes, Micromedical Technologies, UK). Its resolution is 320×240 and sampling frequency or frames per second 30 Hz. The camera system recognized positions of each pupil from successive images of a video stream to detect eye movements. The system records horizontal and vertical signals, but we used the horizontal direction only. We wanted to keep the arrangement as simple as possible for stimulation design so that this was simple for a subject in order to avoid complex stimulations. Furthermore, using simple stimulations means that long recordings are not necessary which is important to see this biometric verification idea as sensible. On the other hand, the more data from each individual, the easier it is perhaps to separate him or her from the group of other subjects. The sampling frequency of 30 Hz was low compared to other typical ones used in eye movement camera systems such as 50 or 60 Hz, occasionally even higher like 200 Hz. Nevertheless, it was

 ISBN: 978-1-61208-227-1

interesting to see whether this low sampling frequency allowed verification. Perhaps using a higher frequency in the future could only better results because of more accurate variable values to be computed. The system included one camera for each eye embedded in the mask attached tightly with a headband. The one of lower noise level of two eye movement signals was used for verification. Usually, both are almost identical.

We used the same stimulation series for every subject. This is, of course, the essential detail for biometric verification so that we can assume that every subject has followed the same stimulation by his or her gaze and we can classify them according to their eye movements. Each subject saw a horizontally jumping LED light dot in front of him or her. The stimulation component of the eye movement recording system included a horizontal LED bar in which one LED was switched on for a while, then switched off and another switched on immediately, and so on, by varying the LED to be next switched on. This way different gaze angles were formed. Intervals between light dot jumps were varying to make them random for a watching subject. Since intervals of 1-3 s were short and varying, the spectator could learn neither them nor varying stimulation angles. Varying, "random" intervals are important to minimize anticipations of a subject while waiting for the next stimulation movement. Anticipation would occur if latency or reaction time from the beginning of a stimulation movement to the beginning of its response, saccade, were shorter than 0.120 s seen as a minimum latency in the physiological sense [7]. It takes some time for the brain to observe a movement and control the response to move the eyes.

The present stimulation arrangement was used to simulate the beginning of a computer session where a user would first sit down to start the machine and to wait for its initialization. We can imagine that the eye movement stimulation would be run immediately after the initialization by stimulating a subject with a few dozen stimulation movements on the screen of a computer or mobile device. Thereafter, the verification procedure would be run.

We used saccades with the largest stimulation amplitudes of around 48 º only since saccades of such large amplitudes contain greater differences between subjects than those with small amplitudes [7]. Great differences between subjects aid in verification. Nonetheless, there were smaller stimulation angles between large to give a random character between stimulations from a spectator's viewpoint. Consequently, we obtained 20 large amplitude saccades from every subject. Values of saccade variables depend on saccade amplitudes. Thus, we used merely the saccades of the largest stimulation amplitude,

For the sake of the low sampling frequency of 30 Hz, we interpolated every signal with a cubic spline method up to 1000 Hz. The purpose here was to simulate a sampling frequency of the newest, expensive high resolution eye movement cameras and, most of all, to estimate values of eye movement variables more precisely than enabled by the original signals sampled at 30 Hz.

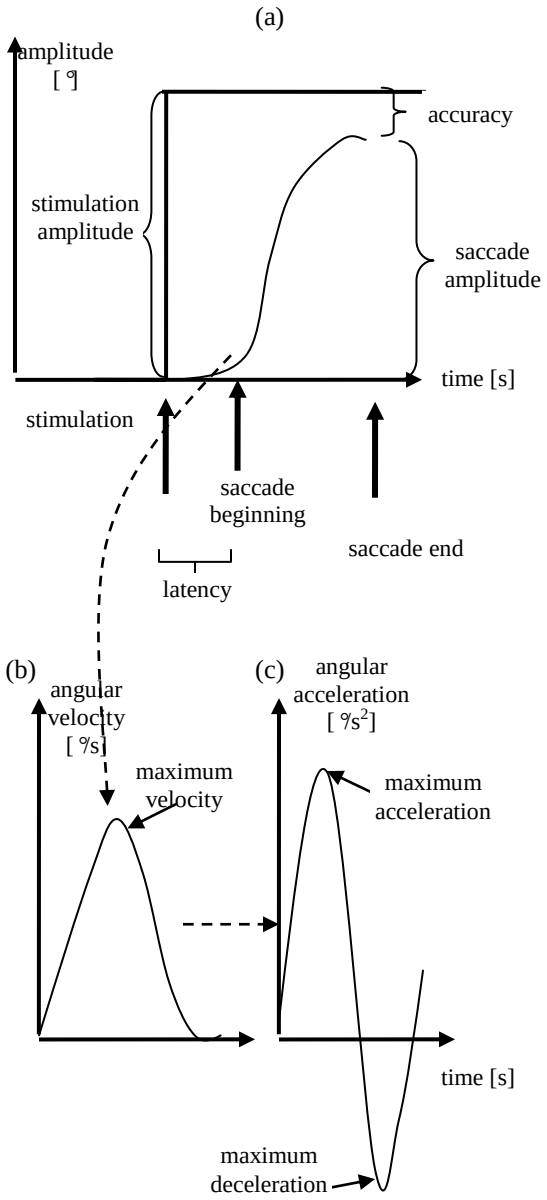


Figure 1. (a) The step (broken line) is a stimulation movement produced by a horizontally jumping light dot from the left (down in the figure) to the right (up). A saccade as a response follows it after a latency. The difference between amplitudes determines a negative accuracy, because the saccade amplitude is smaller here. A positive accuracy is also possible, but is more infrequent than negative. In our tests these values were used as absolute. Accuracy, amplitude and latency were three variables used. (b) From the saccade signal the first derivative approximation of the velocity curve is computed from which (c) the second derivation of the acceleration curve is approximated. The maximum velocity, maximum acceleration and maximum deceleration were other three useful variables to be computed.

III. SIGNAL ANALYSIS AND DATA PREPROCESSING

Fig. 1 depicts an ideal saccade and its stimulation as a schema. The first signal analysis task is to detect the exact beginning and end of every stimulation movement and those of the following response eye movement, saccade.

 ISBN: 978-1-61208-227-1

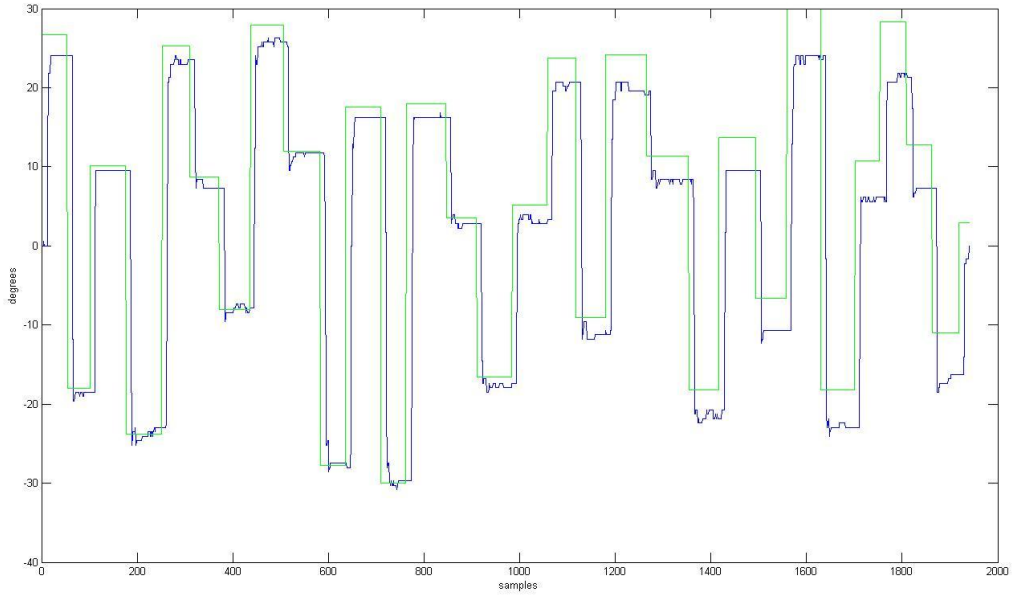


Figure 2. A smooth (green) stimulation signal of 64 s sampled at 30 Hz and its (blue) response with saccades.

The former is easy to detect since it is a clear step in a signal. The latter may rarely be somewhat corrupted by noise or artifacts such as blinks; See Fig. 2, including horizontal saccades.

If a saccade is inaccurate, its amplitude clearly differs from that of its stimulation. The brain can rapidly produce a corrective saccade with a small amplitude to correct the gaze closer to the objective. One cannot sense this correction movement, but it is "automatic". We did not include possible, quite infrequent corrective saccades, but determined the accuracy of a saccade along with the primary saccade as usual. A response to its stimulation movement had to resemble a real saccade sufficiently to be accepted for further use in signal analysis. In principle a subject might not occasionally follow the target with the gaze. This would yield no saccade at all. Anticipation as a too early eye movement including a latency value less than 0.120 s or even a saccade before a stimulation would be rejected as no actual responses to stimulations. The quality of signals given by the camera system was high with low noise. Thus, rejections of eye movements from signals were infrequent, no more than a few per cent of all saccades.

The same stimulation movements (Fig. 2) were run for every recording, so that eye movements of subjects were comparable with each other. A stimulation series included four stimulations with the largest amplitude of 48 º. Five recordings were run successively from every subject giving 20 large saccades for a subject.

The five recordings of each subject formed our data for biometric verification tests. The stimulation series also contained saccades of smaller amplitudes between those four large to make the stimulation series more random-like for a subject not able to guess the direction or amplitude of a stimulation movement or an interval between two successive stimulations. Intervals were 1-3 s within a recording of 64 s.

After the interpolation of signals, the first derivative and second derivative were computed with approximation formulas such as two-point central difference differentiation [8] from each eye movement signal. A saccade beginning was found provided that absolute velocity values rapidly increased above a threshold of 50 º/s and the corresponding saccade end was found when velocity decreased back below that threshold. After detecting a saccade and ensuring that it was valid according to latency criterion, etc., all its variable values were computed and stored: amplitude, accuracy, latency and maximum velocity, acceleration and deceleration.

During recordings, a sitting, alert, relaxed subject was asked to follow the stimulation light dot by the gaze. In all, we recorded five successive recordings from healthy 132 subjects from whom 33 were females and 99 males. Mean and standard deviation of their ages were 26.2±7.2 years. Neither alcohol nor medications were used during 24 h before a measurement. We wanted to test mainly young subjects in a pretty homogeneous dataset to create a strict testing basis. Age, alcohol or medications may have influence on values of saccade variables. Means and standard deviations were the following: amplitude 48.0±13.4 º, accuracy 3.2±8.4 º, latency 0.269±0.057 s, maximum velocity 1038±322 º/s, maximum acceleration 47591±23166 º/s^2 and maximum deceleration 44845±24745 º/s^2.

IV. VERIFICATION PROCEDURE

In biometric or whatever user verification, we have to prepare two opposite conditions: a subject attempting to log in is either authenticated user or impostor. Thus, we built our test procedure to take these two conditions into account.

 ISBN: 978-1-61208-227-1

When machine learning algorithms are used, we have to construct a training set and its corresponding test set. The content of these two sets are varied on the basis of available data. In the current case, our eye movement data were quite limited. Although there were several subjects, the bottle neck for tests was the small number 20 of saccades of the largest amplitude. Therefore, we implemented two experimental test settings called Alternatives 1 and 2. In the one of them for every subject there were either three recordings (12 saccades) in a training set and the rest of two recordings (q=8 saccades) in the corresponding test set. In the other there were four recordings (16 saccades) in a training set and one recording (q=4 saccades) in its test set. Since from every subject there were five recordings all in all, we obtained c=10 different combinations of a training set and a test set from five recordings for the former Alternative 1 and c=5 combinations for the latter Alternative 2. These were prepared for every of n=132 subjects. Our aim was to test our data as broadly as possible as conventional while applying data mining methods for classification.

Our verification task (Fig. 3) comprised two classes. Therefore, it was best that the number of saccades of an authenticated user and that of other subjects now called nonusers were not very imbalanced. We had either m=12 (Alternative 1) or m=16 (Alternative 2) saccades of an authenticated user in a training set. We then took one saccade randomly from either 2m=24 or 32 nonusers to test Condition 1 (an authenticated user) and, in addition, still one saccade to represent an impostor from q=8 or 4 other random subjects. Nonusers and impostors were naturally represented by different random subjects from among n-1=131 subjects (an authenticated user excluded). At first, we implemented tests with this approach since we may assume that randomly selected subjects represent a more extensive area in the variable space than one authenticated. Nonetheless, we noticed that better results could be obtained by once copying the saccades of an authenticated user to balance the class size of an authenticated user's class and that of nonusers to be equal 2m. Copying once m saccades of the former increased the density of these saccades in a dataset.

In the verification procedure, the following symbols are also employed. All tests were repeated r=10 times since there were random choices of saccades of nonusers and impostors and also random initializations, among others, in multilayer perceptron networks. To test the remaining q saccades were taken to a test set where q was equal to 8 (Alternative 1) or 4 (Alternative 2). Symbols TP and FN equal the numbers of true positive and false negative decisions in classifications and FP and TN those of false positive and true negative decisions. On the basis of the two former, a decision for a subject is made whether a test subject is an authenticated user (Condition 1). Correspondingly, the two latter are used for a decision whether a test subject is an impostor (Condition 2).

$C1_1$=$C2_1$=$C1_2$=$C2_2$=0; % counters for correct classifications of authenticated users and those of impostors

For h=1:r % iterations of the main loop

For i=1:n % one by one as an authenticated user
TP_2=TN_2=FP_2=FN_2=0 (Alternative 2);
For j=1:c % c combinations of recordings
Take m saccades from 3 (Alternative 1) or 4 (Alternative 2) recordings of an authenticated user to a training set;
Copy these m saccades in the training set;
Take randomly 2m nonusers and one saccade from each and add these saccades to a training set;
Train a model with 4m saccades of two classes: an authenticated user and nonusers;
TP_1=TN_1=FP_1=FN_1=0 (Alternative 1);
For j=1:q % tests of Condition 1
Classify a test saccade of an authenticated user into either correct class
TP=TP+1
or incorrect class
FN=FN+1;
End
For k=1:q % tests of Condition 2
Classify a test saccade of an impostor into either correct class
TN=TN+1
or incorrect class
FP=FP+1;
End
% Follow majority vote for decision
If $TP_1 \geq FN_1$ **then** $C1_1$=$C1_1$+1 (Alternative 1);
If $TN_1 > FP_1$ **then** $C2_1$=$C2_1$+1 (Alternative 1);
End
% Follow majority vote for decision
If $TP_2 \geq FN_2$ **then** $C1_2$=$C1_2$+1 (Alternative 2);
If $TN_2 > FP_2$ **then** $C2_2$=$C2_2$+1 (Alternative 2);
End
End
(Alternative 1)
Accuracy of authenticated users=100 % · $C1_1/(r \cdot n \cdot c)$
Accuracy of impostors=100 % · $C2_1/(r \cdot n \cdot c)$
(Alternative 2)
Accuracy of authenticated users=100 % · $C1_2/(r \cdot n)$
Accuracy of impostors=100 % · $C2_2/(r \cdot n)$

Figure 3. Verification procedure for authenticated users (Condition 1) and impostors (Condition 2). Two different test settings are called Alternatives 1 and 2.

V. CLASSIFICATION RESULTS AND DISCUSSION

The main data mining task was to classify test saccades into two classes: an authenticated user or nonusers. There were n=132 subjects and r=10 main iterations in the verification procedure yielding 13200 decisions in Alternative 1 and 1320 decisions in Alternative 2.

 ISBN: 978-1-61208-227-1

TABLE I. CLASSIFICATION ACCURACIES OF MLP NETWORKS WITHOUT NORMALIZATION: MEANS AND STANDARD DEVIATIONS IN PERCENTS (*ON* EQUALS THE NUMBER OF OUTPUT NODES AND *C* CONDITIONS 1 AND 2)

Accuracies for two test alternatives, output node numbers *ON* and conditions *C*						
Alter-native	***ON***	***C***	***Number of hidden nodes***			
			4	6	8	10
1	1	1	71.8±0.8	70.8±0.9	**71.0±1.6**	70.4±0.8
1	1	2	64.9±0.7	65.2±1.8	**66.5±1.4**	66.4±0.9
1	2	1	72.1±1.2	71.8±1.3	**72.2±0.8**	71.7±1.1
1	2	2	66.8±1.5	66.8±1.5	**66.8±1.7**	67.0±1.5
2	1	1	78.5±3.3	79.6±3.4	78.9±2.5	**78.5±2.5**
2	1	2	74.2±3.1	78.0±3.4	79.8±3.6	**80.8±2.6**
2	2	1	81.9±2.8	**82.6±3.2**	82.2±3.5	79.8±2.8
2	2	2	77.8±1.9	**78.6±3.0**	78.6±2.7	79.2±3.0

TABLE II. CLASSIFICATION ACCURACIES OF MLP NETWORKS WITH NORMALIZATION AND ALTERNATIVE 2: MEANS AND STANDARD DEVIATIONS IN PERCENTS

Accuracies for output nodes and conditions					
Output nodes	***Condition***	***Number of hidden nodes***			
		4	6	8	10
1	1	81.1±2.7	78.4±3.9	**78.9±2.6**	78.4±2.5
1	2	75.8±2.4	79.5±3.2	**81.1±3.6**	80.5±3.2
2	1	80.5±1.8	80.1±4.2	80.0±4.5	**79.6±2.6**
2	2	77.3±2.4	81.7±3.6	80.2±4.3	**82.7±2.5**

We applied multilayer perceptron (MLP) networks [9] with 6 input nodes (6 variables), 4, 6, 8 or 10 hidden nodes and 1 or 2 output nodes for two classes. A validation error was used for MLP networks. It automatically stopped training after 9 or 10 epochs to avoid overtraining. Since we used the backpropagation algorithm in Matlab (MathWorks Inc., USA) also used for all tests of our research, we experimented with its training procedure variations including the adaptive learning rate, Powell-Beale restarts, batch gradient descent with momentum and Levenberg-Marquardt algorithm [10]. For actual tests we used the last method that yielded slightly better results than those of the other.

At first, we investigated possible differences between test results of Alternatives 1 and 2. Since the number of 5 recordings (20 saccades) of each subject was small subject to build training and test sets in data mining, it was important to test more than one alternative. However, the scarcity of the data did not allow more alternatives than the aforementioned two. We also varied the number of output nodes from 1 to 2. On the basis of the best results written in Bold in Tables I and II 2 output nodes produced accuracies 1-4% superior to those of 1 node.

TABLE III. CLASSIFICATION ACCURACIES OF LOGISTIC DISCRIMINANT ANALYSIS AND SVM WITH NORMALIZATION: MEANS AND STANDARD DEVIATIONS IN PERCENTS.

Accuracies for two test alternatives *A* and conditions *C*						
A	***C***	LogDA	***SVM kernels***			
			Linear	2nd deg.	3rd deg.	Gaussian
1	1	78.5±105	80.0±0.5	75.6±1.0	69.6±1.2	**84.9±0.7**
1	2	65.7±1.7	62.2±1.3	63.6±1.7	61.8±1.7	**73.0±1.3**
2	1	86.6±1.7	88.0±2.9	82.7±2.1	74.1±2.5	**92.1±1.9**
2	2	77.4±3.4	73.9±4.8	77.1±2.9	73.3±4.5	**84.8±1.9**

The scales of the variables markedly differed from each other. We tested MLP networks without and with normalization into interval [0,100]. The accuracies obtained without or with normalization had virtually no differences on an average. The results of the former are showed in Table I. Those of the latter are in Table II with Alternative 2 only, since Alternative 2 with the larger training set than with Alternative 1 indicated to be 7-13% better in Table I. The similar observation was gained for all later results. Note that while evaluating results we always have to look at both conditions at the same time, because they both are equally critical objectives. Note also that 50% is seen as a baseline result for Conditions 1 and 2. Because there are two classes of equal size, a random guess between them would be correct with probability 0.5. The number of the hidden nodes from 6, 8 or 10 yielded the best results for the pairs of Conditions 1 and 2.

We ran support vector machines (SVM) with the linear, quadratic, third degree polynomial and radial basis function (Gaussian) kernels. Table III shows results for SVM kernels and logistic discriminant analysis (LogDA). We ran tests for all four SVM kernels and logistic discriminant analysis by using both Alternatives 1 and 2 with and without normalization. Alternative 2 again generated higher results than Alternative 1. The use of normalization according to Table III did not affect average results seemingly at all compared with those not presented without normalization, mostly less than ±1%. SVM with the radial basis function (Gaussian) kernel was the best choice here, but differences were small compared with a few other kernels.

TABLE IV. CLASSIFICATION ACCURACIES OF RBF NETWORKS WITH NORMALIZATION: MEANS AND STANDARD DEVIATIONS IN PERCENTS

Accuracies for two test conditions				
Condition	***Spread and goal***			
	15 0.05	15 0.08	20 0.08	20 0.1
1	75.4±4.1	77.8±0.1	83.4±2.6	**88.5±1.8**
2	92.6±1.6	94.7±1.6	88.9±3.9	**88.9±1.9**

Ultimately, we exploited RBF networks by running system parameters of spread 10, 15, 20, 25, 30, 35, 40, 45 and 50, and goal 0.005, 0.02, 0.03, 0.05, 0.08 and 0.1. The best combinations of these were spread equal to 15 or 20 and goal equal to 0.05, 0.08 or 1.0. Final results of RBF networks

 ISBN: 978-1-61208-227-1

are presented in Table IV. For the RBF networks, our data required normalization, because our tests (not presented here) without it favoured Condition 2 and almost entirely failed with Condition 1. Thus, the results in Table V were computed with normalization and using Alternative 2.

Since our final objective to develop a biometric verification procedure on the basis of eye movements included a criterion that computing time should be fast, it is important to look at running times of the preceding tests. There were 132×10×5=6600 models trained for every test type or structure (cell) in the case of Alternative 2. For Alternative 1 there were 132×10×10=13200 models trained, correspondingly. The training and test time of an MLP network was around 0.5 s on an average. For RBFs that time of one network was around 4 s and for SVMs and LogDA less than 0.05 s. Let us remember that these execution times also included training not always necessary to do while applying a data mining method in actual applications, except when the system is used for the first time and then adaptively, say, after a successful login. In any case, even the use of the slowest method here was fast enough. Of course, additional computation is needed before the data mining phase to perform signal analysis. Still, this is also very fast, because its time complexity is linear and the length of eye movement signals is short, no more than a few thousand samples, say 1-3 minutes. Consequently, the running time would be minimal compared to such a recording time. At the beginning, in the course of a recording the eye movement camera system also makes image processing, but this is also close to real time. The camera system used consisted of only an initial calibration when taken into use. Thus, calibration required no additional processing time here.

VI. CONCLUSION

The MLP networks produced their best results with Alternative 2, 2 output nodes, 6 hidden nodes in Table I and 10 hidden nodes in Table II. The use of normalization did not improve the results obtained which were around 8% poorer than the best of SVMs and RBFs in Tables III-V. The Gaussian kernel was the best choice with SVMs. RBFs were very sensitive to normalization needed apart from the other being very insensitive to normalization.

The best results obtained were fairly good as 89% of the best results in Tables IV and V. We may assess that the best realistic accuracies based on various biometric verification references are around 95%. Thus, the results of this quite novel way to perform a biometric verification task are promising although more research has to be made to improve verification accuracies. A clear chance here is to collect a larger set of recordings from each individual. There were only five recordings with four large saccades per a subject. Forming a larger training set from each subject than now it is quite probable that we are able to improve classification results based on data mining methods. To compare with other scarce results presented thus far, our results were equal or better than various values 50-90% given in [11, 12].

The eye movement camera system used included a low sampling frequency of 30 Hz (frames per second). Still, verification was fairly successive. The low sampling frequency was, however, interesting since it was similar to that often used in cheap web cameras. We may except that in the future eye movement cameras are installed in computers or mobile devices to follow a user's gaze for various human-computer interface tasks [13]. If their sampling frequencies will be higher, e.g., 200 Hz, biometric verification with eye movements may well be realistic.

ACKNOWLEDGEMENT

The authors thank prof. Ilmari Pyykkö, M.D., from the Department of Otorhinolaryngology, Tampere University Hospital, Finland, for advice subject to eye movements. The first author acknowledges the support given by Tampere Doctorial Program in Information Science and Engineering.

REFERENCES

[1] X. Tan, B. Bhanu, and Y. Lin, "Fingerprint classification based on learned features," IEEE Trans. Syst. Man Cybern, Part C: Appl. Rev., vol. 35, no 3, pp. 287–299, August 2005.

[2] V. Conti, C. Militello, F. Sorbello, and S. Vitabile, "A frequency-based approach for features fusion in fingerprint and iris multimodal biometric identification systems," IEEE Trans. Syst. Man Cybern, Part C: Appl. Rev., vol. 40, no 4, pp. 384–395, July 2010.

[3] K. Venkataramani, S. Qidwai, and B. V. K. Vijayakumar, "Face authentication from cell phone camera images with illumination and temporal variations," IEEE Trans. Syst. Man Cybern, Part C: Appl. Rev., vol. 35, no 3, pp. 411–418, August 2005.

[4] Z. Sun, Y. Wang, T. Tan, and J. Cui, "Improving iris recognition accuracy via cascaded filters," IEEE Trans. Syst. Man Cybern, Part C: Appl. Rev., vol. 35, no 3, pp. 435–441, August 2005.

[5] R.W. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," IEEE Computer, vol. 33, no 2, pp. 64–69, February 2000.

[6] M. Juhola, "A syntactic method for analysis of sacadic eye movements," Pattern Recogn., vol 19, pp. 353-359, 1986.

[7] M. Juhola, H. Aalto, and T. Hirvonen, "Using results of eye movement signals in the neural network recognition of otoneurological patients," Comp. Meth. Progr. Biomed.. vol 86, pp. 216-226, 2007.

[8] S. Usui and I. Amidror, "Digital low-pass differentation for biological signal processing," IEEE Trans. Biomed. Eng., vol. 29, pp. 686-693, 1982.

[9] S. Haykin, Neural Networks, A Comprehensive Foundation, Second Edition, Prentice Hall, 1999.

[10] H. Demuth, M. Beale and M. Hagan, Neural Network Toolbox™ 6 User's Guide, The MathWorks, 2009.

[11] P. Kasprowski and J. Ober, "Eye movements in biometrics," Biometric Authentication: Int. Workshop, Springer-Verlag, LNCS, vol. 3087, pp. 248-258, 2004.

[12] O. V. Komogortsev, S. Jayarathna, C. R. Aragon, and M. Mahmoud, "Biometric identification via an oculomotor plant mathematical model," Proc. 2010 Symp. Eye Tracking Research & Applications, pp. 57-60, 2010.

[13] A. Holzinger, R. Geierhofer, and G. Searle, "Biometric signatures in practice: A challenge for improving human-computer interaction in clinical workflows," A. M. Heinecke, H. Paul (Eds.), Mensch & Computer, Oldenbourg Verlag, München, Germany, pp. 339-347, 2006.

 ISBN: 978-1-61208-227-1

Extracting Transportation Information and Traffic Problems from Tweets during a Disaster

Where do you evacuate to?

Aya Ishino, Shuhei Odawara, Hidetsugu Nanba, Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
Hiroshima, Japan
{ishino, odawara, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

***Abstract*—In a disaster, one of the most important issues for victims is how to find evacuation routes to safety from hazardous areas. To offer such routes, we propose methods automatically extracting transportation information and traffic problems from tweets written in Japanese and posted during a disaster. To investigate the effectiveness of our methods, we conducted some experiments using tweets posted during the Great Eastern Japan Earthquake in March 2011. From the experimental results, we obtained precision of 78.2% and recall of 53.4% in automatic extraction of transportation information. For extracting traffic problems, we identified tweets containing relevant information (we call them traffic problem tweets), and extracted traffic problem from them. In identifying traffic problem tweets, we obtained precision of 77.7% and recall of 70.7%. In extracting traffic problems, we obtained precision of 87.0% and recall of 57.1%. Thus, we have constructed a system for providing transportation information and traffic problems in a disaster.**

Keywords-disaster; evacuation routes; information extraction.

I. INTRODUCTION

Disasters occur frequently throughout the world. For instance, there were the large earthquakes in Haiti in January 2010 and in Sumatra, Indonesia in December 2004. In March 2011, a massive earthquake of magnitude 9.0 struck off the coast of eastern Japan. This earthquake is called the Great Eastern Japan Earthquake. It caused tsunamis and an accident at a nuclear power plant, and forced large numbers of people to evacuate from their homes and towns. In such disasters, one of the most important issues for victims is how to find evacuation routes to safety from hazardous areas. To offer such evacuation routes, there is a need to collect transportation information and traffic problems from other victims. Because they are so widely used, we focused on extracting information from tweets.

After the Great Eastern Japan Earthquake, 18 million tweets were posted on Twitter in a day, which is 1.8 times as much as normal. Some tweets contained useful information about transportation and traffic problems. In this paper, we propose methods for extracting transportation information and traffic problems automatically from tweets posted during disasters. In addition, we construct a system for presenting the extracted information. We believe that the system can offer safe evacuation routes for disaster victims and transportation routes for relief materiel.

The remainder of this paper is organized as follows. Section II describes the system behavior using snapshots. Section III describes related work. Section IV explains our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section V reports on these and the results. We present some conclusions in Section VI.

II. SYSTEM BEHAVIOR

In this section, we describe our prototype system, which (1) provides transportation information, and (2) identifies traffic problems. Fig. 1 shows transportation information from disaster victims. Arrows with icons indicate transportation information. The arrow extends from a departure place (shown as ① in Fig. 1) to a destination (②). Each icon depicts a transportation method. If the user clicks the icon (③), the system shows details of transportation information (④). The user can discover that a disaster victim evacuated from Ishinomaki City to Ichinoseki City by car.

Fig. 2 shows traffic problems. An arrow with an icon indicates a traffic problem. A traffic problem is indicated by an arrow with an icon. The arrow shows that a traffic problem has occurred between one end of the arrow (shown as ① in Fig. 2) and the other end the arrow (②). If the user clicks the icon (③), the system shows details of the traffic problem (④). The user can discover that a traffic problem has occurred between Sendai City and Yamagata City on Route 48. In this paper, we describe the methods used by our system for extracting transportation information and traffic problems.

III. RELATED WORK

In this section, we describe some related studies on information mining in disasters and extracting transportation information.

A. Information Mining in Disasters

In time of disaster, vast amounts of data are generated via computer-mediated communication; however, it is difficult to extract useful information from them for users. There have been some studies of information mining in several disasters.

Verma *et al*. [1] collected tweets from four different disasters, and automatically detected tweets that could contribute to situational awareness automatically. They obtained over 80% accuracy.

Sakaki *et al.* [2] considered each Twitter user as a sensor, and detected disaster events based on sensory observation.

 ISBN: 978-1-61208-227-1

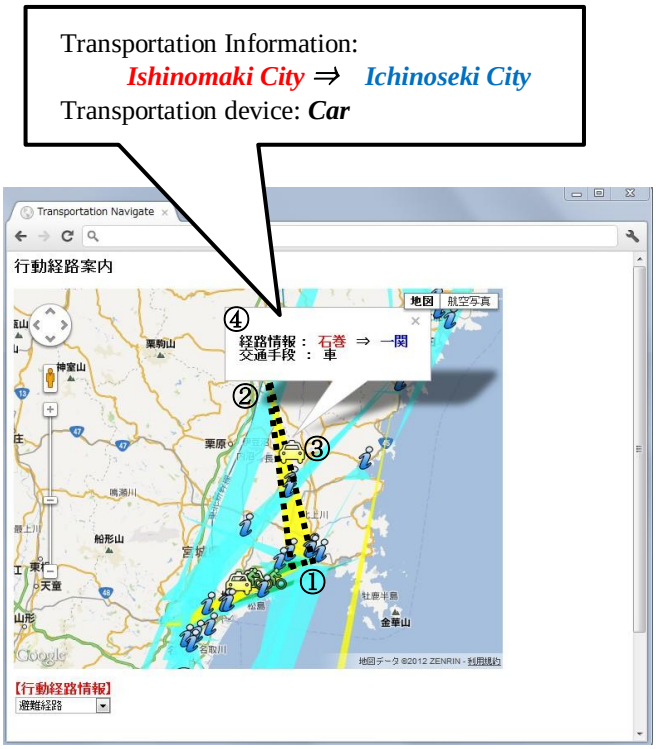


Figure 1. The system that provides transportation information.

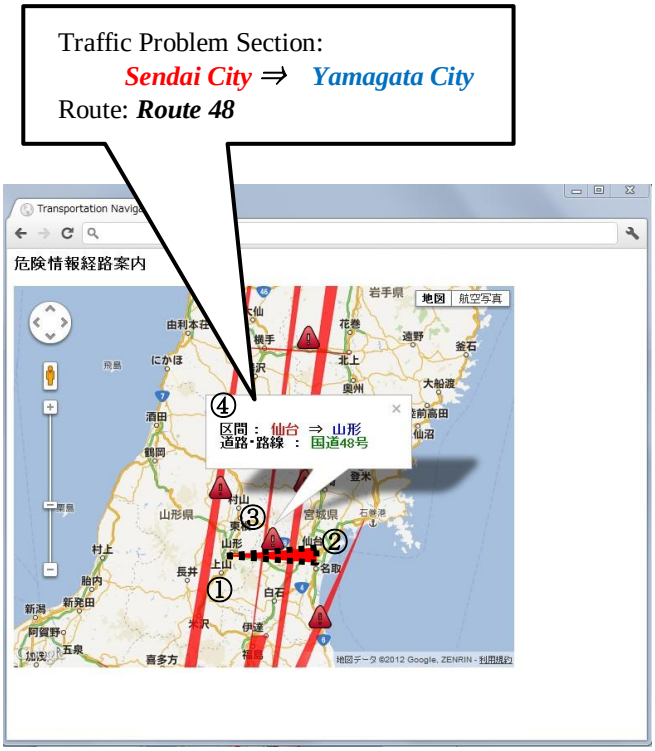


Figure 2. The system that provides traffic problem.

They targeted disaster events such as earthquakes and typhoons. As an application, they constructed an earthquake reporting system.

There are some studies about evacuation in disasters. Iwanaga *et al.* [3] build an earthquake evacuation ontology from twitter and provided the most suitable evacuation center. Troung *et al.* [4] presented a novel framework that manages and provides various types of context information required for adapting processes in emergency management systems.

Soon after the Great Eastern Japan Earthquake, many Natural Language Processing (NLP) researchers, engineers, and students from all over Japan created a working group, called "ANPI_NLP." "ANPI" means "safety" in Japanese. ANPI_NLP tried to collect tweets with hash tags, such as "#anpi (safety information)" or "#hinan (evacuation)", and extracted information about the safety of people. [5] In this paper, we use the tweet corpus provided by ANPI_NLP, and extract transportation information and traffic problems from it.

B. *Extracting Transportation Information*

There have been a number of studies of extracting transportation information. Davidov [6] presented an algorithm framework that enabled automated acquisition of map-link information from the Web based on surface patterns such as "from X to Y." Given a set of locations as initial seeds, they retrieved an extended set of locations from the Web and produced a map-link network that connected these locations using transport-type edges. In this paper, we propose a method for extraction of transportation information via machine-learning techniques.

Ishino *et al.* [7] extracted traveler's transportation information automatically from travel blog entries written in Japanese using machine-learning techniques. They used cues related to travel, such as "観光" (sightseeing tour) or "旅行" (travel) for machine learning. In this paper, we aim to extract transportation information from disaster victims. Therefore, we collect cues related to disasters for machine learning.

IV. EXTRACTING TRANSPORTATION INFORMATION AND TRAFFIC PROBLEMS

In this paper, we propose methods for extracting transportation information and traffic problems from tweets written in Japanese and posted during the Great Eastern Japan Earthquake. We explain our methods for extracting transportation information in Section IV-A, and for traffic problems in Sections IV-B and IV-C.

A. *Extracting Transportation Information*

In this section, we describe our method for extracting transportation information from tweets. We use machine learning to extract information, such as "a departure place", "a destination", or "a transportation method", from tweets. First, we define the tags used in our examination. Fig. 3 is a tagged example.

- FROM tag includes a departure place.
- TO tag includes a destination.
- METHOD tag includes a transportation method.

 ISBN: 978-1-61208-227-1

[Original]
(Tweet 1)
私の祖母は<FROM>山田岡</FROM>在住です。津波被害はなくガラスが数枚割れたと聞きました。避難勧告を受けて、楢葉の<METHOD>バス</METHOD>で<TO>いわき市の草野中学校</TO>に避難しています。
(Tweet 2)
義弟の安否確認が取れました。<FROM>石巻</FROM>から<METHOD>徒歩</METHOD>で<TO>仙台市内</TO>の家まで帰って来たそうです。

[Translation]
(Tweet 1)
My grandmother lives in <FROM>Yamadaoka</FROM>. The tsunami caused little damage there. She was urged to evacuate, and went to <TO>Kusano junior high school in Iwaki City</TO> by <METHOD>bus</METHOD>.
(Tweet 2)
I found out my brother-in-law is safe. He came back home in <TO>Sendai City</TO> from <FROM>Ishinomaki City</FROM> on <METHOD>foot</METHOD>.

Figure 3. Examples of tagged tweets.

We formulate the identification of the class of each word in a given sentence and solve it using machine learning. For the machine-learning method, we opted for the Conditional Random Fields (CRF) method [8]; its empirical success has been reported recently in natural language processing. The CRF-based method identifies the class of each entry. Features and tags are used in the CRF method as follows: (1) k tags occur before a target entry; (2) k features occur before a target entry; and (3) k features follow a target entry. We used the value k = 6, which was determined via a pilot study. We used the following 13 features for machine learning. A sequence of nouns (a noun phrase) was treated as a noun. We used MeCab [9] as a Japanese morphological analysis tool to identify the part of speech.

- A word.
- Its part of speech.
- Whether the word is a quotation mark.
- Whether the word is a cue phrase, as shown in Table I.

B. Identifying Traffic Problem Tweets

We proposed a method for extracting traffic problems from tweets posted during the Great Eastern Japan Earthquake. In a pilot study, we investigated the number of tweets containing traffic problem information (we call them traffic problem tweets) and found some examples. Therefore, this task is divided into two steps: (1) identifying traffic problem tweets from a tweet corpus; and (2) extracting traffic problems from the traffic problem tweets. We explain Step 1 in this section and Step 2 in Section IV-C.

In this section, we explain our method for identification of traffic problem tweets automatically. Fig. 4 shows examples of traffic problem tweets. They contain words related to traffic problems, such as “通行止め” (closed to traffic) or “停止” (shut down), and the name of the relevant road. We employed Support Vector Machine (SVM) [10] as a machine-learning technique to identify traffic problem tweets. We use the following features for machine learning.

TABLE I. CUE PHASES FOR EXTRACTION OF TRANSPORTATION INFORMATION

Tag	Cue phase	The number of cues
FROM	Whether the word is a cue that often appears immediately after the FROM tag, such as “から” (from) or “を出発” (left).	5
FROM TO	Whether the word is frequently used in the name of a shelter, such as “学校” (school) or “公民館” (community center).	23
	Whether the word is a cue that the FROM tag and the TO tag do not contain, such as “方向” (directions) or “沿い” (along).	7
	Whether the word is the name of a station, provided by ANPI_NLP.	8619
	Whether the word is the spot name in eastern Japan, provided by ANPI_NLP.	1755
	Whether the word is the name of a school in eastern Japan, provided by ANPI_NLP.	806
	Whether the word is the name of a train line, provided by ANPI_NLP.	569
TO	Whether the word is a cue that often appears immediately after the TO tag, such as “まで” (to) or “へ避難” (evacuate).	30
METHOD	Whether the word is a cue that often appears immediately after the METHOD tag, such as “で行く” (by).	19
	Whether the word is the name of a transportation device, such as “飛行機” (airplane) or “自動車” (car).	37

 ISBN: 978-1-61208-227-1

- The word relates to traffic problem, such as “通行止め” (closed to traffic) or “停止” (shut down) (19).
- The word relates to a road, such as “自動車道” (Expressway) or “インターチェンジ” (Interchange) (13).
- The word relates to transportation devices, such as “新幹線” (Shinkansen bullet train) or “地下鉄” (subway) (9).

C. Extracting Traffic Problems

In this section, we explain our method for extracting traffic problems from traffic problem tweets identified in Section IV-B. We use machine learning to extract information such as “a road” or “a train line”, or “traffic problem section”, from tweets. First, we define the tags used in our examination. Fig. 5 is a tagged example.

- LINE tag includes a road or a train line.
- LOC tag includes a traffic problem section.

We use CRF for the machine learning. Features and tags are used in the CRF method as follows: (1) k tags occur before a target entry, (2) k features occur before a target entry, and (3) k features follow a target entry. We used the value k = 6, which was determined via a pilot study. We use the following 14 features for machine learning. A sequence of nouns (a noun phrase) was treated as a noun. We used MeCab as a Japanese morphological analysis tool.

- A word.
- Its part of speech.
- Whether the word is a quotation mark.
- Whether the word is a mark, such as “~”, or “→”.
- Whether the word is a cue phrase, as shown in Table II.

[Original]
(Tweet 1)
地震で中央自動車道も上野原―勝沼インターチェンジ間などが通行止め。
(Tweet 2)
新幹線 浜松～品川停止中。

[Translation]
(Tweet 1)
After a large earthquake, the Chuo Expressway is closed to traffic between Uenohara city and the Katsunuma Interchange.
(Tweet 2)
Shinkansen (Bullet Train) are shout down. Hamamatsu – Shinagawa

Figure 4. Example of traffic problem tweets.

[Original]
(Tweet 1)
地震で<LINE>中央自動車道</LINE>も<LOC>上野原</LOC>―<LOC>勝沼インターチェンジ</LOC>間などが通行止め。
(Tweet 2)
<LINE>新幹線</LINE> <LOC>浜松</LOC>～<LOC>品川</LOC>停止中。

[Translation]
(Tweet 1)
After a large earthquake, <LINE>the Chuo Expressway</LINE> is closed to traffic between <LOC>Uenohara city</LOC> and <LOC>the Katsunuma Interchange</LOC>.
(Tweet 2)
<LINE>Shinkansen (Bullet Train)</LINE> are shout down. <LOC>Hamamatsu</LOC> – <LOC>Shinagawa</LOC>

Figure 5. Example of tagged traffic problem tweets.

TABLE II. CUE PHASE FRO EXTRACTION OF TRAFFIC PROBLEMS

Tag	Cue phase	The number of cues
LINE	Whether the word is frequently used in the name of a road and a train line, such as “道路” (road) or “号線” (line).	23
	Whether the word is the name of a train line, provided by ANPI_NLP.	569
	Whether the word is the name of a bypass, collected from Wikipedia.	1301
	Whether the word is the name of an express way.	60
	Whether the word is the name of a toll road.	181
LINE LOC	Whether the word frequently used in traffic problems, such as “通行止め” (closed to traffic) or “停止” (shout down) .	51
	Whether the word frequently used in traffic problems, such as “通行可能” (available for traffic) or “復旧” (restoration) .	19
LOC	Whether the word is the name of a station, provided by ANPI_NLP.	8619
	Whether the word is the spot name in eastern Japan, provided by ANPI_NLP.	1755
	Whether the word is a cue that often appears in the LOC tag, such as “駅” (station) or “インターチェンジ” (interchange).	10

 ISBN: 978-1-61208-227-1

V. EXPERIMENTS

A. Ectracting Transportation Information

Data Sets and Experimental Settings

We randomly selected 1303 tweets written in Japanese from the tweet corpus provided by ANPI_NLP, and tagged them manually, as described in Section IV-A. The numbers of manually assigned tags are shown in Table III. We used CRF++ [11] software as the machine-learning package. As a base line method, we used only a word as a feature for machine learning. We used precision and recall as evaluation measures, calculated as follows.

$$\text{Precision} = \frac{\text{The number of correctly extracted tags}}{\text{The number of tags that the system extracted}} \quad (1)$$

$$\text{Recall} = \frac{\text{The number of correctly extracted tags}}{\text{The number of tags that should be extracted}} \quad (2)$$

TABLE III. NUMBERS OF MANUALLY ASSIGNED TAGS IN THE EXTRACTED TRANSPORTATION INFORMATION

Tag	Training	Test
FROM	237	71
TO	425	120
METHOD	61	17
Total	723	208

Results and Discussion

The evaluation results are shown in Table IV. Our method obtained higher precision and recall than the baseline method. We discuss the experimental results as follows.

TABLE IV. EVALUATION RESULTS FOR EXTRACTING TRANSPORTATION INFORMATION

Tag	Our method		Baseline method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
FROM	78.4	40.9	72.4	29.6
TO	76.3	59.2	73.2	43.3
METHOD	91.7	64.7	80.0	23.5
Total	78.2	53.4	73.3	37.0

[Original]
(Correct) 助川小学校に避難された方がいらっしゃいましたら、現在どのような状況か情報頂きたいです！
(Analysis result) <TO>助川小学校</TO>に避難された方がいらっしゃいましたら、現在どのような状況か情報頂きたいです！

[Translation]
(Correct) If victims evacuate to Sukegawa Elementary School, please let me know what is going on!
(Analysis result) If victims evacuate to <TO>Sukegawa Elementary School</TO>, please let me know what is going on!

Figure 6. Example of a failure in extracting transportation information.

First, we discuss a typical error causing low precision. Fig. 6 shows an example of a failure in extracting transportation information. The TO tag was mistakenly assigned to "助川小学校" (Sukegawa Elementary School), which might not be an actual evacuation site. This was because the "TO" cue "避難" (evacuate) appears immediately before it. To improve the performance of extracting transportation information, we should consider language structure.

Next, we discuss a typical error causing low recall. A typical error is the lack of cues. In particular, we could not collect the names of some facilities or places cyclopedically. When preparing for a disaster, we must collect the names of facilities and places all over the world.

B. Identifying Traffic Problem Tweets

Data Sets and Experimental Settings

For our examination, we identified traffic problem tweets among 1750 tweets written in Japanese provided by ANPI_NLP. The number of manually identified traffic problem tweets is shown in Table V. We performed a four-fold cross validation test. We used a standard SVM package, TinySVM (http://chasen.org/~taku/software/TinySVM/). We used precision and recall as evaluation measures.

TABLE V. NUMBER OF MANUALLY IDENTIFIED TRAFFIC PROBLEM TWEETS

Traffic Problem Tweets	Others	Total
166	1584	1750

Results and Discussion

Table VI shows the experimental results. Our method obtained a higher recall than baseline method. We now discuss the low recall of our method. A typical reason for low recall is the lack of cues. For machine learning, we used manually selected cues, as described in Section IV-B. To improve the coverage of cues, a statistical approach, such as applying n-gram statistics to a larger tweet corpus, will be required.

TABLE VI. EVALUATION RESULTS FOR IDENTIFYING TRAFFIC PROBLEM TWEETS

	Precision (%)	Recall (%)
Our method	77.7	70.7
Baseline method	80.4	61.9

C. Extracting Traffic Problems

Data Sets and Experimental Settings

We manually assigned tags to traffic problem tweets, as described in Section IV-C, and used them for our examination. Table VII shows the numbers of manually assigned tags. We used CRF++ software as the machine-learning package. As a baseline method, we used only a word as a feature for machine learning. We used recall and precision as evaluation measures, calculated as equations (1) and (2).

 ISBN: 978-1-61208-227-1

TABLE VII. NUMBERS OF MANUALLY ASSIGNED TAGS FOR EXTRACTING TRAFFIC PROBLEMS

Tag	Training	Test
LINE	126	39
LOC	176	67
Total	166	243

Results and Discussion

The evaluation results are shown in Table VIII. Our method obtained higher recall than the baseline method.

We now discuss the low recall of our method. Errors in our method have been found in tweets that contain both problem information and safety information about traffic states. Fig. 7 shows an example of failure in the extracting traffic problems. In the example, the "LINE" tag should be assigned to "県内高速道路" (expressway in the prefecture), but our method did not assign any tags to this word. This tweet contains the cue "通行止め" (closed to traffic), and the cue "通行可能" (can pass). In this case, we should consider language modification relationships of cues.

TABLE VIII. EVALUATION RESULTS FOR EXTRACTING TRAFFIC PROBLEMS

Tag	Our method		Baseline method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
LINE	89.7	68.4	80.0	31.6
LOC	85.0	50.8	92.0	34.3
Average	87.0	57.1	87.5	33.3

[Original]
(Correct) 能代南ＩＣ－二ツ井ＩＣ間は通行可能。そのほかの<LINE>県内高速道路</LINE>通行止め。
(Analysis result) 能代南ＩＣ－二ツ井ＩＣ間は通行可能。そのほかの<line>県内高速道路</line>通行止め。

[Translation]
(Correct) You can pass between the Noshiro Minami Interchange and the Futatsui Interchange. Another <LINE>expressway in the prefecture</LINE> is closed to traffic.
(Analysis result) You can pass between the Noshiro Minami Interchange and the Futatsui Interchange. Another expressway in the prefecture is closed to traffic.

Figure 7. Example of a failure in extracting traffic problems.

VI. CONCLUSION

To offer evacuation routes to safety for disaster victims, we have proposed methods for extracting transportation information and traffic problems from tweets posted during disasters. To investigate the effectiveness of our methods, we conducted some experiments using tweets posted during the Great Eastern Japan Earthquake. From the experimental results, we obtained precision of 78.2% and recall of 53.4% in automatic extraction of transportation information. For extracting traffic problems, we identified tweets containing information about traffic problems (we called them traffic problem tweets), and extracted traffic problems from them. In identifying traffic problem tweets, we obtained precision of 77.7% and recall of 70.7%. In extracting traffic problems, we obtained precision of 87.0% and recall of 57.1%. Thus, we have constructed a system for providing transportation information and identifying traffic problems in disasters. We consider that the system can offer evacuation routes for disaster victims and transportation routes for relief materiel.

In this paper, we used tweets that posted during the Great Eastern Japan Earthquake. Therefore, we used the names of facilities or places in eastern Japan as cues for machine learning. When preparing for disasters anywhere, we must collect the names of a facilities or places all over the world.

In this paper, we focused on tweets written in Japanese. In our future work, we will translate cue phrases from Japanese into other languages, and apply our method to tweets written in various languages.

REFERENCES

[1] S. Verma, S. Vieweg, W. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, and K.M. Anderson, "Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets During Mass Emergency," Proc. the Fifth International AAAI Conference on Weblogs and SocialMedia, Barcelona, Spain, pp. 385–392, July 2011.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors," Proc. 19th International World Wide Web Conference, NC, USA, April 2010.

[3] I. Shizu Miyamae Iwanaga, T. M. Nguyen, T. Kawamura, H. Nakagawa, Y. Tahara, and A. Ohsuga, "Building an Earthquake Evacuation Ontology from Twitter," Proc. 2011 IEEE International Conference on Granular Computing, Kaohsiung, Taiwan, November 2011.

[4] H. L. Truong, L. Juszczyk, A. Manzoor, and S. Dustdar, "Escape - an Adaptive Framework for Managing and Providing Context Information in Emergency Situations," Proc. 2nd European Conference on Smart Sensing and Context, vol. 4793, pp. 207–222, Lake District, UK, October 2007.

[5] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety Information Mining — What can NLP do in a disaster —," Proc. 5th International Joint Conference on Natural Language Processing, pp. 965-973, Chiang Mai, Thailand, November 2011.

[6] D. Davidov, "Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns," Proc. 2009 Conference on Empirical Methods in Natural Language Processing, pp. 267-275, Singapore, August 2009.

[7] A. Ishino, H. Nanba, and T. Takezawa, "Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries," Proc. 18th International Conference on Information Technology and Travel & Tourism, Innsbruck, Austria, January 2011.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. 18th Conference on Machine Learning, pp. 282-289, MA, USA, June 2001.

[9] http://mecab.sourceforge.ne

[10] C. Corinna, and V. Vladimir, "Support-Vector Networks," J. Machine Learning, vol.20, No.3, pp. 273-297, 1995.

[11] http://www.chasen.org/~taku/software/CRF++

 ISBN: 978-1-61208-227-1

Classification of Time-Interval and Hybrid Sequential Temporal Patterns

Mohammed GH. I. AL Zamil
Department of Computer Information Systems
Yarmouk University
Irbed, Jordan
Mohammedz@yu.edu.jo

***Abstract*— Due to the rapid growth of information systems that manage temporal data, efficient and automated classification techniques are of great importance. For instance, timely and accessible temporal data enhances critical financial operations such as predicting future stock prices. Similarly, in medical domain, classifying temporal data, which is relevant to patients or critical operations, leads to efficient control and recovery from severe problems. Therefore, time is an essential dimension to many domain-specific problems. This research introduces Temporal-ROLEX; a framework to categorize temporal data that effectively induces semantic temporal patterns. This paper presents an efficient rule-based classification approach for categorizing temporal data. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns. The proposed framework extends ROLEX-SP approach to handle the classification of temporal data in different domains. To illustrate the design, the article provides a detailed mathematical description that relies on set-theory to model the framework of Temporal-ROLEX. Furthermore, this paper provides a detailed description of proposed algorithms to facilitate implementing and reproducing the results. To evaluate the effectiveness of the Temporal-ROLEX, we performed extensive experiments on a weather temporal dataset. Also, the F-measure and support values on weather dataset are reported as well as a scalability and sensitivity analysis to assess the capability of Temporal-ROLEX to work with temporal datasets. Findings indicate a significant improvement of Temporal-ROLEX over some existing techniques. Specifically, Temporal-ROLEX achieves significant enhancement using sequential temporal pattern over existing state-of-the-art techniques. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns. Finally, the results have been analyzed and justified the factors that affect the performance in both cases.**

Keywords-Temporal Data Analysis; Classification of Temporal Data; Lexical Patterns.

I. INTRODUCTION

Due to the rapid growth of information systems that manage temporal data, efficient and automated classification techniques of temporal data are of great importance. Time is an essential dimension to many domain-specific problems such as financial and medical domains. This paper presents an efficient rule-based classification approach for categorizing temporal data. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns.

ROLEX-SP has been introduced by M. ALZamil and A. Can [1] to categorize domain specific knowledge using specialized rule-based induction and learning methods to produce efficient classification of domain specific knowledge. ROLEX-SP automates the induction and the learning processes by extracting lexical patterns and constructing specialized form of association rules. Such technique handles the problems of multiclass classification and feature imbalance problems. This research introduces Temporal-ROLEX; a framework to categorize temporal data that effectively induces semantic temporal patterns

Temporal-ROLEX is intended to find temporal relationships such as: during, after, overlap, start, finish and equal. However, it defines a form of association rules that generate not only efficient patterns to classify events, but also minimize the margin error to enhance the overall performance of the classification task. Figure 1 shows during temporal event, in which event e1 starts and finishes during the execution of event e2. The work in this paper is restricted to during relationship, since it is able to represent hybrid relations among temporal events. In other words, a hybrid relationship is able to describe before and after relationships.

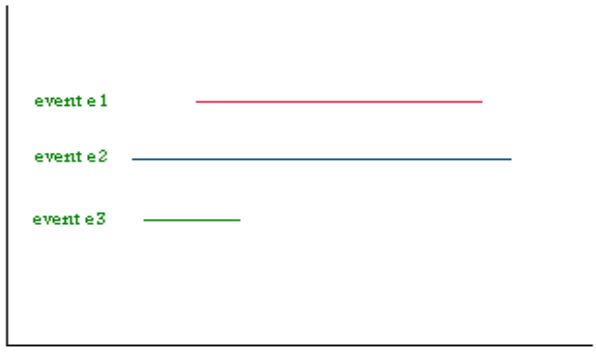


Figure 1. During sequential pattern (i.e., events e1 starts and finishes during the execution of e2)

 ISBN: 978-1-61208-227-1

This paper is structured as follows: Section I introduces the problem understudy and proposes the direction of the research solution. Section II provides a literature review of strictly related work of other researchers and compares them with the proposed method. Section III defines the formal model of the proposed methodology as a background to theoretical and empirical work. Section IV details the framework that is followed to classify temporal data and the algorithm that has been applied to produce the empirical results. Section V provides a description of the experimental setup. Finally, Section VI discusses the conclusion of this work and justifies the results.

II. RELATED WORK

Attempts have been made to construct temporal features in order to construct association rules such as those discussed in Bruno and Garza [2], Miao et al. [3], and Chiang et al. [4]. In Bruno and Garza [2], association rules have been developed to cope with outlier detection using functional quasi dependency. The technique does not model time-delay as a part of association rules. The delta function that associate two temporal attributes X and Y assumes no delay. The technique in Bruno and Garza [2] handled time-delay explicitly, which affect the overall performance as well as efficiency of the classification process; which is not crucial in outlier detection task.

Chiang et al. [4] have proposed a mathematical model to extract temporal patterns to track customer buying habits. The model is developed to capture temporal characteristics of business data in single-point-of-time events. Our proposed methodology focuses on time intervals as well as single point of time events. Similarly, our proposed technique benefits from the formal definition in [4], in that we formulate the temporal patterns using similar mathematical aspects.

Zhang et al. [5] have proposed a method to extract *during* temporal patterns. A during temporal pattern (DTP) is a special case of interval temporal patterns. These patterns provide valuable information in broadcasting future information such as weather and stock broadcasting. Kong et al. [6] have presented the notion of multi temporal patterns using predicates: before, during, equal and overlap.

Winarko and Roddick [7] have Introduced ARMADA, an algorithm to discover interval time temporal rules. ARMADA research asserts that time-stamps relationships such as *during* could be more useful than solid time interval. Classifying time-interval events into temporal clusters provide meaningful information in different application areas such as financial analysis and weather broadcasting. Unlike ARMADA association rules, our work relies on discovering hybrid temporal rules that could be represented using *during* relation.

Although performance plays a significant role in assessing classification techniques, pre-processing tasks might be crucial in many applications in terms of scalability and efficiency. However, temporal datasets dimensions are characterized as huge ones. Techniques to reduce such dimensionality are important to produce scalable temporal mining systems. We applied methods in Stacey and McGregor [8] and Wang and Megalooikonomou [9] to reduce the dimensionality of time series.

In the literature, there are many data mining and knowledge discovery techniques on medical domain and biomedical data [10, 11, 12, 13]. The contribution of this article over existing ones is the ability of Temporal ROLEX to handle timely information regardless of its domain. Further, the method proposed in this article deals with time as a classification feature. The later might negatively affect the classification performance, but it adds the advantage of enhancing timely information classification

III. BACKGROUND

For the purpose of defining the formal model of the temporal classification problem, Inductive-Logic-Programming (ILP) [14] is used as follows: given

1. A finite set TC of unrelated temporal classes of the form $\{Tc_1, Tc_2, ..., Tc_k\}$ where $k > 1$, meaning that there are many temporal classes and the assigned label of a class do not affect the labeling of other classes. For instance, an event might be classified under *during* class and *overlap* class at the same time if this event belongs to different set of events.
2. A set $E = \{e_1, e_2, \ldots, e_n\}$ of events such that $\forall(j)\exists(\ Q \subseteq TC \wedge |\ Q\ | = v) : e_j \in Q$ where $1 \le v \le k$ and $1 \le j \le N$, meaning that an event might belong to more than one temporal class; Q is a subset of the set of temporal classes.
3. A set of states $S = \{s_1, s_2, \ldots, s_m\}$ each of which represents a state of the current environment such as: *raining and shining* in the weather dataset.
4. A set of time-intervals $T = \{t_1, t_2, \ldots, t_n\}$, where $t_i = \{st_i, et_i\}$ represents the start and end time of a given event e_i
5. A set P_{ci}^{+} of positive patterns consisting of atomic facts of the form $p_{ci}^{+} \in E_{Tci}$ such that $(p_{ci}^{+} \in e \wedge e \in E_{Tci}) \Rightarrow e \in Tci$; a positive pattern under class Tci that occurs in the subset E_{Tci}, which represent a set of events that belong to the class Tci.
6. A set P_i^{-} of negative data patterns; patterns that represent an event but does not refer to

class Tci. In other words, they represent outliers or rare cases.

7. The function $[g(a_\alpha) = \{ e_1(a_\alpha, t_1), e_2(a_\alpha, t_2), \ldots, e_k(a_\alpha, t_k)\}]$ includes all the interval times in which the state a_α occurs.

construct a classifier H_{ci} that consists of all positive and negative facts. In other words, the classifier represents a set of association rules to forecasting a temporal class or a set of temporal classes of a given set of events based on the presence or absence of some facts in that set.

The learning task of Temporal-ROLEX generates association rules such that: given a category $Tc_i \in TC$, a positive pattern $p^+_{Tci} \in P_{Tci}{}^+$ associates with class Tci, and a set of negative patterns $P_i^- (P^- \cap P^+ = \phi)$, where P^- is the set of all negative patterns and P^+ is the set of positive patterns, the classifier H_{Tci} of class Tci is defined as a set of rules. We used the rule's representation in [15] as follows:

$$\left[Tc_i \leftarrow p^+_{Tci} \in g(a_\alpha),\ \neg(p^-_{i1} \in g(a_\alpha)) \wedge \neg(p^-_{i2} \in g(a_\alpha)) \wedge \ldots \wedge \neg(p^-_{im} \in g(a_\alpha))\right]$$

If a positive example p^+_{ci} occurs in document $g(a_\alpha)$ and none of the negative patterns occur in $g(a_\alpha)$, the classifier will assign event e under class Tci. Notice that, negative patterns are prevented from undoing the effect of other categories' positive ones.

IV. FRAMEWORK

Let $e_j = \{s_i, t_j\}$ and $e_k = \{a_l, t_k\}$ be two events in the temporal dataset. Both e_j and e_k are called during events if e_j has executed during the execution of e_k. For any two given states a_i and a_k, a_i is called to be during a_k denoted as $a_i \Rightarrow^d a_k$. Our goal is to define a set of positive and negative predicates to predict during temporal patterns.

Instead of the accuracy formula that has been applied in the previous version of ROLEX-SP, the function support that has been defined in [16] has been used to induce positive and negative patterns as well. Given $|g(a_\alpha)|$, the number of the time intervals included in all instances (records in the dataset) of a_α, the maximum number of time intervals among all states $|g_0|$:

$$Support(a_\alpha) = \frac{|g(a_\alpha)|}{|g_0|} \quad (1)$$

It represents the relative frequency of time intervals for a given state with respect to the number of time intervals for a most frequent state.

The proposed induction algorithm is shown in Figure 2.

LSP_Generator

- **Goal**: to extract positive and negative syntactic patterns from the set TS
- **Input**: TS, C.
  - **TS**: the set of labeled training set of data records
  - **C**: the set $C = \{c_1, c_2, \ldots c_k\}$ of categories where $k > 1$.
- **Output**: P^+, P^-
  - $P^+ = (P^+_{c1}, P^+_{c2}, \cdots, P^+_{ck})$ A set of positive patterns relevant to all categories.
  - $P^- = (P^-_1, P^-_2, \cdots, P^-_k)$ A set of negative patterns relevant to all categories.
- **Method**: Apply the following instructions

```
Begin
1     P+ = (), P- = ()
2     For each ci ∈ C
3       For each d ∈ TSc
4         P = parse(d, Rec(ci))
5         For each p ∈ P
6               Support(a_α) = |g(a_α)| / |g_0|
                if Support(a_α) ≥ threshold then
7                 P+_ci = P+_ci ∧ p
                Else
                  P-_i = P-_i ∧ p
8         Next p
9       Next d
10    Next ci
11    return(P+, P-)
End
```

Figure 2. Induction Algorithm

V. EXPERIMENT AND ANALYSIS

In this section, the results have been collected from applying Temporal-ROLEX on a weather dataset. The data set has been obtained from a weather station in Jordan in 2009. The dataset consists of 14 attributes: wind direction, average wind speed, maximum wind gust, average hourly temperature, percentage relative humidity, global hourly radiation, hourly sunshine duration, hourly precipitation duration, hourly precipitation amount, horizontal visibility, fog, snow, etc. Most of the collected values were continuous. To proceed, the pre-processing techniques to discriminate and convert the records into temporal ones have been applied, which are consisting of event name, start time, end time, and state.

A. F-Measure

First, we compute the recall and precision values relevant to every category according to the following formulas:

 ISBN: 978-1-61208-227-1

$$\mathrm{Pr} = \sum_{c \in C} |TP_c| / \sum_{c \in C} |TP_c| + |FP_c| \quad (2)$$

$$\mathrm{Re} = \sum_{c \in C} |TP_c| / \sum_{c \in C} |TP_c| + |FN_c| \quad (3)$$

where $|TP_c|$ is the number of correctly classified records in the testing set under category c, $|FP_c|$ is the number of incorrectly classified records in the testing set under category c, and $|FN_c|$ is the number of records in the testing set, which were not classified under category c but should have been. The F-measure is defined as follows:

$$F = \mathrm{Pr} \times \mathrm{Re} / (1-\alpha)\,\mathrm{Pr} + \alpha\,\mathrm{Re} \quad (4)$$

where $\alpha \in [0,1]$

$$Average\ F_{macro} = \sum_{i=1}^{|C|} \frac{F_i}{|C|} \quad (5)$$

where |C| is the number of categories in the dataset.

The results indicate that Temporal-ROLEX achieves 67.8% average F-Measure. The experiments show that Temporal-ROLEX achieves significant enhancement using sequential temporal pattern over existing state-of-the-art techniques on the same dataset such as DTP [16] that achieve 66.2%. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns; i.e., 63.7 while DTP achieve 65.1%. The results have been analyzed and justified the factors that affect the performance in both cases

B. *Sensitivity Analysis*

In order to evaluate the results, the analysis task considers four sensitivity attributes, which measure the quality of empirical results. These attributes include the number of valid patterns, the effect of rules on average F-measure, the execution time versus number of rules, and the execution time versus the number of events.

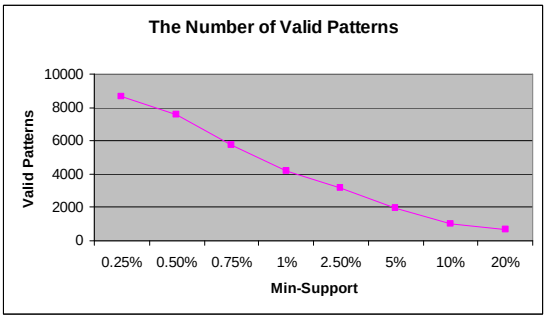


Figure 3. The Number of valid Patterns and Minimum Support

Figure 3 shows that Temporal-ROLEX performs well at low percentage of support measure. In other words, the rules induced using the proposed induction algorithm are able to classify valid patterns correctly among low number of time intervals in the training set.

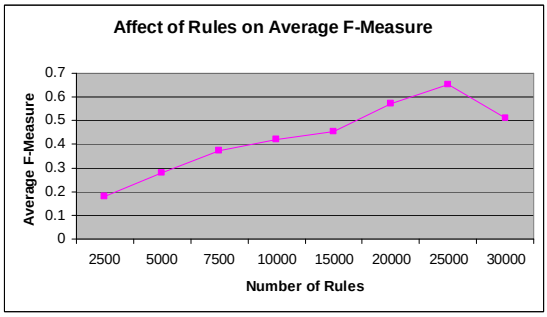


Figure 4. The Number of Rules and The average F-Measure

Figure 4 concludes a positive relationship between f-measure and the number of rules; the higher the number of rules, the higher the f-measure. This property demonstrates that in order to achieve high performance, the induction algorithm has to be fed with large training set to produce rules that cover all, or at least, most patterns.

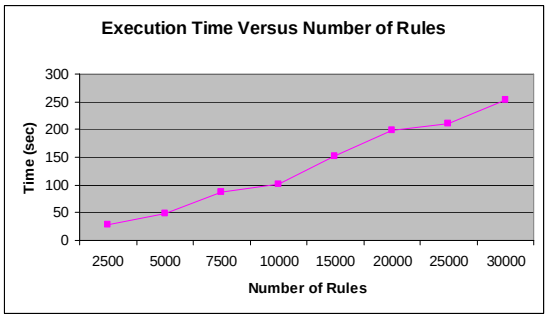


Figure 5. The Execution Time and The Number of Rules

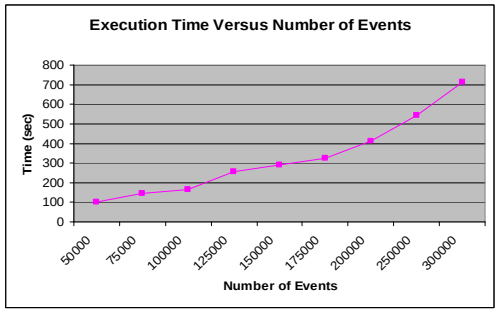


Figure 6. The Execution Time and Number of Events

Finally, Figures 5 and 6 show that the execution time increased as the number of rules and/or events increased.

VI. CONCLUSION

This paper presented a rule-based method for categorizing temporal records. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns. Experiments have been performed on a weather dataset in order to evaluate the proposed method and compare our work with well known algorithms in the literature. Findings indicate a significant improvement of Temporal-ROLEX over a well known technique; DTP. Specifically, Temporal-ROLEX achieve significant enhancement using sequential temporal pattern. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns.

Furthermore, Temporal-ROLEX achieved statistically significant improvement. Applying syntactic patterns, both positive and negative, enhances the accuracy of Temporal-ROLEX over the other method.

The article also provided a sensitivity analysis to the performance of Temporal-ROLEX as a function to the number of association rules and the number of data elements in the training set. The results indicated that Temporal-ROLEX was affected by the number of rules positively. On the other hand, the observations during experiments indicated that the number of records in the training set does not affect the overall performance of the learning process.

REFERENCES

[1] Al Zamil, M. and Betin-Can, A. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems 24 (2011) 58–65.

[2] Bruno, G. and Garza, P. TOD: Temporal outlier detection by using quasi-functional temporal dependencies. Data & Knowledge Engineering 69 (2010) 619–639

[3] Miao, O., Li, O., and Dai, R. AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications 36 (2009) 7192–7198.

[4] Chiang, D., Wang, Y., and Chen, S. Analysis on repeat-buying patterns. Knowledge-Based Systems 23 (2010) 757–768.

[5] Zhang, L., Chen, G., Brijs, T. and Zhang, X. Discovering during-temporal patterns (DTPs) in large temporal databases. Expert Systems with Applications 34 (2008) 1178–1189.

[6] Kong, X., Wei, O. and Chen, G. An approach to discovering multi-temporal patterns and its application to financial databases. Information Sciences 180 (2010) 873–885.

[7] Winarko, E. and Roddick, J. ARMADA – An algorithm for discovering richer relative temporal association rules from interval-based data. Data & Knowledge Engineering 63 (2007) 76–90.

[8] Stacey, M. and McGregor, C. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine (2007) 39, 1—24.

[9] Wang O. and Megalooikonomou, V. A dimensionality reduction technique for efficient time series similarity analysis. Information Systems 33 (2008) 115–132

[10] Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011) Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Proceedings of Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare. Dublin, IEEE, 550-554

[11] Holzinger, A., Simonic, K. M., andYildirim, P. (2012) Disease-disease relationships for rheumatic diseases. COMPSAC 2012. Izmir,Turkey

[12] Holzinger, A., Scherer, R., Seeber, M., Wagner, J., and Müller-Putz, G. (2012) Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm, C. (Ed.) International Conference on Information Technology in Bio- and Medical Informatics - ITBAM 2012 Heidelberg, Berlin, New York, Springer, 166-168

[13] Holzinger, A. (2012). On Knowledge Discovery and interactive intelligent visualization of biomedical data: Challenges in Human–Computer Interaction & Biomedical Informatics. DATA - International Conference on Data Technologies and Applications, Rome-Italy.

[14] Lavrac, N. and Dzeroski, S. Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (1994).

[15] P. Rullo, V. Policicchio, C. Cumbo, S. Iiritano, Olex: Effective rule learning for text categorization, IEEE Transactions on Knowledge and Data Engineering 21 (8) (2009) 1118‑1132.

[16] Wu, S. and Chen, Y. Discovering hybrid temporal patterns for interval-based events, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 742-758.

 ISBN: 978-1-61208-227-1

Optimized Class Association Rule Mining using Genetic Network Programming with Automatic Termination

Eloy Gonzales, Bun Theang Ong, Koji Zettsu
Information Services Platform Laboratory
Universal Communication Research Institute
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
Tel: +81-774-98-6866, Fax: +81-774-98-6960
e-mail: {egonzales, bt_ong, zettsu}@nict.go.jp

***Abstract*—Association rule mining is one of the tasks of data mining and it has been extensively studied during the last years. As a consequence, recently, several methods for extracting association rules have been developed. Some methods use Evolutionary Algorithms to extract association rules. Among them, a relatively new method using Genetic Network Programming (GNP) has been developed and its effectiveness has been shown, which outperforms other conventional algorithms. However, there still remain some issues mainly focused on performance.**

To improve the conventional GNP data mining algorithmic efficiency without loss of reliability, a GNP enhanced with an automatic termination criteria named AT-GNP is proposed in this paper. Indeed, in an effort to save computational resources, the objective is to stop the search right before unnecessary function evaluations are performed. The concept of Gene Matrix (GM) is used to direct the search and to stop it at a proper instant.

An extensive comparison between the conventional GNP-based association rule mining and AT-GNP is performed in the simulations to evaluate the performance. Finally, the association rules extracted using both methods are applied to the classification problems and the prediction accuracies of them are compared with other conventional approaches.

Keywords-Association rule mining; classification; evolutionary computation; genetic network programming; termination criteria.

I. INTRODUCTION

Among several methods of extracting association rules that have been reported, a relatively new Evolutionary Computation (EC) method named Genetic Network Programming (GNP) has also been developed recently and the effectiveness of applying it to the data mining is shown by several authors such as Gonzales et. al. [1][2][3] for diverse types of datasets.

However, it still suffers of performance issues especially concerned to the processing time. This is mainly due to the termination criteria which is basically after a fixed number of generation in the evolution of GNP. Empirically have been demonstrated by Shimada et. al. [4][5] that most of the association rules are extracted during the initial generations of the GNP, but the problem is to determine exactly when the algorithm has to terminate without loss of reliability.

There is not much work yet in the research of EAs dealing with the question of the termination criteria. Nevertheless, it is recognized that in many real world applications, saving computational resources is extremely important. There are only a few recent works on termination criteria for EAs. Giggs et. al. [6], empirically studied the problem characteristics in an attempt to determine the maximum number of generations. Kwok et. al. [8] used statistics to terminate the search when it is estimated that no further improvement in terms of solution quality can be expected. Jain et. al. [7] studied eight termination criteria with clustering techniques that examine the distribution of individuals in the search space at a given generation. Ong and Fukushima [9][10] introduced the concept of the Gene Matrix (GM). The GM is a matrix that represents subranges of the possible values of each variable. It gives an indication on the distributions of the variables over the search range. This information is used to provide the search with new diverse solutions and to let the search know how far the exploration process has been performed in order to terminate it. Our proposed method also takes advantage of the GM. The particularity of using GM when compared to the existing methods that deal with the question of the termination criteria (see above) is that the algorithm is expected to terminate without a priori knowledge of any desirable or available solution range, and of any specific number of iterations or function evaluations. Actually, the termination instant after completion of adequate exploration and exploitation is determined by the algorithm itself.

The aim of this paper is to extend the conventional GNP-based mining method [4][5] by using a variation of the GM to guide the search and the evolution of the GNP individuals. That is, a mechanism similar to the GM is applied to GNP. This mechanism ensures that all judgment and decision nodes have been mutually joined to each other at least a given number of times before terminating the search. Consequently, the diversity of the solutions is favored. Concurrently, the depth of the graph structure of GNP is not altered, thus preserving the quality of the final

 ISBN: 978-1-61208-227-1

solutions.

The following sections of this paper are organized as follows: In Section II, a brief description of association rules is presented; the outline of GNP is briefly reviewed in Section III, where also the enhanced method for rule extraction using GM is presented. Simulation results are described in Section IV, and finally, conclusion and future work are given in Section V.

II. Association Rules

Zhang et. al. [11] introduced a formal statement of the problem of mining association rules. Let $I = \{A_1, A_2, \ldots, A_l\}$ be a set of l distinct attributes. Let T be a transaction which contains a set of attributes such that $T \subseteq I$. D be a database with different transaction records T. A transaction T contains X, a set of some attributes in I, if $X \subseteq T$.

An association rule is an implication of the form of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. X is called antecedent and Y is called consequent of the rule. In general, the set of attributes X and Y are called itemsets.

There are two important basic measures for association rules, support and confidence. Support of an association rule $X \Rightarrow Y$ is defined as the percentage of records that contain $X \cup Y$ to the total number of records in the database.

Confidence of an association rule $X \Rightarrow Y$ is defined as the percentage of the number of transactions that contain $X \cup Y$ to the total number of records that contain X.

This measure indicates the relative frequency of the rule, that is, the frequency with which the consequent is fulfilled when the antecedent is also fulfilled .

However, the support-confidence framework has been shown not enough to extract interesting association rules, therefore, in this paper, the cosine correlation measure is used in addition to the conventional measurements of support and confidence.

Given two itemsets X and Y, the cosine measure is defined as:

$$cosine(X,Y) = \frac{P(X \cup Y)}{\sqrt{P(X)\ P(Y)}} = \frac{supp(X \cup Y)}{\sqrt{supp(X)\ supp(Y)}} \quad (1)$$

Cosine is a number between 0 and 1. A value close to 1 indicates positive correlation between X and Y. Cosine measure is also a null-invariant measure.

Therefore, the problem of mining class association rules is to find all rules that are highly likely to be interesting, that is, satisfying the minimum support, confidence and cosine thresholds.

$$\begin{aligned} support(X \Rightarrow Y) &\geq min_{supp}, \\ confidence(X \Rightarrow Y) &\geq min_{conf}\ , \text{and} \\ cosine(X \Rightarrow Y) &\geq min_{cosine} \end{aligned} \quad (2)$$

III. Genetic Network Programming

Genetic Network Programming (GNP), introduced by Hirasawa et. al. [12], [13], [14], is one of the evolutionary optimization algorithms, which evolves directed graph structures as solutions instead of strings (Genetic Algorithms) or trees (Genetic Programming). The main aim of developing GNP was to deal with dynamic environments efficiently by using the higher expression ability of graph structures.

The basic structure of GNP is shown in Fig. 1. The graph structure is composed of three types of nodes that are connected on a network structure: a start node, judgment nodes (diamonds), and processing nodes (circles). Judgment nodes are the set of J_1, J_2, …, J_p, which work as *if-then* conditional decision functions and they return judgment results for assigned inputs and determine the next node to be executed. Processing nodes are the set of P_1, P_2, …, P_q, which work as action/processing functions. The start node determines the first node to be executed. The nodes transition begins from the start node, however there are no terminal nodes. After the start node is executed, the next node is determined according to the node's connections and judgment results.

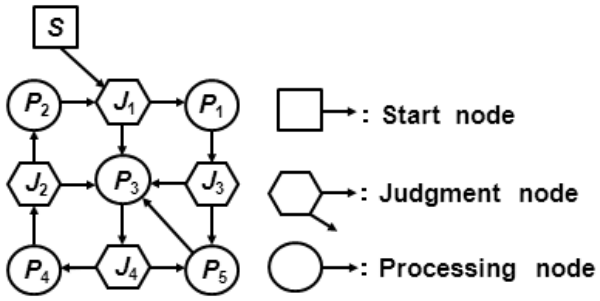


Figure 1. Basic structure of GNP

The gene structure of GNP (node i) is shown in Fig. 2. The set of these genes represents the genotype of GNP-individuals. NT_i describes the node type, $NT_i = 0$ when node i is the start node, $NT_i = 1$ when node i is a judgment node and $NT_i = 2$ when node i is a processing node. ID_i is an identification number, for example, $NT_i = 1$ and $ID_i = 1$ mean node i is J_1. C_{i1}, C_{i2}, …, denote the nodes, which are connected from node i firstly, secondly, …, and so on depending on the arguments of node i. d_i and d_{ij} are the delay time, which are the time required to execute the judgment or processing of node i and the delay time of transition from node i to node j, respectively. In this paper, the execution time delay d_i and the transition time delay d_{ij} are not considered.

A. Class Association Rule Mining using GNP

When GNP is applied to class association rule mining [4] [5], attributes of the dataset and their values correspond to the functions of judgment nodes in GNP-individuals. Association rules are represented as the connections of nodes.

 ISBN: 978-1-61208-227-1

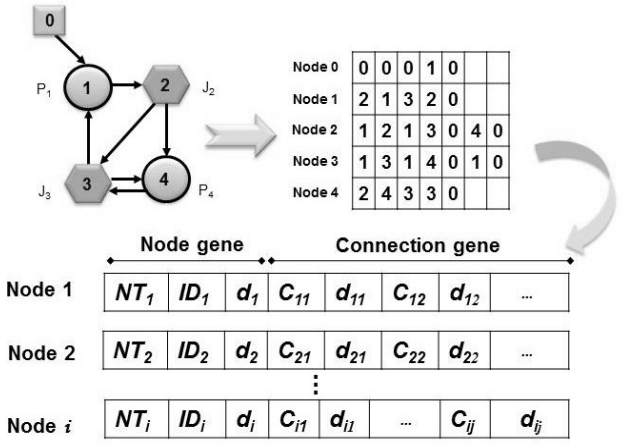


Figure 2. Gene structure of GNP (node i)

Candidate rules are obtained by genetic operations. Rule extraction using GNP is done without identifying frequent itemsets used in Apriori-like methods such as Agrawal et. al. [15] and stored in a pool through generations. The fundamental difference with other evolutionary methods is that GNP evolves in order to store new interesting rules in the pool, not to obtain the individual with the highest fitness value.

Let A_i be an attribute in a database and its value be 1 or 0, and C be the set of class labels. The method extracts the following association rules:

$(A_m = 1) \wedge \cdots \wedge (A_n = 1) \Rightarrow (C = k)$,
$(C = 0, 1, 2, \ldots, K)$

1) Genetic Operations: Changing an attribute to another one or adding some attributes in the rules would be considered as candidates of important rules. These rules can be obtained effectively by GNP genetic operations, because mutation and crossover will change the connections or contents of the nodes.

Three kinds of genetic operators are used for judgment nodes: GNP-crossover, GNP-mutation-1 (change the connections) and GNP-mutation-2 (change the function of nodes).

- GNP-Crossover: uniform crossover is used. Judgment nodes are selected as the crossover nodes with the probability of P_c. Two parents exchange the gene of the corresponding crossover nodes.
- GNP-Mutation-1: Mutation-1 operator affects one individual. The connection of the judgment nodes is changed randomly by mutation rate of P_{m1}.
- GNP-Mutation-2: Mutation-2 operator also affects one individual. This operator changes the functions of the judgment nodes by a given mutation rate P_{m2}.

On the other hand, all the connections of the processing nodes are changed randomly. At each generation, all GNP-individuals are replaced with the new ones by the following criteria: The GNP-individuals are ranked by their fitness values and the best one-third GNP-individuals are selected. After that, these GNP-individuals are reproduced three times for the next generation using the genetic operators described before.

If the probabilities of crossover (P_c) and mutation (P_{m1}, P_{m2}) are set at small values, then the same rules in the pool may be extracted repeatedly and GNP tends to converge prematurely at an early stage. If the probability of mutation is set at high values, then some genetic characteristics of the individuals might be lost. These parameter values are chosen experimentally avoiding these issues.

2) Fitness of GNP: The number of processing nodes and judgment nodes in each GNP-individual is determined based on experimentation depending on the number of attributes processed. The connections of the nodes and the functions of the judgment nodes at an initial generation are determined randomly for each GNP-individual. Fitness of GNP is defined by:

$$F = \sum_{r \in R} \{cosine(r) + \alpha_{new}(r) + \beta(n(r) - 1)\} \quad (3)$$

The terms in Eq. (3) are as follows:

R: set of suffixes of extracted important association rules satisfying the minimum support-confidence-correlation measure in a GNP individual

cosine(r): value of cosine correlation of rule r,

$\alpha_{new}(r)$: additional constant defined by

$$\alpha_{new}(r) = \begin{cases} \alpha_{new} & \text{(rule } r \text{ is new)} \\ 0 & \text{(rule } r \text{ has been already extracted)} \end{cases} \quad (4)$$

β: coefficient for the number of attributes.

$n(r)$: the number of attributes in the antecedent of rule r.

cosine, $n(r)$ and $\alpha_{new}(r)$ are concerned with the importance, complexity and novelty of rule r, respectively.

The fitness represents the potential to extract new rules.

B. Termination Mechanism: Gene Matrix and Mutagenesis

In AT-GNP, the termination instant is determined based on the GM, that works mutually with a special mutation operator called "mutagenesis". Mutagenesis is a more artificial mutation operation that allows some characteristic children to improve themselves by modifying their genes in accordance with the status of the GM.

During the search, the information related to the connections between each node is stored within a matrix M. M is initialized as a square diagonal zero matrix of order $p+q$, where p is the number of judgment nodes and q is the number of processing nodes. While the nodes are being connected during the search process, the corresponding entries in M are updated with a non-null value. This information is used in two ways. First, areas of the search space being

 ISBN: 978-1-61208-227-1

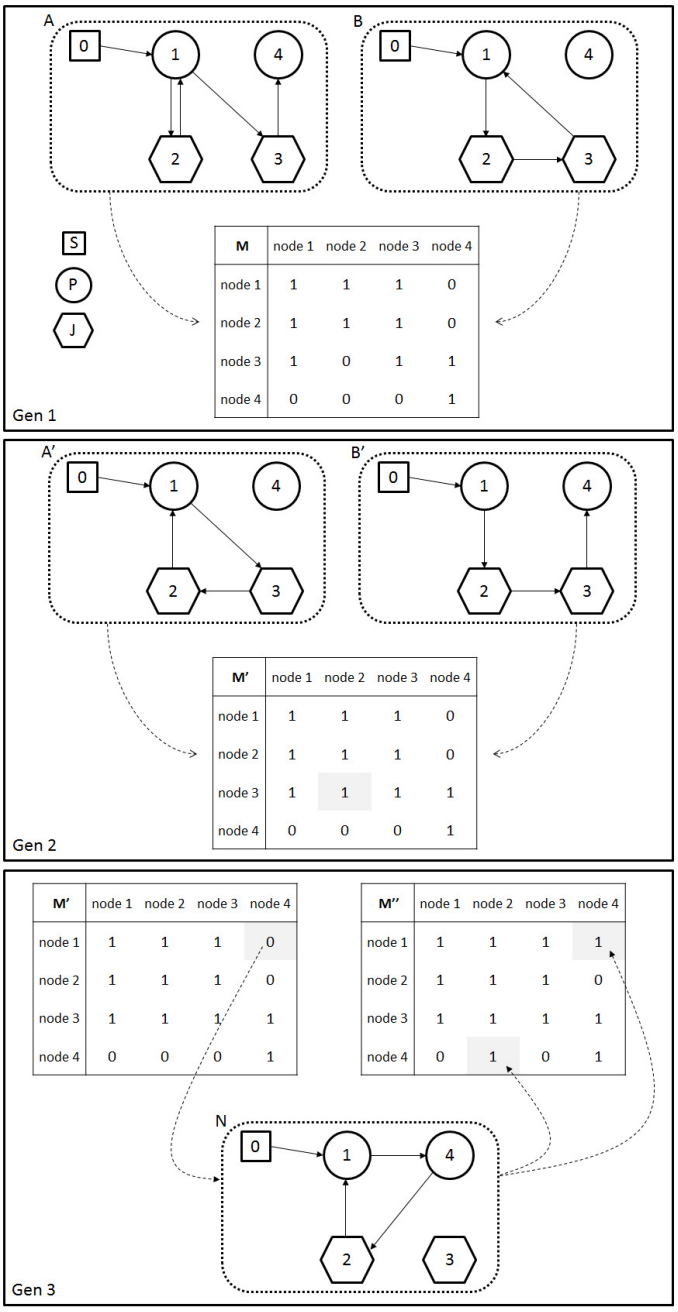


Figure 3. An example of the termination mechanism.

unexplored are revealed during the search. Thus, at each generation, some individuals are mutated in such a way that unexplored regions are visited: The so-called mutagenesis. Second, when M does not contain zero entries anymore, the search is considered to have achieved an advanced exploration process and is stopped.

Mutagenesis operates in two ways in combination with the GM. First, some of the worst individuals that have been selected by the survivor operator to figure in the next generation are altered. To avoid premature convergence, the number of selected individuals should be less than the number of individuals altered by the normal mutation operator. Preliminary tests show that altering the worst or the two worst individuals lead to good performance in most cases. By doing so, we keep genetic diversity and accelerate the exploration process. The difference with the normal mutation operator is that it is not completely random. Indeed, mutagenesis is guided by the status of the GM. Specifically, a zero-position in GM is randomly chosen, say the position (i, j). Then the considered individual sees one of its connections altered with a link between nodes i and j. Hence, there is a chance for the crossover operation to explore different combinations of solutions containing this setting. Afterward, the GM is updated since a new connection has been created.

Figure 3 shows an example of this mechanism with three generations, namely, *Gen 1*, *Gen 2* and *Gen 3*. For the sake of simplicity, the population is reduced to two individuals. At generation *Gen 1*, individuals A and B are represented, along with the matrix M. An individual can be formed using two processing nodes and two judgment nodes. Hence, M is a square matrix of order 4. One can see that many entries are still equal to zero: Entry $(3, 2)$ for instance, meaning that node 3 is not followed by node 2 in any individual, although node 2 leads to node 3 in individual B. In generation *Gen 2*, A and B has been evolved to become individuals A' and B', respectively. Although they represent completely different solutions, with hopefully higher fitness, the matrix M' associated with the second generation reveals that the contribution of A' and B' in increasing the diversity of the population is very poor. Indeed, from the first to the second generation, a comparison between M and M' shows that only entry $(3, 2)$ turned to a non-null value. Consequently, we use this information to accelerate the search by specifically generating new solutions so that their composition contains connections between nodes that have not been explored. Let us consider entry $(1, 4)$ for instance. In our example of Figure 3, an original mutation operator modifies during the third generation (normally, at each generation) a candidate solution such that the resulting solution, here referred to as N, will explicitly contain nodes 1 and 4 connected to each other. In this way, M evolves over the generations such that all its entries become non-null. When that point is reached, then the search is terminated.

IV. Methodology

Fig. 4 shows the schema of the proposed method to evaluate the classification accuracy. 10-fold cross validation procedure was performed on the dataset and the results will be given by their average.

The training set and test set are generated randomly from the dataset. Using the training set, the proposed AT-GNP mining method is applied to obtain a pool of class association rules for each class. Two classes are shown in Fig. 4 as an example, that is, Class A and Class B. Finally,

 ISBN: 978-1-61208-227-1

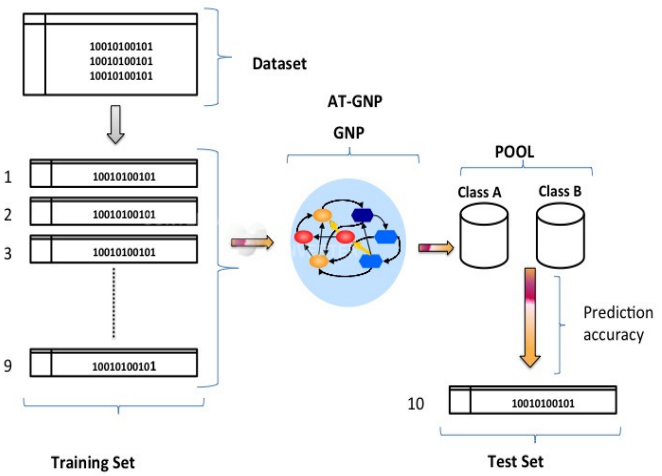


Figure 4. Schema for evaluating the classification accuracy.

these pools are used to evaluate the prediction accuracy of the test set. All algorithms were coded in Java language. Experiments were performed on a 3.2 GHz Pentium PC with 12GB main memory, running Microsoft Windows 7 Ultimate.

A. Numerical Experiments

To assess the performance of the automatically terminated GNP (AT-GNP) against the conventional GNP, the results from 4 widely used databases in the field of DM [16] are compared. Both AT-GNP and GNP were confronted to a classification problem and 10 independent runs were performed in the same experimental conditions for each database. Table I reports the obtained averaged results in terms of number of generations, number of extracted rules and running time for each considered class as well as in terms of accuracy level, for databases "labor", "crx", "hepatitis" and "vehicle", respectively.

Table II shows the comparison of the classification accuracy using several conventional methods. The results of C4.5. [17], Ripper [18], CBA [19], CMAR [20] and CPAR are taken from the Yin and Han paper [21].

For the four considered databases, it can be seen that for an equivalent or slightly improved accuracy, the number of generations obtained by AT-GNP right before automatic termination in all classes is on average 55% less expensive than what is required by GNP. The computation time is directly proportional to the number of generations and sees a reduction of 60% on average. The amount of extracted rules by AT-GNP is in almost all cases lower or of same order than of GNP. However, as indicated by the accuracy level, this does not have a negative impact on the quality of the final solution. For the last database, it is very interesting to notice that where the number of generations obtained by automatic termination is higher than of GNP, the number of extracted rules was particularly low. This may be explained by the fact that AT-GNP is automatically adjusting the effort in an attempt to extract more rules.

During our experiments, we have also considered allowing artificially much more and much less number of generations to assess whether or not AT-GNP suffered from premature convergence or did not terminate without unnecessary computation. However, as partially indicated by the comparison with GNP and a doubled number of generations, it is clear that AT-GNP did dot suffer from any of them.

V. Conclusion and Future Work

By equipping GNP with an automatic termination, we could alleviate the need to specify a given number of generations before running the algorithm. The numerical experiments demonstrated that our mechanisms could determine a proper termination instant without prior knowledge of the database to be handled. For an equivalent or slightly superior accuracy level, AT-GNP requires on average half the number of generations needed by GNP and is thus also reducing the computing time by half.

For future work, the method will be extended to deal with large and heterogeneous scientific databases combined with web data. Also, the authors will study the circumstances under which our termination technique is accurate. It involves the distribution of the data statistics of the databases.

References

[1] E. Gonzales, T. Nakanishi and K. Zettsu, "Large-Scale Association Rule Discovery from Heterogeneous Databases with Missing Values using Genetic Network Programming", in *Proc. of the First International Conference on Advances in Information Mining and Management (IMMM2011)*, pp. 113-120, Barcelona, Spain, October 2011.

[2] E. Gonzales and K. Zettsu, "Association Rule Mining from Large and Heterogeneous Databases with Uncertain Data using Genetic Network Programming", in *Proc. of the Fourth International Conference on Advances in Databases, Knowledge and Data Applications (DBKDA2012)*, pp. 74-80, Saint Gilles, Reunion Island, March 2012.

[3] E. Gonzales, K. Taboada, K. Shimada, S. Mabu, and K. Hirasawa, "Combination of Two Evolutionary Methods for Mining Association Rules in Large and Dense Databases", *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 13, No. 5, pp. 561-572, 2009.

[4] K. Shimada, R. Wang, K. Hirasawa and T. Furuzuki, "Medical Association Rule Mining Using Genetic Network Programming", *IEEJ Trans. EIS*, Vol. 126, No. 7, pp. 849-856, 2006.

[5] K. Shimada, K. Hirasawa, and T. Furuzuki, "Genetic Network Programming with Acquisition Mechanisms of Association Rules", *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 10, No. 1, pp. 102-111, 2006.

[6] M. S. Giggs, H. R. Maier, G. C. Dandy, and J. B. Nixon, *Minimum number of generations required for convergence of genetic algorithms*, in "Proceedings of 2006 IEEE Congress on Evolutionary Computation", Vancouver, BC, Canada, pp. 2580-2587, 2006.

 ISBN: 978-1-61208-227-1

Table I
COMPARISON RESULTS BETWEEN GNP AND AT-GNP FOR SEVERAL DATASETS.

labor	Class "Good"			Class "Bad"		
	Generations	Num. of Rules	Time (s)	Generations	Num. of Rules	Time (s)
GNP	101 (fixed)	12363	103.4	101 (fixed)	2991.2	21
AT-GNP	43 (auto)	6347.6	32.4	43.4 (auto)	2237	10.2
CRX	Class "+"			Class "–"		
	Generations	Num. of Rules	Time (s)	Generations	Num. of Rules	Time (s)
GNP	101 (fixed)	562	13	101 (fixed)	1317.6	17.2
AT-GNP	48.4 (auto)	444.6	7	47.6 (auto)	854	8.6
hepatitis	Class "live"			Class "die"		
	Generations	Num. of Rules	Time (s)	Generations	Num. of Rules	Time (s)
GNP	101 (fixed)	34983.4	428	101 (fixed)	2819.6	16.6
AT-GNP	42.6 (auto)	15727.4	97.2	45.2 (auto)	1847.8	8.2

vehicle	Class "bus"			Class "saab"			Class "opel"			Class "van"		
	Gen.	Num. Rules	T(s)	Gen.	Num. Rules	T(s)	Gen.	Num. Rules	T(s)	Gen.	Num. Rules	T(s)
GNP	101	596	11.2	101	4.6	7.8	101	1	7.8	101	174	9.6
AT-GNP	68.4	486.4	8	173.8	5	13	186.2	1	15.4	65.6	159.4	6.6

Table II
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS

Dataset	AT-GNP	GNP	C4.5	Ripper	CBA	CMAR	CPAR
Labor	90.66%	89.66%	79.30%	84.00%	86.30%	89.70%	84.70%
CRX	84.01%	83.44%	84.90%	84.90%	84.70%	84.90%	85.70%
Hepatitis	86.15%	86.11%	80.60%	76.70%	81.80%	80.50%	79.40%
Vehicle	76.12%	76.12%	72.60%	62.70%	68.70%	68.80%	69.50%
AVERAGE	84.23%	83.83%	79.35%	77.07%	80.37%	80.97%	79.82%

[7] B. J. Jain, H. Pohlheim, and J. Wegener, *On termination criteria of evolutionary algorithms*, in "Proceedings of the Genetic and Evolutionary Computation Conference", Morgan Kaufmann Publishers, pp. 768, 2001.

[8] N. M. Kwok, Q. P. Ha, D. K. Liu, G. Fang, and K. C. Tan, *Efficient particle swarm optimization: a termination condition based on the decision-making approach*, in "Proceedings of the IEEE Congress on Evolutionary Computation", Singapore, pp. 25-28, 2007.

[9] B. T. Ong and M. Fukushima, *Genetic algorithm with automatic termination and search space rotation*, Memetic Computing, **3**, pp. 111-127, 2011.

[10] B. T. Ong and M. Fukushima, *Global optimization via differential evolution with automatic termination*, Numerical Algebra, Control and Optimization, **2**, pp. 57-67, 2012.

[11] C. Zhang and S. Zhang, *Association Rule Mining: models and algorithms*, Springer, 2002.

[12] S. Mabu, K. Hirasawa, and J. Hu, "A Graph-Based Evolutionary Algorithm: Genetic Network Programming (GNP) and Its Extension Using Reinforcement Learning", Evolutionary Computation, *MIT Press* , Vol 15, No. 3, pp. 369-398, 2007.

[13] K. Hirasawa, T. Eguchi, J. Zhou, L. Yu, J. Hu, and S. Markon, "A Double-deck Elevator Group Supervisory Control System using Genetic Network Programming",*IEEE Trans. on System, Man and Cybernetics, Part C*, Vol. 38, No. 4, pp. 535-550, 2008.

[14] T. Eguchi, K. Hirasawa, J. Hu, and N. Ota, "A study of Evolutionary Multiagent Models Based on Symbiosis",*IEEE Trans. on System, Man and Cybernetics, Part B*, Vol. 36, No. 1, pp. 179-193, 2006.

[15] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", in *Proc. of the 20th VLDB Conf.*, pp. 487-499, 1994.

[16] Frank, A. Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [Last Access: May 14th, 2012]

[17] J. R. Quinlan, "4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[18] W. Cohen, "Fast effective rule induction", In Proc. of the ICML'95, pp. 115-123, Tahoe City, CA, July 1995.

[19] B. Liu, W. Hsu and Y. Ma, "Integrating Classification and Association Rule Mining", In Proc. of the KDD'98, pp. 80-86, New York, NY, Aug. 1998.

[20] W. Li, J. Han and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules", In Proc. of the ICDM'01, pp. 369-376, San Jose, CA, Nov. 2001.

[21] X. Yin and J. Han, "CPAR:Classification based on Predictive Association Rules", In Proc. of the 2003 SIAM International Conference on Data Mining (SDM'03), 2003.

 ISBN: 978-1-61208-227-1

Information Mining Over Significant Interval on Historical Data

A Study on World Major Indexes

Sim Kwan Hua
School of Engineering, Computing & Science
Swinburne University of Technology
Kuching, Malaysia
e-mail: khsim@swinburne.edu.my

Abstract—**Despite the popularity of financial charting software catalyzed by the advancement in computing technology over the past decade, the analysis of financial historical data through charting software remains at the surface of statistical description. Analysis of historical data presented typically on a price chart should be elevated further to information mining that could interpret the fundamental condition on the ground, as an important effort to support a good decision making process. This paper introduces a new way of interpreting historical financial data by calculating the mean value of the historical data over a significant interval, and eventually mining the high intensity price level. Experiment was conducted on historical data of six major world indexes over the period of ten years to assess the competency of the use of mean value over significant intervals in comparison to static interval used in conventional moving averages. The outcome of the experiment reveals the relevancy of the use of mean value over significant intervals on all the six major world indexes. This study institutes and demonstrates a new way of mining fundamental information and insight from a historical data set; the finding stimulates an innovative way on how data can be interpreted to derive information that is crucial in financial decision making process.**

Keywords-Data Analysis; Time Series Analysis; Data Mining; Statistical Analysis; Technical Analysis

I. INTRODUCTION

Over the past one decade, financial charting software has been a primary tool used in analyzing exchange markets such as stocks, futures, commodity and foreign currency exchanges. It is reported that more than 75% of professionals in the exchange markets of various instruments rely on financial charts at some points in making their trading decisions [1]. Therefore, mining the information required from the price chart forms part of the critical components in the decision making process of financial market participants.

The advancement of computing and software technology has made financial charting software widely accessible to institution and individual retailers; some of financial software providers include MetaTrader from MetaQuates, Routers MetaStock, TD Ameritrade, Thinkorswin from Bloomberg and Lauchpad & Charts from TradeStation. It is a very common practice nowadays for brokerage firms and investment banks to integrate real-time financial charting components into their trading software for use by their dealers and clients.

Furthermore, massive volatility in the more recent year in most of the financial instruments worldwide has quite a tremendous impact not only on economics but also on both social and political aspects. It provokes a major challenge confronting everyone ranging from investors, speculators, businesses, and even to the level of governmental policy makers. As a matter of fact, financial markets are affected by many highly interrelated variables such as economics, political and even psychological factors in a very complex manner, making financial time series one of the most difficult analyses among all other time series analysis [2].

Market participants worldwide have always shown keen interest trying to predict the future of the overall market movement in order to maximize their returns, at the same time hedging and mitigating their risks again potential pitfalls. Real-time charts provided by charting software showing the latest price change has become the primary tools in helping them in their day to day decision making process. Nevertheless, price chart is only a graphical representation of historical data from the past. Thus, information mining from the price chart becomes extremely vital to conclude a reliable decision which is in alignment with the future price movement.

Numerous studies have been attempted trying to analyze the financial markets by using various time series analysis techniques, alongside with few other popular time series models and stochastic models used in signal processing. Every single data point on a price chart is a successful transaction recorded at certain point in time, and it can be regarded as a signal reflecting the equilibrium states of all the correlated factors, may it be fundamental, technical or even emotional factor [3].

Nevertheless, most of the approaches introduced so far utilize historical data by performing learning or training on them in order to identify certain patterns, or to compute probabilities, or even optimizations which involve mostly the use of Artificial Intelligent techniques. All these are based on the assumption that the history will repeat itself. But, the

 ISBN: 978-1-61208-227-1

existence of randomness element in financial market claimed by Efficient Market Hypothesis (EMH) is undoubtedly hampering the performance of those approaches.

To date, very little effort has been made to mine information over a selected interval on the price chart. In most cases, the whole set of historical data will be used or fed into a model without any effort of information mining to pre-process the data. Even in technical analysis, which is starting to gain some grounds, majority of the technical indicators remain as a statistical summary of a set of data points in the past.

This paper aims to mine the underlying transaction information from historical data set of a price chart by identifying significant interval. In other words, the amount of data in the past is not the sole determinant in order to mine the required information. Hence, it does not operate on the assumption that history will repeat itself, and essentially eliminating the concern of random element in financial market.

The experiment and testing are done on six major world indexes on data over the past twelve years, starting from year 2000, and the results will be analyzed and discussed.

This paper begins with Section I as an introduction; Section II concerns the background; Section III elaborates on the mean over significant interval approach; Section IV describes the experiment conducted; at the same time discusses analytical results. Section VI presents the conclusion.

II. BACKGROUND

Generally, there are three schools of thought in financial market analysis; they are generally known as Efficient Market Hypothesis (EMI), fundamental analysis and technical analysis.

Efficient Market Hypothesis believes that no one can achieve above average advantages based on any historical and present data. This is also backed by another prominent, Random Walk Hypothesis, which states that prices of financial instruments wander in a purely random way. Likewise, Efficient Market Hypothesis advocates that all available information is fully reflected on the price itself [4]. Both theories dictate that the previous change in the value of a variable, such as price, is unrelated to future or past changes. As a result, statistical data collection implicitly defines each data point as independent. Based on such contextual assumptions, the data can appear random when the data points are treated as discrete events. However, S. Taylor (1986), Russell and Torbey (2008) provide compelling evidence to reject both of these theories [5].

Fundamental analysis from the second school of thought studies the underlying intrinsic value of a financial instrument by analyzing fundamental factors such as economics, financial, accounting and business environments, and their effects on its future value. Though fundamental analysis possesses the longest and profound history in the world of financial analysis, it is not meant for studying the volatility and fluctuation of prices around the underlying intrinsic value of a financial instrument. The huge volatility in financial and commodity markets lately has denoted the inadequacy of fundamental analysis to attest those huge fluctuations.

The third school of thought is technical analysis with practitioners who analyze primarily upon charts that based solely on market-delivered data such as price and volume. They perform statistical study rather than examine the economics fundamentally driven information in analyzing a financial instrument. Consequently, technical analysis does not receive sufficient level of scrutiny from academic researchers. As such, it served more as a secondary tool in financial market analysis [6].

Recently, researchers have introduced many different approaches in various fields of study as an effort to derive more useful information from historical data, which may include both fundamental data and technical data in order to analyze the financial markets.

Together with some other Artificial Intelligence (AI) techniques, Artificial Neural Network (ANN) has been one of the popular models for predicting financial markets. Cao Qing, et al. [7] and R. M. Rahman, et al. [8] in their respective studies, had used the neural networks to predict the future movements of various financial instruments. The results showed that the performance of ANN technique in forecasting financial instruments was very convincing and outperformed conventional linear models.

Manish Kumar and Thenmozhi M [9] endeavored to the use of Support Vector Machines (SVM) on S&P CNX NIFY; their result showed that SVM outperforms neural networks, discriminate analysis and logic model used in the study. Besides, stochastic model is another prevalent preference; the work of B. Kaushik [10] specified a two-state Markov Chain Model for discredited returns and proposed a measure for efficiency by using the modulus of the second highest Eigen value of the transition matrix and relates it to the speed of convergence of the Markov chain. In a more recent study, S. Vasanthi, et al. [11] explored the capability of Markov Chain Analysis in predicting indexes of emerging markets, and the result showed that Markov model outperforms the conventional moving averages in technical analysis.

Apparently, models were built from various areas of studies over the years by processing a wide range of both fundamental and technical financial data in mining useful information to aid in the decision making process of financial market participants [3]. So far, little attention has been paid to pre-process the data before feeding the data into the model for analysis.

In technical analysis, models have been developed mainly based on historical price data, statistical calculation such as mean, standard deviation and the rate of change will be performed to compute the statistical description required to form an understanding of the latest state of a given data set. Coherently, different data set will derive different statistical description; even selecting two different intervals from the same data set will produce two very different statistical descriptions.

 ISBN: 978-1-61208-227-1

Typically, statistical analysis will be represented graphically through charting software in the form of technical indicators such as Moving Average, Relative Strength Index, Stochastic and Moving Average Convergence Divergence. Analysis is normally done by selecting a technical indicator to be applied on a set of historical data plotted on a price chart.

Basically, a price chart is composed from price feed supplied by the exchange market, where vertical axis represents the price level and horizontal axis states the time when the transaction is made. In other words, price chart is just a graphical representation of historical data plotted on a chart. Applying a statistical analysis such as an *n*-period Moving Average on a chart simply insinuates the deriving of mean value from the last *n* data points [12]. The average or mean value of a set of price data points derived in the past has very little provision on the future value, and to the direction of the new price value in the future.

This is also supported by Efficient Market Hypothesis and Random Walk Theory which believe that price is a discrete event, thus any information derived from past data, including the mean value provides no influential trace on the possible value of price in the future.

Instead of pondering around the typical vindication of statistical description, this study aims at mining market information from the mean values by focusing on the selection of appropriate intervals used in calculating mean value. Fundamental information can be mined from mean value if the interpretation is based on significant intervals from the highest or the lowest price level recorded on a price chart. Since moving average is the main way of obtaining mean value in financial charting, the investigation will therefore, propagate from Moving Averages in technical analysis.

III. INTERVAL OF MOVING AVERAGES

In statistics, moving average is a type of finite impulse response filter used to analyze a set of data points by creating a series of averages of different subsets of the full data set [12].

Moving average smoothes a data series over the fluctuation for a specified time period to delineate and spot the overall trend of the past data. In financial series analysis, moving average is defined as the average (mean) price of an instrument over a specified time period.

Given a sequence $\{a_i\}_{i=1}^{N}$, an *n*-moving average is a new sequence $\{s_i\}_{i=1}^{N-n+1}$ defined from the a_i by taking the average of subsequences of *n* term.

$$s_i = \frac{1}{n} \sum_{j=i}^{i+n-1} a_j \tag{1}$$

In simple terms, it is calculated by adding the instrument prices for the most recent *n* data points and then dividing by *n* [12].

Over the years, research efforts have been directed to improve analytical power of moving averages through the introduction of various types of moving averages, such as exponential moving average, weighted moving average, accumulative moving average, triangular weighting, double smoothing, triple exponential moving average (TEMA), geometric moving average and many more [13]. Nonetheless, three most commonly used moving averages are simple moving average, exponential moving average and weighted moving average [14].

Exponential moving average (EMA) is a type of infinite impulse response filter that applies weighting factors which decreases exponentially. The EMA for a series *Y* is calculated recursively with:

$$S_t = \alpha \times Y_t + (1 - \alpha) \times S_{t-1} \tag{2}$$

Where the coefficient represents the degree of weighting decrease, it may be expressed in terms of *N* interval, where $\alpha = 2/(N+1)$. Y_t is the observation at a time period *t* and S_t is the value of the EMA at any time period *t* [13].

A weighted moving average is any average that has multiplying factors to give different weights to data at different positions in the sample window; it is expressed in the following general form:

$$W_t = \frac{w_1 P_t + w_2 P_{t-1} + \cdots + w_n P_{t-n+1}}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i P_{t-i+1}}{\sum_{i=1}^{n} w_i} \tag{3}$$

This gives the weighted moving average at time *t* as the average of the previous *n* prices (P), each with its own weighting factor w_t [13].

The introduction of a variety of moving average serves as a strong evidence of attempts to improve the meagerness of moving average via mathematical approaches. However, besides optimization techniques to find the most optimal interval, literally no other attempt has been made on the selection of intervals for calculating moving average.

As price chart records data points of prices transacted over a specific period of time, calculating the mean of the data points signifies the average transaction price over that period of time, which brings forth a very important information; the average overall price level of all the participants who involved in the transaction during that interval.

Hence, the selected mean interval is the key in mining useful mean or average transaction price. If the interval used to calculate mean from the highest or lowest price onward, then the portfolio condition of one who has traded during that interval can be inferred easily. In other words, a useful average transaction price level can be mined by selecting a significant interval which starts from a highest or lowest price level.

As the overall average position over a significant interval can be known, decision to be made by majority of the market

 ISBN: 978-1-61208-227-1

participants can then be anticipated. According to decomposition effect of Prospect Theory, intensity of reactions and behaviors of market participants will mount when the new price level approaches their overall average position, which unquestionably will affect the position of their portfolios [15].

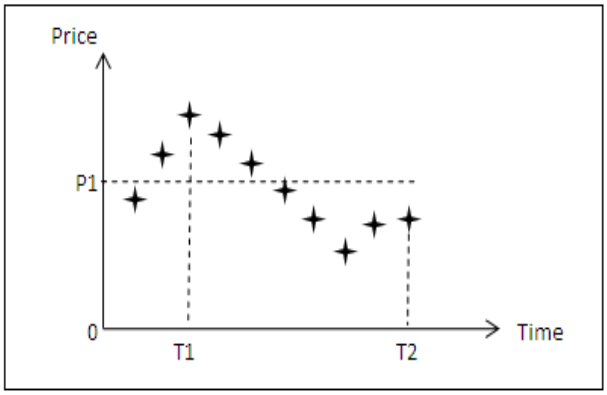


Figure 1. Mean over significant interval.

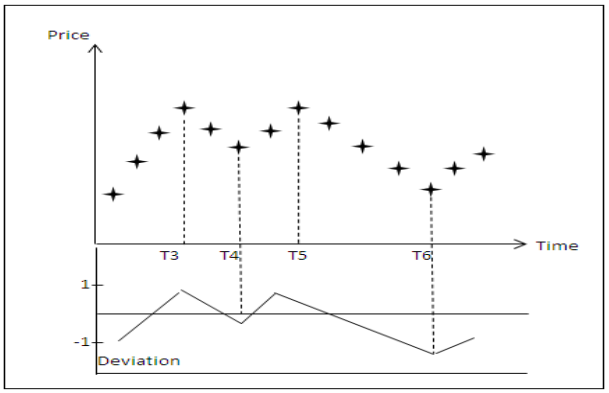


Figure 2. Defining highest and lowest point.

As illustrated in Figure 1, since the mean can be interpreted as the average transaction level over a specific interval, it is critical to mine the mean for the interval from T1 onward, interval from T1 onward can be categorized as significant interval because the highest price is recorded at T1. For that reason, all the participants who entered after T1 will have a negative impact on their portfolios. The mean transaction level for this group of participants is collective at price P1 when time reaches T2. In other words, if the price at T2 were to reach P1, the portfolios of this group of participants will turn from negative to positive, allowing them to reach their breakeven point, and high intensity of reaction can be anticipated in the decision making process of this particular group of people.

However, prices of financial instruments especially those with high liquidity will fluctuate at various degrees of magnitude, causing the highest and lowest price to be very subjective when it comes to identifying significant interval as time moves on. Thus, it is necessary and imperative to exercise objectivity in identifying the highest and lowest price.

Therefore, a significant interval used in this study is defined as the duration of movement of price exceeding one standard deviation away from the 50-period mean, which denotes a high intensity of consensus among market participants has been reached on the forming of a trend [16]. As illustrated in Figure 2, duration from T3 onward will not be considered as a significant interval because price at T4 is less than 1 standard deviation away below 50-period mean. On the other hand, interval from T5 onward is considered as a significant interval after T6, where price movement has exceeded one standard deviation away at T6. So, the mean value over a significant interval should be calculated after T6, and this mean value carries vital information on the overall position of market participants. It is extremely useful in making a favorable decision when price approaches that mean value at any point of time after T6.

IV. EXPERIEMENT AND EVALUATION

The experiment was conducted to observe the reaction of price when it approaches the mean value over a significant interval. This experiment was done on six major world indexes over a period of twelve years starting from 2000 to the end of 2011. These major world indexes include Dow Jones Industrial Average (DJI), Deutscher Aktien-Index (DAX), CAC40, FTSE100, Hang Seng Index (HSI) and Nikkei225.

The experiment data used is the daily data of those six major world indexes, downloaded from Yahoo Finance and the experiment was carried up by using Meta Stock's system tester.

The evaluation was performed by comparing the reaction of price between various conventional moving averages with typical static intervals and mean value derived over significant intervals dynamically. Conventional static intervals used for comparison in this experiment range from 10-period to 200-period moving average.

In order to differentiate between the normal price fluctuation and decent price reaction upon approaching high intensity price level, the average daily range of all the indexes needs to be computed. This is to identify and estimate the normal daily range of price fluctuation for all the six major indexes, so that the abnormal price fluctuation can be classified.

Since the historical data used in this experiment is the daily end-of-day data, the average daily range of all six major indexes over twelve years testing period was calculated and summarized in the following table:

 ISBN: 978-1-61208-227-1

TABLE I. AVERAGE DAILY RANGE

Index	CAC	FTSE	HSI	NIKKEI	DJI	DAX
Average Daily Range (%)	1.78	1.63	1.53	1.52	2.35	1.99

It is shown in Table I that the averages of daily range for all the indexes studied are logged within range of 1.52% to 2.35%. This implies that on average, most indexes have a normal daily fluctuation of within approximately two percent of the index itself.

Consequently, in this experiment, a valid price reaction toward high intensity price level is defined as a price reversal into opposite direction for more than five percent from the mean value. It is set to be at least twice of the normal daily fluctuation range, so that the validity of the price reaction can be attested, hence minimizing the possibility that the price reaction is caused by normal a daily fluctuation.

Thus, once the new price data reaches the mean value, a reaction will be captured and recorded if the price reacted or bounced into opposite direction for more than five percent.

However, it will be recorded as no reaction if the new price level continues to move in its preceding direction for more than two percent after hitting a mean value. Likewise, the two percent buffer is to allow the normal daily fluctuation of the indexes.

Simulations were run on twelve years historical data of the six major world indexes individually, and evaluations on the price reaction toward mean value of significant interval along with conventional static moving averages were recorded. Correspondingly, the observation and evaluation of price reaction during the simulation were done for both upward and downward movements.

The outcomes obtained from the experiments are presented in the following tables:

TABLE II. UPWARD MOVEMENT

Interval	CAC	FTSE	HSI	NIKKEI	DJI	DAX
MA10	30.06%	31.68%	28.85%	30.49%	34.65%	31.01%
MA25	33.67%	30.88%	34.44%	25.49%	36.71%	29.69%
MA50	28.95%	28.00%	28.57%	28.57%	32.35%	27.54%
MA100	30.43%	25.81%	36.11%	27.91%	24.00%	26.83%
MA150	30.00%	34.78%	44.00%	25.93%	20.51%	38.71%
MA200	20.83%	30.00%	35.00%	22.22%	12.50%	42.86%
Significant Interval	58.06%	56.00%	60.87%	61.25%	55.56%	55.13%

TABLE III. DOWNWARD MOVEMENT

Interval	CAC	FTSE	HSI	NIKKEI	DJI	DAX
MA10	27.47%	30.28%	30.72%	33.54%	31.72%	27.34%
MA25	27.36%	25.00%	24.00%	39.39%	27.27%	28.81%
MA50	30.67%	33.33%	23.53%	30.00%	26.67%	27.63%
MA100	30.95%	34.48%	25.64%	37.14%	36.36%	27.66%
MA150	32.14%	33.33%	26.09%	45.83%	40.00%	21.21%
MA200	42.86%	42.11%	27.27%	30.43%	42.42%	11.11%
Significant Interval	61.29%	55.93%	65.91%	59.26%	68.29%	59.38%

The result in Table II exhibits that in an upward movement, there is a probability of approximately 30% for price to bounce into opposite direction in response to the moving averages with static intervals ranging from 10 periods to 200 periods. The best probability recorded among static interval moving averages is 44% from a 150-period moving average on HSI, while the lowest probability deeps as low as 12.50% from a 200-period moving average on DJI.

Notably, the mean from significant intervals yields a substantially higher probability of well above 50%; indeed it ranges from 55.13% to 60.87% across all the six major indexes evaluated in this experiment.

Similarly, Table III presents the outcome for downward price movement, and the result obtained is very consistent with the result attained from the upward price movement in Table II. All the conventional moving averages with static intervals have also recorded a probability of around 30%. Again, mean over significant intervals has achieved a probability of well above 50%. Moreover, the highest probability for mean over significant intervals in a downward movement reached as high as 68.29% on DJI.

A promising reaction of price toward mean value over significant intervals has been observed on all the six major world indexes over the twelve years testing period, which obviously cover different phrases of market condition. Conversely, such price reaction fails to be observed on any of the conventional moving averages with static intervals. Clearly, the mean value over significant intervals has the capability to mine a high intensity price level that causes noticeable price reaction.

In summary, the results reveal that the mean value over significant intervals has absolute higher probability of getting the price to bounce into opposite direction compared to all the conventional static intervals of moving averages. The obvious differences signify that information mining from calculating mean of significant intervals deserve a serious attention.

V. CONCLUSION

This study explored a new dimension of using technical analysis to mine information that represents the underlying condition, and not just based purely on statistical description. It is believed that the mean calculated from a

 ISBN: 978-1-61208-227-1

significant high or low can be interpreted as a high intensity average price level of majority of the market participants. The experiments and preliminary results suggest that mean over significant intervals demonstrate a very promising accomplishment compared to the other conventional moving averages with static intervals. This has also initiated a new epoch of how technical analysis can be interpreted to mine information from a price chart, and to elevate the performance and superiority of charting software in financial market analysis. Last but not least, future research should explore the possibility of extending this concept of mining fundamental information right from price chart to other technical indicators, with an expectation for more precise information mining in financial decision making.

REFERENCES

[1] T. Gehrig and L. Menkhoff, "Extended Evidence on The Use of Technical Analysis in Foreign Exchange", International Journal of Finance & Economics, Vol 11, Issue 4, 2006, pp. 327-338.

[2] G. Boetticher, "Teaching Financial Data Mining using Stocks and Futures Contracts", *Journal of Systemic, Cybernetics and Informatics*, Vol 3, no 3, 2006, pp. 26-32.

[3] J. V. Hansen, J.B. McDonald, and R. D. Nelson. "Some Evidence of Forecasting Time-series with Support Vector machines", Journal of Operational Research society, vol. 57, no.9, 2006, pp. 0153-1063.

[4] M. G. Kendall and H. Bradford, "The Analysis of Economic Time-Series- Part1: Prices". Journal of the Royal Statistical Society, 116(1), 1953, pp. 11-34.

[5] W. M. Martin, "Technical Anaysis: The interface of Retional and Irrational Decision Making" Business Review, Cambrige, 11. 2, 2008, pp. 48-54.

[6] K. P. Hanley, "Scientific Frontiers and Technical Analysis", Journal of Technical Analysis, vol. 64, 2006, pp. 20-33.

[7] Q. Cao, K. B. Leggio, and Schniederjans, "A Comparison between Fama and French's model and Artificial Neural Networks in Predicting the Chinese Stock Market", Computer and Operations Research 32, 2005, pp. 2499-2512.

[8] R. M. Rahman, R. K. Thulsiram, and P. T. Thulasiraman, "Performance Analysis of Sequential and Parallel Neural Network Algorithm for Stock Price Forecasting", International Journal of Grid and High Performance Computing, 3(1), 2011, pp. 45-68.

[9] K. Manish and M. Thenmozhi, "Predictability and Trading Efficiency of S&P CNX Nifty Index Returns Using Support Vector Machines and Random Forest Regression", Journal of Academy of Business and Economics, Vol.7(1), 2007, pp. 15-23.

[10] B. Kaushik, "A Measure of Relative Effieciency of Financial Markets from Eigen value based Mobility Indices", Finance India, Vol.XVI, No.4, 2002, pp. 1419-1425.

[11] S. Vasanthi, M. V. Subha, and Thirupparkadal Nambi, "An Empirical Study on Stock Index Trend Prediction using Markov Chain Analysis", Journal of Banking Financial Services and Insurance Research, Vol 1, 2011, pp. 72-91.

[12] M. F. Triola, "Essientials of Statistics, 4th Edition", Addison-Wesley, Longman Inc, 2011, pp. 112-126.

[13] P. J. Kaufman, "Trading Systems and Methods", John Wiley & sons, 1998, pp. 66-88.

[14] L. Stevens, "Essential Technical Analysis: Tools and Technique to Spot Market Trend", John Wiley and Sons, 2002, pp. 218-238.

[15] M. Massa and N. W. Geotzmann, "Disposition Matters: Volume, Volatility and Price Impact of a Behavioral Bias", Yale ICF Working Paper, No. 03-01, 2003.

[16] K. H. Sim, "Recognizing the Formatioin of Trend: A standard Deviation Approach", Proceeding of 4th International Conference of Interation Sciences: IT, Human and Digital Content", IEEE press, 2011, pp. 136-142.

 ISBN: 978-1-61208-227-1

A Fast Short Read Alignment Algorithm Using Histogram-based Features

Qiu Chen, Koji Kotani*, Feifei Lee, and Tadahiro Ohmi

New Industry Creation Hatchery Center, Tohoku University
** Department of Electronics, Graduate School of Engineering, Tohoku University*
Aza-Aoba 6-6-10, Aramaki, Aoba-ku, Sendai 980-8579, JAPAN
e-mail: qiu@fff.niche.tohoku.ac.jp

***Abstract*—Current new generation of DNA sequencers have had the ability to generate billions of short reads rapidly and inexpensively. How to solve fast and robust short read alignment problem become one of the most important challenges in bioinformatics research area. The current solutions for short-read alignment have limitations that alignment algorithms such as MAQ and Bowtie have few capabilities to align reads with insertions or deletions. In this paper, we propose an efficient hierarchical alignment algorithm to reduce it. For a given short read, first, a fast histogram search method is used to scan the reference sequence. Most of locations in reference sequence with low similarity will be excluded for latter searching. The Smith-Waterman alignment algorithm is then applied to each remainder location to search for exact matching. Experimental results show the proposed method combining histogram information and Smith-Waterman algorithm is a faster and accurate algorithm for short read alignment.**

Keywords-Short read; Alignment; Fast search; Smith-Waterman; Histogram-based feature

I. INTRODUCTION

The decipherment of 3-billion-base human genome sequence was finally completed by the international cooperation in April 2003 [1][2]. Since this achievement of human genome project, researchers around the world are now having a very keen competition on clarification of the structure and performance analysis of the protein, genes and protein networks, and new gene sequences are clarified every day. The enormous quantity of data has been accumulated in the database like GenBank [3], EMBL [4], and DDBJ [5], etc. Moreover, the volume of data of Genome Database still increases in exponential [6].

Current new generation of DNA (Deoxyribonucleic Acid) sequencers have had the ability to generate billions of short reads rapidly and inexpensively. The Illumina/Solexa sequencing technology typically generates 50-200 million 32-100 bp reads on a single run of the machine [13], which is transforming genomic science. These new machines are quickly becoming the technology of choice for whole-genome sequencing and for a variety of sequencing-based assays, including gene expression, DNA-protein interaction, human resequencing and RNA splicing studies [7].

In resequencing, a reference genome is already available for the species and one is interested in comparing short reads obtained from the genome of one or more donors (individual members of the species) to the reference genome. Therefore, the first step in any kind of analysis is the mapping of short reads to a reference genome.

How to map large amount of short reads to a reference sequence (e.g., the human genome) has become a challenging topic to the existing sequence alignment programs. A lot of new alignment algorithms have been developed to meet the requirement of efficient and accurate short read mapping.

Current available main algorithms for short read alignment include Bowtie [9], SOAP [10], SOAP2 [11], MAQ [12], BWA [13], mrFAST [14], mrsFAST [15], Novoalign [16] and SHRiMP [17], etc.

There are 4 types of the DNA nucleotides, namely, A (adenine), C (cytosine), G (guanine) and T (thymine), which are utilized to encode DNA. Due to sequencing errors and/or genetic variations, many reads map to the reference sequence approximately but not exactly, and therefore, to map a read to the reference sequence, read mapping programs should allow a certain number of mismatches between the read and a candidate location.

But current solutions for short-read alignment have limitations while implementing in an actual alignment application. SOAP [10], Novoalign [16], etc. can be easily parallelized with multi-threading, but large memory are usually necessary for building an index for the human genome [13]. SOAP2 [11], MAQ [12], Bowtie [9] and BWA [13] have few capabilities to align reads with insertions or deletions. If number of mismatches increases, it will take more time to align billions of reads to a large reference, and the accuracy will be reduced.

In this paper, we propose an efficient hierarchical alignment algorithm using Histogram-based Features and Smith-Waterman algorithm (HF-SW) that can tolerate moderate mismatches. For a given short read, first, a fast histogram search method is used to scan the reference sequence. Most of locations in reference sequence with low similarity will be excluded for latter searching. The Smith-Waterman alignment algorithm [18] is then applied to each remainder location to search for the exact matching. The effects will be demonstrated by using simulated data as well as real data.

This paper is organized as follows. In Section II, we will first introduce the proposed alignment algorithm using histogram-based features for short read in detail. Experimental results using both simulated data and real data

 ISBN: 978-1-61208-227-1

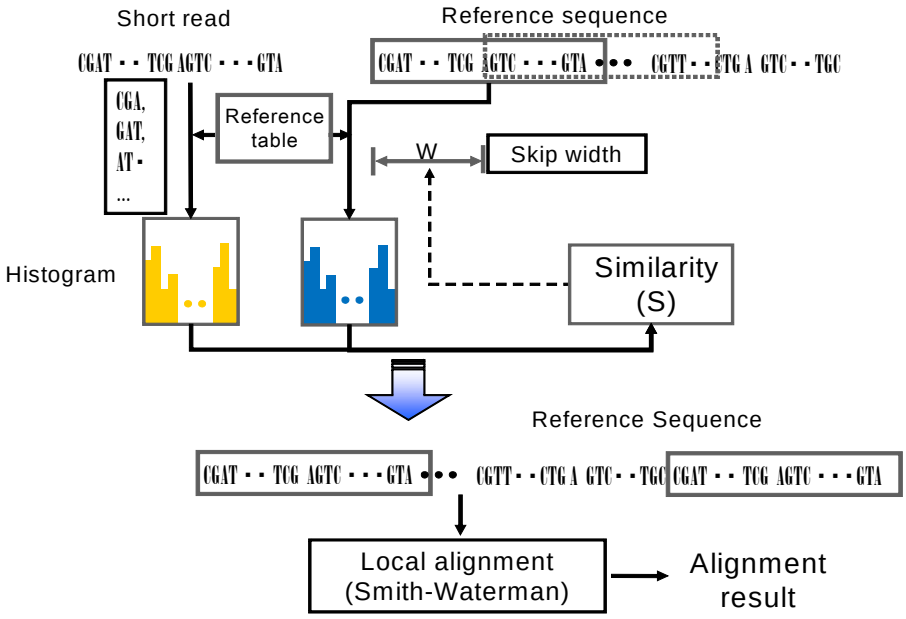


Figure 1. Processing steps of proposed method.

will be discussed in Section III. Finally, conclusions are given in Section IV.

II. PROPOSED METHOD

In this paper, we present a new short read alignment algorithm for short reads mapping in a large size of reference sequence. Histogram-based features of a given short read are firstly used to compare with the reference sequence and similarity scores would be obtained. Only the locations whose similarities exceeded a given threshold are then aligned using exhaustive Smith-Waterman dynamic programming algorithm [18].

TABLE I. REFERENCE TABLE.

CCC	CCT	CCG	CCA	CTC	CTT	CTG	CTA
0	1	2	3	4	5	6	7
CGC	CGT	CGG	CGA	CAC	CAT	CAG	CAA
8	9	10	11	12	13	14	15
TCC	TCT	TCG	TCA	TTC	TTT	TTG	TTA
16	17	18	19	20	21	22	23
TGC	TGT	TGG	TGA	TAC	TAT	TAG	TAA
24	25	26	27	28	29	30	31
GCC	GCT	GCG	GCA	GTC	GTT	GTG	GTA
32	33	34	35	36	37	38	39
GGC	GGT	GGG	GGA	GAC	GAT	GAG	GAA
40	41	42	43	44	45	46	47
ACC	ACT	ACG	ACA	ATC	ATT	ATG	ATA
48	49	50	51	52	53	54	55
AGC	AGT	AGG	AGA	AAC	AAT	AAG	AAA
56	57	58	59	60	61	62	63

Figure 1 shows the processing steps of our proposed method. When an unknown short read is input, it will be divided into small sequence, for instance, ACT and CGG, etc. A small sequence can be considered as a three dimensional vector. This processing overlaps over the entire short read. After that, the histogram feature is calculated. There are only 4 types of DNA bases, so the number of combination of 3-dimensional vector is 64. A reference table with the size of 64 is shown in Table I, by which the index number of the 3-dimensional vector is very easy and fast to be determined. The number of vectors with same index number in each separate partial sequence is counted and feature vector histogram is easily generated, and it is used as histogram feature of the short read.

In the mapping stage, the windows are applied to both the short read and the reference sequence. Corresponding histogram-based features of the short read and the partial sequence of the reference sequence in the window are generated as described above. The similarity between these histograms is then calculated. If the similarity exceeded a threshold value given previously, the location candidate will be detected and located. Otherwise, the window on the reference sequence will be skipped to the next position determined by the similarity in current position and the threshold value. In the last step, the window on the reference sequence is shifted forward and the mapping proceeds.

Here, histogram intersection is used as the similarity measure [20], and is defined as formula (1).

 ISBN: 978-1-61208-227-1

$$S_{SR} = S(H_S, H_R) = \frac{1}{N}\sum_{l=1}^{L}\min(h_{Sl}, h_{Rl}) \quad (1)$$

where h_{Sl}, h_{Rl} are the numbers of feature vectors contained in the *l-th* bin of the histograms for the short read and the partial reference sequence, respectively, L is the number of histogram bins, and N is the total number of feature vectors contained in the histogram. The skip width w is shown by formula (2).

$$w = \begin{cases} floor(N(\theta - S_{SR})) + 1 & (S_{SR} < \theta) \\ 1 & otherwise \end{cases} \quad (2)$$

where *floor(x)* means the greatest integral value less than x, and θ is a given threshold.

When reference sequence scanning is finished, location candidates whose similarities exceeded a given threshold are selected. The Smith-Waterman alignment algorithm [18] is then applied to each remainder location to search for the exact matching.

III. EXPERIMENTS AND DISCUSSIONS

To evaluate our proposed short read alignment algorithm using Histogram-based Features and Smith-Waterman algorithm (HF-SW), we compared its performance with one of the main short read alignment algorithm named Burrows-Wheeler Alignment tool (BWA) which achieves better performance than other main alignment algorithms such as Bowtie [7], SOAP2 [9], and MAQ [10].

BWA is a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. For short read alignment against the human reference genome, BWA is an order of magnitude faster than MAQ while achieving similar alignment accuracy [13].

We performed all of the experiments on a conventional PC@3.2GHz (12G memory). The algorithm was implemented in ANSI C.

A. Evaluation on simulated data

We simulated reads from the human genome using the wgsim program that is included in the SAMtools package [19] and ran the both programs to map the reads back to the human genome. Because the exact coordinate of each read, we are able to calculate the alignment error rate.

Table II shows that HF-SW achieved similar alignment accuracy with error rate of 0.117% as BWA at short read length of 70. HF-SW is more accurate than BWA when short read length is longer than 125.

The mapping time spending of HF-SW algorithm at 70 bp is about 854 seconds, which is similar with BWA. As shown in Table II, HF-SW algorithm can finish mapping only 937 seconds at 125 bp, which is about 2.2 times faster than BWA algorithm.

TABLE II. COMPARISON BETWEEN BWA AND PROPOSED METHOD USING EMULATED DATA.

Algorithm	Time(s)	Err(%)
BWA-32	569	0.3
HF-SW-32	746	0.53
BWA-70	1093	0.12
HF-SW-70	854	0.117
BWA-125	2104	0.05
HF-SW-125	937	0.044

TABLE III. COMPARISON BETWEEN BWA AND PROPOSED METHOD USING REAL DATA.

Algorithm	Time(h)	Conf(%)
BWA-51	3.2	88.9
HF-SW-51	3.15	88.7

B. Evaluation on real data

To evaluate the performance on real data, we downloaded about 12.2 million pairs of 51 bp reads from European Read Archive (AC:ERR000589). These reads were produced by Illumina for NA12750, a male included in the 1000 Genomes Project. Reads were mapped to the human genome NCBI build 36.

As shown in Table III, HF-SW confidently mapped 88.7% of all reads in 3.15 hours, which achieved similar performance compared with BWA algorithm.

IV. CONCLUSIONS

In this paper, we proposed a novel short read alignment algorithm combining histogram features and Smith-Waterman dynamic programming algorithms. Experimental results using emulated data as well as real data show proposed alignment algorithm will give more robust resulting and the proposed method is more efficient compared with conventional algorithms for short read alignment.

 ISBN: 978-1-61208-227-1

REFERENCES

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., "The sequence of the human genome," Science, vol. 291, no. 5507, pp. 1304 -1351, 2001.

[2] F. S. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from Large-Scale Biology," Science, vol. 300, no. 5617, pp. 286-290, 2003.

[3] GenBank, ftp://ftp.ncbi.nih.gov/genbank/.

[4] EMBL, http://www.embl.org/

[5] DDBJ, http://www.ddbj.nig.ac.jp/

[6] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

[7] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," Nature Biotechnology, vol. 27, 2009, pp. 455-457.

[8] Q. Chen, K. Kotani, F. Lee, and T. Ohmi, "A Fast Retrieval of DNA Sequences Using Histogram Information," 2009 Int'l Conf. on Future Information Technology and Management Engineering (FITME 2009), pp. 529-532, Sanya, China, Dec., 2009.

[9] B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," Genome Biol., vol. 10, 2009, R25.

[10] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," Bioinformatics, vol. 24, 2008, pp. 713-714.

[11] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment," Bioinformatics, vol. 25, 2009, pp. 1966-1967.

[12] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," Genome Res., vol. 18, 2008, pp. 1851-1858.

[13] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics," vol. 25, no. 14, 2009, pp. 1754–1760.

[14] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, et al., "Personalized copy number and segmental duplication maps using next-generation sequencing," Nature Genetics, vol. 41, 2009, pp. 1061-1067.

[15] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, and S C. Sahinal, "mrsFAST: a cache-oblivious algorithm for short-read mapping," Nature Methods, vol.7, 2010, pp. 576-577.

[16] Novocraft, http://www.novocraft.com/.

[17] K.R. Rasmussen, J. Stoye, and E. W. Myers, "Efficient q-gram filters for finding all e-matches over a given length," Lecture Notes in Computer Science, Springer, vol. 3500, 2005, pp. 189-203.

[18] T. F. Smith and M. S.Waterman, "Identification of common molecular subsequences", Journal of Molecular Biology, vol. 47, pp. 195-197, 1981.

[19] SAMtools, http://samtools.sourceforge.net.

[20] V.V. Vinod and H. Murase, "Focused color intersection with efficient searching for object extraction", Pattern Recognition, vol. 30, no.10, 1997, pp. 1787-1797.

[21] 1000 Genomes Project, http://www.1000genomes.org.

 ISBN: 978-1-61208-227-1

AgileKDD

An Agile Knowledge Discovery in Databases Process Model

Givanildo Santana do Nascimento
Federal University of Sergipe
São Cristóvão, Brazil
gsnascimento@petrobras.com.br

Adicinéia Aparecida de Oliveira
Federal University of Sergipe
São Cristóvão, Brazil
adicineia@ufs.br

***Abstract* — In a knowledge-based society, transforming data into information and knowledge to support the decision-making process is a crucial success factor for all the organizations. In this sense, the mission of Software Engineering is to build systems able to process large volumes of data, transform them into relevant knowledge and deliver them to customers to enable them to make the right decisions at the right time. However, companies still fail to determine a process model to be used in their Knowledge Discovery in Databases and Business Intelligence projects. This article introduces the AgileKDD, an agile and disciplined process for developing systems capable of discovering the knowledge hidden in databases, built on top of the Open Unified Process, KDD Process and CRISP-DM. A case study shows that AgileKDD can increase the success factor of projects whose goal is to develop Knowledge Discovery in Databases and Business Intelligence applications.**

Keywords – Knowledge Discovery in Databases; Business Intelligence; Agile Software Development; Software Process.

I. INTRODUCTION

The Organization for Economic Cooperation and Development (OECD) defined knowledge-based economies as: "economies which are directly based on the production, distribution and use of knowledge and information" [1]. In knowledge-based economies, the global competition is becoming increasingly based on the ability to transform data into information and knowledge in an effective way. Knowledge is equated with the traditional factors of production - land, capital, raw materials, energy and manpower - in the process of wealth creation. Thus, data, information and knowledge constitute key assets for all organizations working in this economic model.

Knowledge Management, Data Mining (DM), Knowledge Discovery in Databases (KDD) and, more generally, Business Intelligence (BI) are key concepts in a knowledge-based economy. BI applications have vital importance for many organizations and can help them manage, develop and share their intangible assets such as information and knowledge, improving their performance. For instance, investments made by Continental Airlines in BI had a Return on Investment (ROI) of 1000% due to increased revenue and reduced costs [2].

However, companies still face problems in determining a process model to be used to develop KDD and BI applications. As business requirements become more dynamic and uncertain, the traditional static, bureaucratic and heavy processes may not be able to deal with them. Recent researches have demonstrated that waterfall lifecycles and traditional software development processes are not successful in BI because they are unable to follow the dynamic requirement changes in a rapidly evolving environment [3]. As a software process is mandatory for KDD and BI development, one possible solution is to use an agile process, which is typically characterized by flexibility, adaptability, face-to-face communication and knowledge sharing.

This article presents AgileKDD, an agile software process designed to guide the KDD and BI applications development in a manner suitable for the current ever-changing requirement environments. The next sections are organized as follows: Section 2 describes the techniques for transforming raw data into information and knowledge. The Section 3 presents the agile software development processes. Section 4 presents the AgileKDD and a case study implemented to verify the AgileKDD applicability. Then, Section 5 presents related work, and, finally, Section 6 presents the conclusion and future work.

II. TRANSFORMING DATA INTO INFORMATION AND KNOWLEDGE

Raw data evolve into information and knowledge as they receive degrees of association, context and meaning [4]. The knowledge gained from the interpretation of data and information drives the knower to action, so knowledge is an important asset for organizations that operate in knowledge-based economies and markets. BI, as well as KDD, has the goal of transforming raw data into knowledge in order to support the decision-making process.

A. Knowledge Discovery in Databases

KDD is a nontrivial process of identifying valid, novel, potentially useful and understandable patterns in data [5]. The discovered knowledge must be correct, understandable by human users and also interesting, useful or new. In addition, the knowledge discovery method must be efficient, generic and flexible (easily changeable).

The KDD systematization effort has resulted in a variety of process models, including the KDD Process [5] and the Cross-Industry Standard Process for Data Mining (CRISP-DM) [6]. They are the most widely used in KDD projects and the most frequently cited and supported by tools. These two processes are considered the *de facto* standards in the

 ISBN: 978-1-61208-227-1

KDD area. Several other process models were derived from them. Figure 1 shows the evolution of 14 DM process models and methodologies. KDD Process can be pointed out as the initial approach and CRISP-DM as the central approach of the evolution diagram [7]. Most of the process models are based on them.

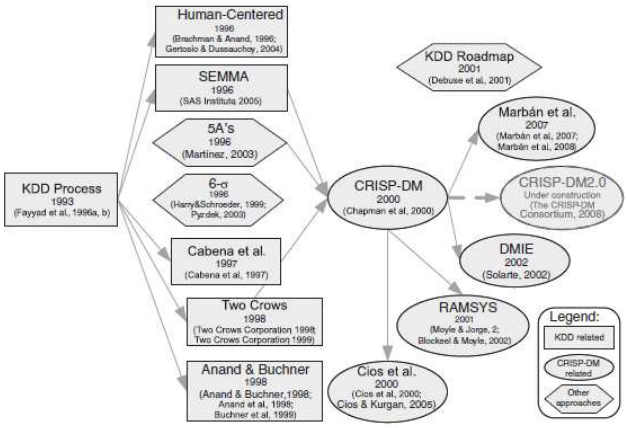


Figure 1. Evolution of data mining process models (Source: [7])

The KDD process models created between 1993 and 2008 were discussed in detail in a survey by Kurgan and Musilek [8] and then categorized by Mariscal, Marbán and Fernández [7] into three groups: (1) KDD related approaches; (2) CRISP-DM related approaches; (3) Other approaches.

Sometime later Alnoukari and El Sheikh [9] continued the older surveys done by Kurgan and Musilek [8] and Mariscal, Marbán and Fernández [7], and proposed a different categorization to the KDD process models: (1) Traditional approach; (2) Ontology-based approach; (3) Web-based approach; (4) Agile-based approach, which integrates agile processes and methodologies with traditional approaches. The main process models in this category are Adaptive Software Development – Data Mining (ASD-DM) [10] and Adaptive Software Development – Business Intelligence (ASD-BI) [1].

Thus, the knowledge discovery process models are evolving from traditional to agile processes, becoming more adaptive, flexible and human-centered [9]. However these processes still lack software engineering capabilities such as requirements management, project management and changes management.

B. Business Intelligence

Business Intelligence is an Information Technology (IT) framework vital for many organizations, especially those which have extremely large amounts of data, which can help organizations manage, develop and communicate their assets such as information and knowledge [2]. According to Mariscal, Marbán and Fernández [7], BI is a broad category of applications and technologies for gathering, storing, analyzing and providing access to data to help enterprise users make better business decisions.

The number of BI projects has grown rapidly worldwide according to Gartner Group annual reports. BI has been on the list of the top ten priorities in IT since 2005 and was at the top of this list for four consecutive years, from 2006 to 2009. In a broader sense, companies have understood that the information and knowledge provided by BI applications are essential to increase their effectiveness, support competitiveness and innovation. Thus, investments into data mining BI applications grew by 4.8% from 2005 to 2006 and by 11.2% from 2007 to 2008 [7] [11].

However, not all KDD and BI results are positive. Regardless of the priority and budgets growth, neither all the projects results were delivered [7] [12]. Many BI projects had failed to achieve their goals or were canceled because they were unable to follow the dynamic requirement changes in rapidly evolving environments. BI left the top of the list of priorities in IT and, in 2010 and 2011, dropped to the fifth position. Technologies with higher productivity, lower risk and faster ROI were prioritized instead [13].

Moreover, many companies still develop BI applications without the guidance of a software process. As any software projects, BI projects need a software process to succeed. Also, the dynamic business requirements, the needs of faster ROI and fluid communication between stakeholders and the team led to agile process as one possible solution.

III. AGILE SOFTWARE ENGINEERING PROCESSES

A software process provides an ordered sequence of activities related to the specification, design and implementation as well as validation and development of software products, transforming user expectations into software solutions [14]. According to Pressman [15], the software processes set the context in which technical methods are applied, the work artifacts (models, documents, data, reports, forms) are produced, the milestones are established, quality is assured and changes are managed.

The traditional software development processes are characterized by rigid mechanisms with a heavy documentation process, which make it difficult to adapt to a high-speed, ever-changing environment [16]. Agile approach is one answer to the software engineering chaotic situation, in which projects are exceeding their time and budget limits, requirements are not fulfilled and, consequently, leading to unsatisfied customers [17].

The Manifesto for Agile Software Development [18] defines the values introduced by the agile software processes. Based on these values, agile processes are people-oriented and have the customer satisfaction as the highest priority through the early and continuous delivery of functioning software. Agile approaches are best fit when requirements are uncertain or volatile; this can happen due to business dynamics and rapidly evolving markets. It is too difficult to practice traditional plan-oriented software development in such unstable environments [16].

Open Unified Process (OpenUP) is a variation of the Unified Process (UP) [19] that applies agile, iterative and incremental approaches within a structured lifecycle. OpenUP is a low-ceremony process that can be extended to address a broad variety of project types [20]. OpenUP has

 ISBN: 978-1-61208-227-1

compliance with the Manifesto for Agile Software Development, is minimal, complete and extensible. Moreover, it increases collaboration and continuous communication between project participants, more than formalities and comprehensive documentation [21].

The development of BI solutions must be guided by a software process. Therefore, it is mandatory to define processes that address aspects of KDD and BI, as well as the disciplines introduced by the software engineering process models. By the other hand, traditional processes are not successful in BI because they are unable to follow requirements in ever-changing environments [3]. Hence, one possible solution is to use an agile process, which is typically characterized by flexibility, adaptability, communication and knowledge sharing.

IV. AGILEKDD

AgileKDD is an agile and disciplined process for the development of KDD and BI solutions. CRISP-DM and KDD Process provide to AgileKDD the activities related to knowledge discovering. OpenUP provides the lifecycle, the phases and the disciplines, which are requirements, architecture, development, test, project management and changes management. OpenUP also adds the agile software development core values and principles, without giving up the management disciplines. The personal effort on an AgileKDD project is organized in micro-increments. They represent small work units that produce measurable steps in the project progress. The process applies intensive collaboration between the actors as the system is built incrementally. These micro-increments provide extremely short cycles of continuous feedback to identify and resolve problems before they become threats to the projects.

AgileKDD divides the projects in planned iterations with fixed time boxes, usually measured in weeks. The iterations drive the team to deliver incremental value to stakeholders in a predictable manner. Iteration plan defines what must be delivered during the iteration and the result is a demonstrable or deliverable piece of the KDD or BI solution. The AgileKDD lifecycle provides stakeholders and project team visibility and decision points at various milestones, until a working application is fully delivered to stakeholders. Figure 2 presents an overview of AgileKDD, highlighting its phases and activities.

The Inception (I) phase has the aim of developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the BI project from the customer's viewpoint. In this phase the project vision and plans are defined and agreed by all project participants. Also, in inception the target data set, or subset of variables and data samples, is selected. The knowledge discovery processes will be performed on the selected target data set. The data quality is a critical success factor for any BI project, so it is verified in Inception phase to indicate the project feasibility and quality constraints. Project management activity consists on high level project planning and governance concerns. Changes and configuration management activity is related to the version control of all the project artifacts, including documentation, sources and binaries.

The Elaboration (E) phase is responsible for the system's architecture and design, data modeling and applications integration.

Once data structures are modeled, the Construct (C) phase starts with Extract-Transform-Load (ETL) activity. ETL routines are built to extract, clean, integrate, transform and load the selected target data into databases. Also, ETL perform data cleaning to removes noise and decide on strategies for handling missing data fields. Thus, the DM techniques that best fit to the data are selected and applied to the information. DM tools search for meaningful patterns in data, including association rules, decision trees and clusters. The team can significantly aid the DM method by correctly performing the preceding steps. The reports, charts and dashboards are built to allow user information access. The verification and validation activities guarantee that the data was extracted, loaded and processed correctly, according to business objectives.

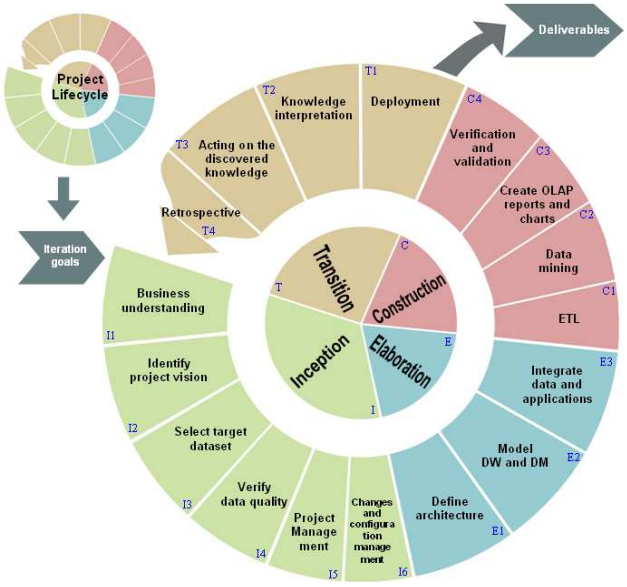


Figure 2. AgileKDD phases and lifecycle.

In Transition (T) phase the deployment of both software and knowledge takes place, the knowledge is interpreted, actions are created and the retrospective discusses lessons learnt during the project to promote continuous process improvement. Interpreting mined patterns involve visualization and storage of the extracted knowledge into knowledge bases, or simply documenting and reporting it to interested parties. This activity also includes checking for and resolving potential conflicts with previously believed knowledge. The AgileKDD process can involve significant iteration, interaction and can contain loops between any phases.

AgileKDD disciplines are the same of OpenUP: requirements, architecture, development, test, project management and configuration and changes management.

 ISBN: 978-1-61208-227-1

Table I shows the AgileKDD disciplines, their purposes and suggested work products.

During a full project cycle, most of the requirements discipline effort is concentrated in the inception phase. The architecture is the main discipline during the elaboration phase. In the same phase, the development is intensified from the definition of the system architecture and continues as the main discipline of construction phase. The tests occur mainly in verification and validation activity of construction phase. The project management discipline is concentrated predominantly in the inception phase. The configuration and change management has greater prevalence in inception and transition phases. Each discipline can be related to a set of work products created during the process phases.

TABLE I. AGILEKDD DISCIPLINES

Discipline	Purpose	Work products[a]
Requirements	Elicit, analyze, specify, validate and manage the requirements for the system being developed.	Vision document. Initial project glossary. Prototypes.
Architecture	Define an architecture for the system components.	Software architecture description. DW and DM models.
Development	Design and implement a technical solution adherent to the architecture that meets the requirements.	Software components. Integrated software increment.
Test	Validate system maturity through the design, implementation, execution and evaluation of tests.	Plan and test procedure. Test record.
Project management	Instruct, assist and support the team, helping them to deal with risks and obstacles faced when building software.	Project plan. Feasibility and risk evaluation.
Configuration and change management	Controlling changes in artifacts, ensuring a synchronized evolution of the set of artifacts that make a software system.	Work items list.

a. All the work products are optional. Only the necessary artifacts must be produced.

AgileKDD applicability has been verified by a case study in oil and gas area. The process was applied to a KDD and BI project that deals with Reservoir Evaluation data and afforded the early delivery of DM results two months after the project kickoff.

The first iteration was dedicated exclusively to the project inception. This phase aimed to identify the product requirements, to the communication with customer, project management, configuration and change management. The second iteration aimed to delivery data mining results related to Reservoir Evaluation (RA) data. The third iteration aimed to calculate the RA performance indicators and present them to users in dashboards. The fourth iteration aimed to deliver the online analytical processing (OLAP) features, including reports, graphs, and *ad hoc* exploration of the data warehouse.

It was observed that AgileKDD process was able to guide the product development since the beginning of the inception iteration to the transition phase of the last iteration performed. At the end of the case study, it was verified that some adjustments were needed in the process to improve its fitness for BI and KDD systems projects. The observations and identified adjustments needs helped to improve the process final version.

V. RELATED WORK

The main work that applies agile methodologies to KDD and BI is [1]. Alnoukari [16] discusses BI and Agile Methodologies for knowledge-based organizations in a cross-disciplinary approach. Alnoukari [22] introduces Adaptive Software Development – Business Intelligence (ASD-BI), a knowledge discovery process model based on Adaptive Software Development agile methodology. Likewise, Alnoukari, Alzoabi and Hanna [10] defined Adaptive Software Development – Data Mining (ASD-DM) Process Model. The main difference between this work and these is the fact that AgileKDD is a software process, not a methodology. As a process, AgileKDD defines what to do instead how to do KDD and BI development. Also, the process proposed by this work defines lifecycle, roles, activities, inputs and outputs regarding agile KDD and BI application development. Moreover, the process AgileKDD contains management disciplines like project, changes and requirements management, which were inherited from OpenUP.

Three surveys about DM and knowledge discovery process models and methodologies are discussed and compared by Mariscal, Marbán and Fernández [7], Kurgan and Musilek [8] and Alnoukari and El Sheikh [9]. All the process models and methodologies presented by these works focus on DM and knowledge discovery, and do not consider other BI components. As BI is more comprehensive than data mining, this work focuses on an agile process modeled to address both KDD and BI software projects, in an adaptable, flexible and systematic manner.

The objective of this work was building a software process capable of guiding KDD and BI projects in an agile and adaptive way. The cornerstone of this work was a software process, the OpenUP, created from the UP, inheriting the maturity of this process in an agile approach. Existing works relied on brand new agile methodologies, which lack of software engineering capabilities and were not scientifically proved yet.

VI. CONCLUSION AND FUTURE WORK

A software process is mandatory for KDD and BI development. However, traditional software development processes are not successful in KDD and BI because they are unable to follow the dynamic requirement changes in an ever-changing environment. Agile processes fit in KDD and BI better than traditional processes because they are characterized by flexibility, adaptability, communication and knowledge sharing.

This work presented AgileKDD, a KDD and BI process based on the Open Unified Process. AgileKDD applicability has been verified by a case study and the results indicate that software development organizations may apply AgileKDD

 ISBN: 978-1-61208-227-1

in implementing knowledge discovery projects. The process brought such benefits as more customer satisfaction through early and continuous delivery of functioning software, better communication between team members and reduced project failure risks.

The main contribution of AgileKDD is its ability to guide the BI solutions development according to the practices present in agile software development processes. AgileKDD can increase the projects success factor and customer satisfaction. The process can be used to guide BI and KDD applications projects in scenarios of continuous requirements evolving and early ROI need.

Future work can validate AgileKDD by more case studies in different areas and improve its capabilities to store the knowledge discovered in ontology bases or knowledge bases.

REFERENCES

[1] A. El Sheikh and M. Alnoukari, Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications. IGI Global, 2012. pp. 1-370.

[2] M. Alnoukari, H. Alhawasli, H. Alnafea, and Amjad Zamreek, "Business Intelligence: Body of Knowledge" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 1-13.

[3] D. Larson, "Agile Methodologies for Business Intelligence" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 101-119.

[4] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, Sixth Edition. Pearson, 2010. pp. 1200 pp.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. "From data mining to knowledge discovery: an overview" in *Proc. Advances in Knowledge Discovery and Data Mining*, 1996, pp. 1–34.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[7] G. Mariscal, O. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies". *The Knowledge Engineering Review*, vol. 25, 2010, pp. 137-166.

[8] L. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models". *The Knowledge Engineering Review*, vol. 21, 2006, pp. 1-24.

[9] M. Alnoukari and A. El Sheikh, "Knowledge Discovery Process Models: From Traditional to Agile Modeling" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 72-100.

[10] M. Alnoukari, Z. Alzoabi, and S. Hanna, "Applying adaptive software development (ASD) agile modeling on predictive data mining applications: ASD-DM Methodology" in *IEEE Proceedings of International Symposium of Information Technology*, 2008, pp. 1083–1087.

[11] M. McDonald, M. Blosch, T. Jaffarian, L. Mok, and S. Stevens, "Growing It's Contribution: The 2006 Cio Agenda". Gartner Group, 2006.

[12] Gartner Group, "Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007". 2005. Available: http://www.gartner.com/it/page.jsp?id=492112 [retrieved: Out., 2012].

[13] Gartner Group, "Gartner Executive Programs Worldwide Survey of More Than 2,000 CIOs Identifies Cloud Computing as Top Technology Priority for CIOs in 2011". 2011. Available: http://www.gartner.com/it/page.jsp?id=1526414 [retrieved: Out., 2012].

[14] I. Sommerville, Software Engineering. Addison Wesley, 2006, pp. 864.

[15] R. Pressman, Software Engineering: A Practioner's Approach. McGraw-Hill, 2005.

[16] M. Alnoukari, "Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications" in *CEPIS UPGRADE: The European Journal for the Informatics Professional*, vol. 12, pp. 56–59, 2011. Available: http://www.cepis.org/upgrade/media/III_2011_alnoukari1.pdf [retrieved: Out., 2012].

[17] Z. Alzoabi, "Agile Software: Body of Knowledge" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. IGI Global, 2012. pp. 14-34.

[18] K. Beck et al Manifesto for Agile Software Development. 2001. Available: http://agilemanifesto.org [retrieved: Out., 2012].

[19] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide. Addison Wesley, 1999.

[20] H. Hristov, Introduction to OpenUP. 2011. Available: http://epf.eclipse.org/wikis/openup/index.htm [retrieved: Out., 2012].

[21] S. Santos, OpenUP: Um processo ágil. 2009. Available: http://www.ibm.com/developerworks/br/rational/local/open_up/index.html [retrieved: Out., 2012].

[22] M. Alnoukari, "ASD-BI: A Knowledge Discovery Process Modeling Based on Adaptive Software Development Agile Methodology" in *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications.* IGI Global, 2012. pp. 1-13.

 ISBN: 978-1-61208-227-1

An Improved Face Recognition Algorithm Using Adjacent Pixel Intensity Difference Quantization Histogram and Markov Stationary Feature

Feifei Lee, Koji Kotani*, Qiu Chen, and Tadahiro Ohmi

New Industry Creation Hatchery Center, Tohoku University
** Department of Electronics, Graduate School of Engineering, Tohoku University*
Aza-Aoba 6-6-10, Aramaki, Aoba-ku, Sendai 980-8579, JAPAN
e-mail: fei@fff.niche.tohoku.ac.jp

***Abstract*—Previously, we have proposed a robust face recognition algorithm using adjacent pixel intensity difference quantization (APIDQ) histogram combined with Markov Stationary Features (MSF), so as to add spatial structure information to histogram. We named the new histogram feature as MSF-DQ feature. In this paper, we employ multi-resolution analysis for the facial image to extract more powerful personal feature. After a set of multi-resolution pyramid images is generated using sub-sampling, MSF-DQ features at different resolution levels are extracted from corresponding pyramid images. Recognition results are firstly obtained using MSF-DQ features at different resolution levels separately and then combined by weighted averaging. Publicly available AT&T database of 40 subjects with 10 images per subject containing variations in lighting, posing, and expressions, is used to evaluate the performance of the proposed algorithm. Experimental results show face recognition using proposed multi-resolution features is very efficient. The highest average recognition rate of 98.57% is obtained.**

Keywords-Face recognition; Adjacent pixel intensity difference quantization (APIDQ); Markov stationary feature (MSF); Multiresolution; Histogram feature

I. INTRODUCTION

In the last two decades, face recognition has been a hot research topic in artificial intelligence and pattern recognition area due to its potential applications in many fields such as law enforcement applications, security applications and video indexing, etc. As a more natural and effective person identification method compared with that using other biometric features such as voice, fingerprint, iris pattern, etc., a lot of face recognition algorithms have been proposed [1]-[14]. These algorithms can be roughly divided into two main approaches, that is to say, structure-based and statistics-based.

In the structure-based approaches [3][4], recognition is based on the relationship between human facial features such as eye, mouth, nose, profile silhouettes and face boundary. Statistics-based approaches [5][6][7] attempt to capture and define the face as a whole. The face is treated as a two dimensional pattern of intensity variation. Under this approach, the face is matched through finding its underlying statistical regularities. Principal component analysis (PCA) is a typical statistics-based technique [5]. However, these techniques are highly complicated and are computationally power hungry, making it difficult to implement them into real-time face recognition applications.

In [18][19], a very simple, yet highly reliable face recognition method called Adjacent Pixel Intensity Difference Quantization (APIDQ) Histogram Method is proposed, which achieved the real-time face recognition. At each pixel location in an input image, a 2-D vector (composed of the horizontally adjacent pixel intensity difference (dIx) and the vertically adjacent difference (dIy)) contains information about the intensity variation angle (θ) and its amount (r). After the intensity variation vectors for all the pixels in an image are calculated and plotted in the r-θ plane, each vector is quantized in terms of its θ and r values. By counting the number of elements in each quantized area in the r-θ plane, a histogram can be created. This histogram, obtained by APIDQ for facial images, is utilized as a very effective personal feature. Experimental results show a recognition rate of 95.7 % for 400 images of 40 persons (10 images per person) from the publicly available AT&T face database [20].

In [17][18], we combine the APIDQ histogram with Markov stationary feature (MSF), which was proposed in [19], so as to encode spatial structure information within and between histogram bins. The MSF extends the APIDQ histogram features by characterizing the spatial co-occurrence of histogram patterns using the Markov chain models and improves the distinguishable capability of APIDQ features to extra-bin distinguishable level [19]. The highest average recognition rate of 97.16% is obtained by using the publicly available database of AT&T [20]. It can be said that the extended MSF-DQ features is more robust for face recognition.

We can imagine that different MSF-DQ features are extracted with different resolutions of the image. Therefore, more comprehensive personal feature information can be obtained by combining multiple recognition results using multi-resolution analysis. In this paper, we employ multi-resolution analysis for the facial image to extract more powerful personal feature.

In Section II, we will first introduce Markov stationary feature (MSF) as well as the Adjacent Pixel Intensity Difference Quantization (APIDQ) histogram feature which

 ISBN: 978-1-61208-227-1

had been successfully applied to face recognition previously, and then describe proposed face recognition algorithm using multi-resolution MSF-DQ features in Section III. Experimental results will be discussed in Section IV. Finally, conclusions will be given in Section V.

II. RELATED WORKS

A. Markov Stationary features (MSF)

The Markov stationary feature (MSF) [19] extends the APIDQ histogram features by characterizing the spatial co-occurrence of histogram patterns using the Markov chain models and improves the distinguishable capability of APIDQ features to extra-bin distinguishable level. We will briefly introduce the MSF in this section.

Let p_k be a pixel in image I, the spatial co-occurrence matrix is defined as $C = (c_{ij})_{K\times K}$ where

$$c_{ij} = \#(p_1 = c_i, p_2 = c_j \| p_1 - p_2 | = d) / 2, \quad (1)$$

in which d (d=1 in this paper) indicates L_1 distance between two pixels p_1 and p_2, and c_{ij} counts the number of spatial co-occurrence for bin c_i and c_j.

The co-occurrence matrix c_{ij} can be interpreted in a statistical view. Markov chain model is adopted to characterize the spatial relationship between histogram bins.

The bins are treated as states in Markov chain models, and the co-occurrence is viewed as the transition probability between bins. In this way, the MSF can transfer the comparison of two histograms to two corresponding Markov chains.

The elements of the transition matrix *P* are constructed from the spatial co-occurrence *C* by formula (2).

$$P_{ij} = c_{ij} / \sum_{j=1}^{K} c_{ij} \quad (2)$$

The state distribution after n steps is defined as $\pi(n)$, and the initial distribution is $\pi(0)$, the Markov transition matrix obeys following rules [19].

$$\pi(n+1) = \pi(n)P,\ \pi(n) = \pi(0)P^n; \quad (3)$$
$$P^{m+n} = P^m P^n$$

where $\pi(0)$ is defined as

$$\pi(0) = c_{ii} / \sum_{i=1}^{K} c_{ii} \quad (4)$$

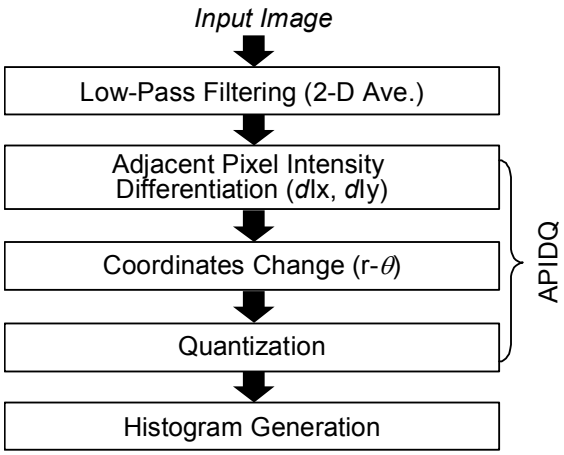


Figure 1. Processing steps of APIDQ histogram method.

According to the formula (3), we can get a distribution of π called a stationary distribution which satisfies

$$\pi = \pi P \quad (5)$$

The stationary distribution becomes the final representation of MSF. Obtaining the MSF of each image, the comparison of two histograms is transferred to the comparison of two corresponding Markov chains.

B. Adjacent Pixel Intensity Difference Quantization (APIDQ)

The Adjacent Pixel Intensity Difference Quantization (APIDQ) histogram method [15] has been developed for face recognition previously. Figure 1 shows the processing steps of APIDQ histogram method. In APIDQ, for each pixel of an input image, the intensity difference of the horizontally adjacent pixels (*d*Ix) and the intensity difference of the vertically adjacent pixels (*d*Iy) are first calculated by using simple subtraction operations shown as formula (6).

$$dIx(i,j) = I(i+1,j) - I(i,j)$$
$$dIy(i,j) = I(i,j+1) - I(i,j) \quad (6)$$

A calculated (*dIx*, *dIy*) pair represents a single vector in the *dIx-dIy* plane. By changing the coordinate system from orthogonal coordinates to polar coordinates, the angle θ and the distance *r* represent the direction and the amount of intensity variation, respectively. After processing all the pixels in an input image, the dots representing the vectors are distributed in the *dIx-dIy* plane. The distribution of dots (density and shape) represents the features of the input image.

 ISBN: 978-1-61208-227-1

Each intensity variation vector is then quantized in the r-θ plane. Quantization levels are typically set at 8 in θ-axis and 8 in *r*-axis (totally 50). Since *dIx*-*dIy* vectors are concentrated in small-r (small-*dIx*, -*dIy*) region, non-uniform quantization steps are applied in *r*-axis. The number of vectors quantized in each quantization region is counted and a histogram is generated. In the face recognition approach, this histogram becomes the feature vector of the human face.

The essence of the APIDQ histogram method can be considered that the operation detects and quantizes the direction and the amount of intensity variation in the image block. Hence the APIDQ histogram contains very effective image feature information. The MSF extends histogram based features with spatial structure information of images, and transfer the comparison of two histograms to two corresponding Markov chains.

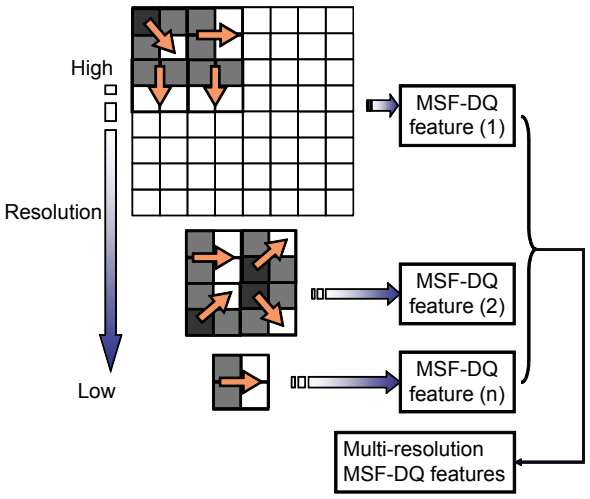


Figure 2. Multi-resolution MSF-DQ features.

III. PROPOSED FACE RECOGNITION ALGORITHM

A. *Multi-resolution analysis*

Because different MSF-DQ features are extracted with different resolutions of the image, more comprehensive personal feature information can be obtained by combining multiple recognition results using multi-resolution analysis. In this paper, we employ multi-resolution analysis for the facial image to extract more powerful personal features. As shown in figure 2, after a set of multi-resolution pyramid images is generated using sub-sampling, MSF-DQ features at different resolution levels are extracted from corresponding pyramid images. Recognition results are first obtained using MSF-DQ features at different resolution levels separately and then combined by weighted averaging.

B. *Proposed algorithm*

The procedure of proposed face recognition algorithm using APIDQ histogram combined with MSF is shown in figure 3. Low-pass filtering is first carried out before APIDQ using a simple 2-D moving average filter. This low-pass filtering is essential for reducing the high-frequency noise and extracting the most effective low frequency component for recognition. After multi-resolution pyramid images are generated using sub-sampling, APIDQ operations are implemented on respective images with different resolution and quantization region number corresponded to each 2x2 image block is calculated. Because each 2x2 image block can be regarded as a pixel of color c_i, the co-occurrence matrix for APIDQ can be computed according to formula (1).

The Markov transition matrix *P* is calculated by formula (2). Then the stationary distribution can be approximated by the average of each row $\vec{a}_i$ of A_n using formula (7).

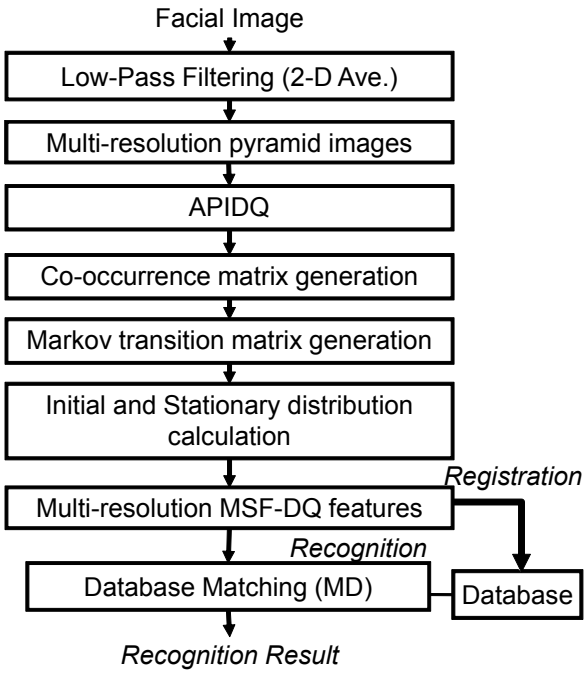


Figure 3. Proposed face recognition algorithm using multi-resolution MSF-DQ features.

$$\pi \approx \frac{1}{K} / \sum_{i=1}^{K} \vec{a}_i \text{, where } A_n = [\vec{a}_1, \cdots, \vec{a}_k]^T, \quad (7)$$

$$A_n = \frac{1}{n+1}(I + P + P^2 + \cdots + P^n) \quad (8)$$

 ISBN: 978-1-61208-227-1

Figure 4. Samples of the database of AT&T Laboratories Cambridge.

n =50 is used as same as in [19]. The initial distribution $\pi(0)$ can be obtained by formula (4). As shown in formula (9), the Markov stationary feature is defined as the combination of the initial distribution $\pi(0)$ and the stationary distribution π after n steps.

$$\vec{h}_{MSF-DQ} = [\pi(0), \pi]^T \tag{9}$$

We call MSF extension of APIDQ histogram as a MSF-DQ feature. The MSF-DQ feature made from each pyramid image is compared with those from the same resolution images in the database by calculating distances (d_i) between them using the same distance calculation formula as in [19]. Then the integrated distances (D) are obtained by weighted averaging as shown in the following formula (10).

$$D = \frac{\sum w_i d_i}{\sum w_i} \tag{10}$$

where w_i is weighting coefficient of the different resolutions, The best match is output as recognition result by searching the minimum integrated distance.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Data sets

The publicly available face database of AT&T Laboratories Cambridge [20] is used for the analysis and recognition experiments. Forty people with 10 facial images each, (totaling 400 images), with variations in face angles, facial expressions, and lighting conditions are included in the database. Each image has a resolution of 92x112. Figure 4 shows typical image samples of the database of AT&T Laboratories Cambridge. From the 10 images for each person, five were selected as probe images and the remaining five were registered as album images. Recognition experiments were carried out for 252 ($_{10}C_5$) probe-album combinations using the rotation method.

B. Experimental results

Comparison of recognition results are shown in Figure 5. Recognition success rates are shown as a function of filter size. The filter size represents the size of the averaging filter core. A size of F3, for instance, represents the filter using a 3x3 filter core. Figure 5 shows the comparison between the recognition results using different resolution MSF-DQ features separately and multi-resolution MSF-DQ features. Average recognition rate is shown here. “bin 50 (original DQ)” stands for the case that original APIDQ utilizes quantization table containing the number of bins of 50 in [15][16]. “bin42_s92x112”, “bin42_s46x56”, “bin42_s23x28”, and “bin42_s11x14” stand for the cases

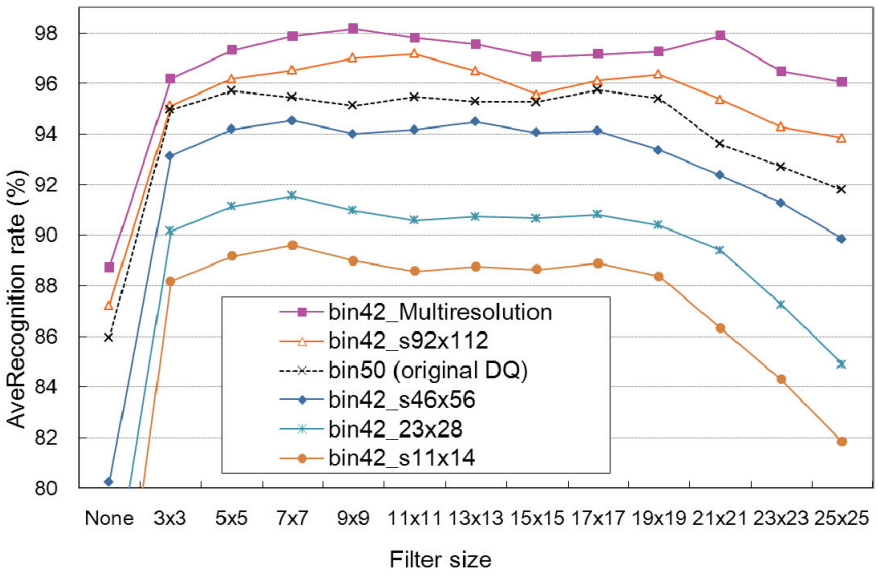


Figure 5. Comparison of results. Average recognition rate is shown here.

 ISBN: 978-1-61208-227-1

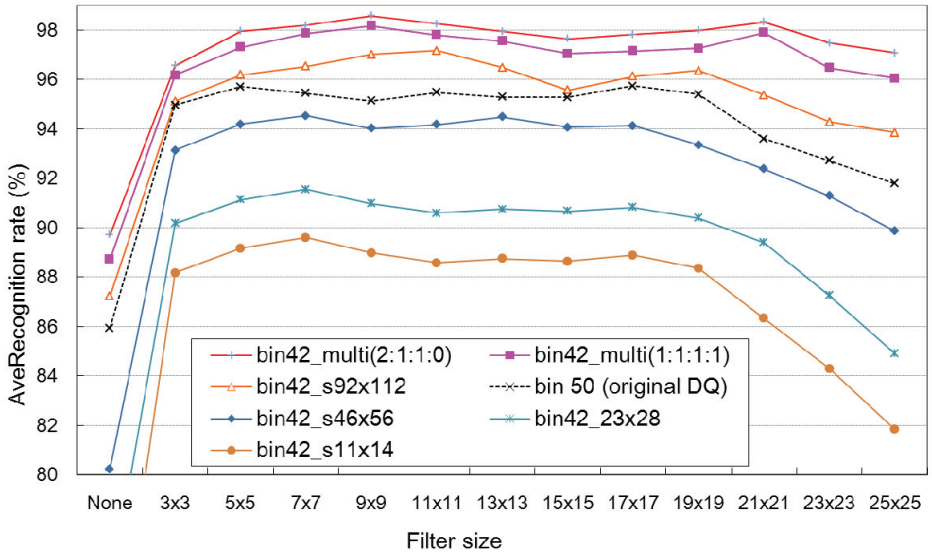


Figure 6. Comparison of results. Average recognition rate is shown here.

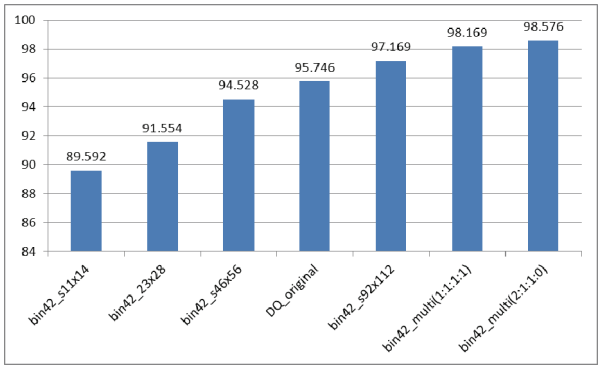


Figure 7. Comparison of results. Maximum average recognition rate is shown here.

using various resolution MSF-DQ features separately. "bin42_Multiresolution" stands for the case using multi-resolution MSF-DQ features proposed in this paper, which weighting coefficient at each resolution level is set as 1.

The best performance of the average recognition rate 97.16% [17][18] is obtained at original image size of 92x112 when using separate single-resolution MSF-DQ features. By using multi-resolution MSF-DQ features with the weighting coefficient at each resolution level of 1, highest recognition rate increases to 98.16%. It can be said that multi-resolution MSF-DQ features is more robust than single-resolution MSF-DQ features.

Figure 6 and 7 also show the results of using single-resolution solely, and those of using some combinations. Maximum of the average recognition rate 98.57% is achieved at the combination of weighting coefficients of 2:1:1:0 with the image resolutions of 92x112, 46x56, 23x28, 11x14, respectively. It can be considered too small image resolution give less contribution for feature generation.

V. CONCLUSION

In this paper, we improved our face recognition using MSF-DQ feature by employing multi-resolution analysis for the facial image to extract more powerful personal feature. Excellent face recognition performance as large as a 98.57% recognition rate has been achieved by using the publicly available database of AT&T. It can be said that multi-resolution MSF-DQ features is more robust for face recognition.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan, Grant-in-Aid for Scientific Research (C), No. 24500104, 2012-2015, and also by research grant from Support Center for Advanced Telecommunications Technology Research, Foundation (SCAT).

REFERENCES

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," Proc. IEEE, vol. 83, no. 5, 1995, pp. 705-740.

[2] S. Z. Li and A. K. Jain, "Handbook of Face Recognition," Springer, New York, 2005.

[3] R. Brunelli and T. Poggio, "Face recognition: features versus templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 10, Oct. 1993, pp. 1042-1052.

[4] L. Wiskott, J.M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 10, 1997, pp. 775-780.

 ISBN: 978-1-61208-227-1

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, Mar. 1991, pp. 71-86.

[6] W. Zhao, "Discriminant component analysis for face recognition," Proc. in the Int'l Conf. on Pattern Recognition (ICPR'00), Track 2, 2000, pp. 822-825.

[7] K.M. Lam, H. Yan, "An analytic-to-holistic approach for face recognition based on a single frontal view," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, 1998, pp. 673-686.

[8] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. on Neural Networks, vol. 13, no. 6, 2002, pp. 1450-1464.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenface vs. Fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, May 1997, pp. 711-720.

[10] B. Moghaddam, A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997, pp. 696–710.

[11] S. G. Karungaru, M. Fukumi, N. Akamatsu, "Face recognition in colour images using neural networks and genetic algorithms," Int'l Journal of Computational Intelligence and Applications, vol. 5, no. 1, 2005, pp. 55-67.

[12] Z. Liu, C. Liu, "Fusion of color, local spatial and global frequency information for face recognition," Pattern Recognition, vol. 43, Issue 8, Aug. 2010, pp. 2882-2890.

[13] H. F. Liau, K. P. Seng, L. M. Ang, and S. W. Chin, "New Parallel Models for Face Recognition," Recent Advances in Face Recognition, Edited by K. Delac etc., InTech, 2008.

[14] Q. Chen, K. Kotani, F. F. Lee, and T. Ohmi, "Face Recognition Using VQ Histogram in Compressed DCT Domain," Journal of Convergence Information Technology, vol. 7, no. 1, 2012, pp. 395-404.

[15] K. Kotani, F.F. Lee, Q. Chen, and T. Ohmi, "Face recognition based on the adjacent pixel intensity difference quantization histogram method," 2003 Int'l Symposium on Intelligent Signal Processing and Communication Systems, D7-4, 2003, pp. 877-880.

[16] F. F. Lee, K. Kotani, Q. Chen, T. Ohmi, "Face Recognition Using Adjacent Pixel Intensity Difference Quantization Histogram," Int'l Journal of Computer Science & Network Security, vol. 9, no. 8, 2009, pp. 147-154.

[17] F. F. Lee, K. Kotani, Q. Chen, T. Ohmi, "A Robust Face Recognition Algorithm Using Markov Stationary Features and Adjacent Pixel Intensity Difference Quantization Histogram," Proc. in 7th Int'l Conf. on Signal Image Technology & Internet Based Systems (SITIS 2011), France, 2011, pp. 334-339.

[18] F. F. Lee, K. Kotani, Q. Chen, T. Ohmi, "Face Recognition Using Adjacent Pixel Intensity Difference Quantization Histogram Combined with Markov Stationary Features," Int'l Journal of Advancements in Computing Technology, in press.

[19] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histograms: Image representation using Markov stationary features," Proc. in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08) , 2008, pp. 1-8.

[20] AT&T Laboratories Cambridge, The Database of Faces, at http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[21] P. J. Phillips, H. Wechsler, J. Huang, & P. Rauss. "The FERET database and evaluation procedure for face recognition algorithms," Image and Vision Computing J, vol. 16, no. 5, 1998, pp. 295-306.

 ISBN: 978-1-61208-227-1

Music Recommendation based on Text Mining

Ziwon Hyung[1], Kyogu Lee[1,2]
[1]*Music and Audio Research Group*
[2]*Advanced Institutes of Convergence Technology*
Seoul National University
Seoul, Korea
Email: ziotoss@gmail.com, kglee@snu.ac.kr

MyoungA Lee
High Technology Division
Lotte Data Communication Co.
Seoul, Korea
Email: maeng111@gmail.com

***Abstract*—Recommending music from millions of items is a challenging problem. In this paper, we propose a novel approach to recommending music given an textual input from the user. To this end, we first mine a large corpus of textual documents from the radio station's Internet bulletin board. Each document, written by a listener, contains a personal story associated with a song request. Assuming that the personal story contains the reason for the song request, we then perform the Latent Semantic Analysis (LSA) on these documents to find the document similarity, which we believe also indicates similar music preference. Our hypothesis is that when the two users request the same song, the situation or context in which they write the associated story is likely to be similar as well, and therefore the two stories will also be similar to each other. Using the mined documents that request the same song as a test set, we show that there is a positive correlation between the document similarity and song similarity, and thus it is possible to recommend music purely based on text mining and analysis.**

Keywords-text mining; Latent Semantic Analysis; music recommendation.

I. INTRODUCTION

Rapid growth in the volume of digital music data raised issues in selecting which music the user would like to listen to. This phenomenon, so-called the Paradox of Choice [13], shows that as the number of options grow, the effort in making a wise selection also increases, resulting in the selection process being a burden. Therefore, recommendation systems are becoming increasingly important due to their ability to filter out the unnecessary or unimportant data from the huge growing volume of accessible data [3]. While there are numerous approaches in music recommendation system, they can be generalized into two categories depending on how they retrieve new items: (1) collaborative filtering based recommender and (2) content-based recommender.

Collaborative filtering based music recommender identifies similar users or items based on prior purchase history and rating to recommend new items. An important requirement for this approach is that the selected item must have enough valid information provided by the users. As a consequence, it is prone to the so-called Cold Start problem [12], which with high probability misses the newly arrived items due to lack of information. Another problem is that the diversity of the recommended item is poor [17]. This problem can be explained by a phenomenon known as Long Tail [2]. Huge concentration of users is focused on popular items while only a small amount demand other items. According to the Digital Music Report 2012 [4], the combined sales of the top ten digital singles marked about 86.2 million copies. Considering the total amount of digital music sold, this number is significant. This indicates that the music industry follows the Long Tail phenomenon. Since collaborative filtering method is based on the preference of users as a criterion for recommendation, this results in using only a small portion of the music data when recommending new items.

Content-based filtering music recommender uses metadata such as genre, artist, and lyrics [9], [10], [16], and/or acoustic features [7], [8] to find similar items. While this approach is immune to the cold start problem and popularity bias of the CF approach, it faces other issues such as computational power. Since the music database is extremely large and still expanding, using content-based recommendation approach requires a huge amount of time to analyze the content and recommend similar music, and thus it is inefficient for commercial use. Another problem is that the system must be provided with an input music in order to compare the content and provide a recommendation list. This again leads to the cold start problem and also the paradox of choice.

While CF method and content-based method have its own issues, a common problem in both approach is that they neglect an important criterion; the user's situational information when one seeks to listen to music. Recently, people tend to write their daily situational information via social network services. From this observation, we thought of using such textual information to extract the contextual information when recommending music. The idea of our algorithm is to perform Latent Semantic Analysis (LSA) [6] on the documents retrieved from the radio station's Internet bulletin board to discover similar stories. The audience of the radio channel writes their own story in the bulletin board and requests a song they would like to listen to as a consequence of the story. In this paper, we will use the term document

 ISBN: 978-1-61208-227-1

to indicate the stories and song request written in the radio station's Internet bulletin board. Our hypothesis is that when people request the same song, the situation or context in which they write the associated story is likely to be similar as well. Since each document contains a song request as well, by discovering similar documents, the system can recommend the songs linked to the similar documents.

There has been several approach in using contextual information as a criterion for recommending music. However, using textual stories written in the radio station bulletin board for music recommendation has not been attempted to the authors' knowledge. Another contribution would be that by implementing this approach to many existing SNS could provide a song that suits the message.

The remainder of the paper is organized as follows. In the next section, we first summarize recent music recommendation algorithms and address the problems of current approaches. We also introduce the characteristic of the stories written in the Korea radio station bulletin board. In Section III, we explain our system in detail. In Section IV, we provide a statistical evaluation of the system and in Section V, we present the results. We conclude the paper with a summary and directions for future work in Section VI.

II. BACKGROUND

There are several approaches in the music recommendation field. Commonly, the methods can be grouped into three categories depending on the algorithm: collaborative filtering method, content-based filtering method, and a hybrid method. Since our focus is on taking consideration of the user's input without any content analysis, we will discuss about some of the CF methods in this section.

A. Collaborative Filtering-based Recommendation Systems

There are two different types of collaborative filtering method: the memory-based recommender system and the model-based recommender system. The memory-based recommender system again can be categorized into user-based CF and item-based CF depending on the focus of the algorithm. The user-based CF predicts the user's interest in a new item based on rating information from similar user profiles. The item-based CF works in a similar way but instead of using similar user profiles, it uses similarity between items [15].

On the other hand, the model-based recommender uses the collection of ratings to learn a model. Using the learned model, the expected rating for a new item is estimated. Some of the widely used models are the cluster model, Bayesian networks, statistical model, and machine learning models [1]. The performance of this method is greatly affected by the model and thus to create a model that improves the quality of the recommender system is still an ongoing issue.

While these CF methods are widely used in commercial nowadays, the effectiveness of the recommendation is questioned as the recommender systems confront some problems. Since the music database is extremely large, the rating information of the user is very sparse. This sparse user rating information can lead to biased suggestions. Another problem is that it neglects the contextual information of the listener. This problem has been tackled previously and will be explained in the following section.

B. Context-aware Recommendation Systems

Reynolds *et al.* introduce the need to take into consideration of the user's contextual information. Through a survey experiment, they showed that activity, one of the contextual information, has a great impact on the listener's mood. From this research, it was shown that the activity one is involved in has great impact on the choice of the music one wants to listen to [11].

Su *et al.* proposed another method on using contextual information to recommend music [14]. Their system uses contextual information such as heartbeat, body temperature, air temperature, noise volume, humidity, light, motion, time, season, and location. Along with this contextual information, their system performs a content analysis on the music data to build a pattern database which links music with the user. Using this link and the contextual information, it proved to provide a more effective recommendation list.

However, the system suggested by Su *et al.* has an off-line preprocessing step which is to generate the pattern database via content analysis. Again, this confronts a scalability issue. A more critical problem is that while all the suggested contextual information might implicitly infer an activity state, it doesn't actually indicate what activity one is doing. In order to overcome the listed problems we suggest a recommendation system that uses the documents that the users themselves created. Within the document the user requests a song and describes the background for requesting the song. As mentioned above, since activity has great impact on the choice of the music one wants to hear, we believe that by using this document we could recommend song depending on the situation one is in.

C. Characteristic of Korean Radio Broadcasting

What made our research possible was the characteristic of Korea radio broadcasting system. There are three participants in Korea radio channel: the DJ, celebrity guests, and the audience. The DJ and the celebrity guests direct their program and plays music that satisfy the theme for each section. The audience, mostly radio channel listeners, posts their personal stories along with a song request on the radio station bulletin board and these stories act as the pool of music to be selected by the radio DJ. Each story, which is written in Korean by a listener, is associated with a song request and a background for such request. We believe that

 ISBN: 978-1-61208-227-1

the background information contains situational information. Thus, these documents, posted on the Korean radio channel's Internet bulletin board, provide a link between music and the contextual information. Therefore, we aim to use document similarity to find similar music.

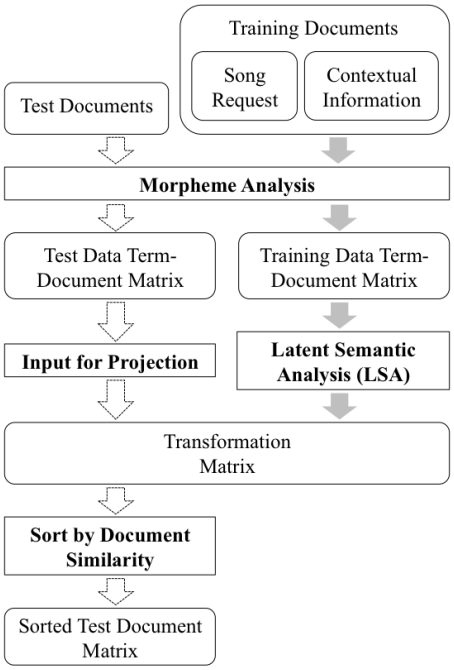


Figure 1. Overall process of our system. The evaluation process is shown altogether. The bold text indicates the steps of the process. The shaded arrow indicates the process of creating a transformation matrix and the unshaded arrow indicates the process of the evaluation.

III. Proposed System

In this section, we describe our system that uses the textual data to extract the contextual information when recommending music. This approach can be expanded broadly as social network services are overwhelming these days and people tend to write about their situational status often. The overall system is shown in Figure 1. Amongst the stories gathered we divide them into test documents and training documents. Then, we perform a morpheme analysis on both data set. Latent Semantic Analysis (LSA) is performed on the training set to create a transformation matrix. Using this transformation matrix, the test documents are projected and ranked for evaluation. In the following sections we will talk about the core algorithm of our system in more detail.

A. Morpheme Analysis

In our system, we use document similarity to generate a recommendation list. In order to find similarity between the documents, we first need a comparable representation of each story. This is accomplished by using the vector model. Each story is represented as a vector where each element represents the occurrence of the words used in the story. However, using all the words has two major problems. One problem is that the complete word set contains stop-words. Stop-words are words such as 'and', 'the', 'at', etc. These words are uninteresting words and the presence of them can degrade the performance of LSA. Another problem is that the complete word set contains stemmed words. For example, 'learn', 'learning', 'learned', and 'learnable' all come from the same stem 'learn' but is regarded as distinct words. This again causes the performance of LSA to go down. To avoid these problems, morpheme analysis is performed prior to vectorizing the documents [5]. The morpheme analysis tool removes the stop-words and also discovers the word stem.

B. Latent Semantic Analysis

Through the morpheme analysis, each document is represented as a vector of word occurrence where the stop-words and stemmed words are removed. However, to use the vector as it is leads to another problem. Compared to the total word pool which is the overall words used in all of the gathered documents, the number of words used in each document is extremely small causing the vector to be sparse. Using this sparse vector for comparison would not be accurate. In the research field of language processing, Latent Semantic Analysis (LSA) has been used as a proper tool when comparing sparse data [6]. LSA processes the sparse matrix to discover the latent meaning of the documents or the words. With the processed word-document matrix containing latent meaning, we are now able to find similar documents by using distance metrics.

An important algorithm used when performing LSA is the singular value decomposition (SVD). After creating a word-document matrix by combining all the story vectors, the matrix is decomposed into a set of rotation and scale matrices. The result of the decomposition is shown in (1).

$$M = USV^T \tag{1}$$

where $M \in \mathbb{R}^{t \times d}$ is the original word-document matrix, $U \in \mathbb{R}^{t \times t}$ is the matrix representing the words, S is a diagonal matrix of size $\mathbb{R}^{t \times d}$ containing singular values, and $V \in \mathbb{R}^{d \times d}$ is the matrix representing the documents. Both U and V are orthogonal.

Once the decomposition is done, the diagonal matrix S and the document relevant orthogonal matrix V^T are multiplied to find the semantic discriminations between the documents. The parameter that can be altered is the number of singular values to use. The number of singular values determines the dimension of the vector space where the reduced document vector will be projected to. The result of the projection can be shown as: $D' = S' \times V'^T$ where $D' \in \mathbb{R}^{k \times d}$ is the reduced approximation matrix, S' is the reduced version of the diagonal matrix using k singular

 ISBN: 978-1-61208-227-1

values, and V' is the reduced document relevant orthogonal matrix. After the projection, a distance metric is used to calculate the distance between the document vectors. In our experiments, we used cosine and Euclidean distance metrics.

C. Music Recommendation based on Document Similarity

By performing LSA, a transformation matrix that would project the test document vector to the same vector space for comparison is created. We used 10,000 singular values for the reduction explained in Section III-B. Using the transformation matrix, the input matrix, which would be a set of test document vectors, is projected to the vector space and the distance between each vector is compared to generate a ranked list. The closer the vector is, the more similar the document will be. Therefore, assuming that people prefer similar music in similar situation, recommending music requested in similar documents would be a feasible recommendation.

IV. Evaluation

In this section, we explain the dataset and the metrics used for evaluating our system. As mentioned in Section III-C, our assumption is that people in similar situation would prefer similar songs. Since our system extracts contextual information from individual stories, if our assumption is correct then the stories that request the same song would be similar. In order to validate our assumption, from the mined documents, we manually marked documents that requested the same song and regarded these as relevant to each other. We then applied conventional precision and recall approach and reciprocal rank to evaluate our system.

A. Dataset

Individual stories are posted in the radio channel's Internet website. We data mined 14,000 documents from the bulletin board of the radio program held between 2:00 pm and 4:00 pm. Amongst the 14,000 documents, 10,000 documents were used to train the transformation matrix. The remaining 4,000 documents were used for evaluation. We first extracted the requested song for all 4,000 test documents. This was a semi-auto process since we had to manually mark the requested song for each document. After marking the data, we ran a program to check which music was requested how many times. Since the title and the musician can be a noise data, the program also deleted them after the counting process was performed. After counting the songs requested by the 4,000 test documents, we collected documents that were linked to the top 10 most frequently requested songs shown in Table (I). There were 291 documents altogether and these documents were used as a test set. From here on, documents requesting the same song will be denoted as relevant documents.

Song ID	Song Title	Number of Documents
1	Happy Birthday to You	41
2	Heartbreaker	40
3	Can't I Love You	36
4	Relief	37
5	There Isn't Anyone Like You	28
6	Will you Marry Me	24
7	Cheer Up	22
8	I Don't Care	22
9	Tears are Bitter	21
10	Love Rain	20

Table I
TOP 10 MOST FREQUENTLY REQUESTED SONGS.

B. Metrics

Taking each document as an input, the remaining 290 documents were ranked based on document similarity. To measure similarity, we used Euclidean distance and cosine distance. We calculated three metrics to evaluate our system; mean average precision at 10 (MAP10), mean average precision at 5 (MAP5), and mean reciprocal rank (MRR). We compared the MAP5, MAP10 and MRR of our algorithm with that obtained when the documents were ranked randomly. From here on, the randomly generated MAP and MRR will be denoted as MAP5r, MAP10r and MRRr respectively.

1) Mean Average Precision: Precision and recall is an evaluation metric that is widely used in information retrieval. For each document, using equations (2) and (3), we calculated the precision until the recall rate reached 1. Then, it was averaged to find the average precision for each document. Once the average precision was calculated, we averaged the average precision for each test song.

$$Precision = \frac{relevant\_docs \cap retrieved\_docs}{retrieved\_docs} \quad (2)$$

$$Recall = \frac{relevant\_docs \cap retrieved\_docs}{relevant\_docs} \quad (3)$$

For each test song, we calculated the precision at 5 and 10. Since each song is associated with several relevant documents that requested the song, we calculated the precision at 5 and 10 for each input document and calculated the average of the mean precision for each input of the relevant document. These results were compared to the Mean Average Precision at 5 and 10 of that generated randomly. The random generation was assumed to have a uniform distribution and the result is show in Figure 2 and Figure 3.

2) Mean Reciprocal Rank: Another conventional metric in Information Science is the Mean Reciprocal Rank (MRR). In order calculate MRR, we first find the rank of the first relevant document for each document. After finding the first appearance of the relevant document for each test document, the average of the inverse of the rank is calculated. This

 ISBN: 978-1-61208-227-1

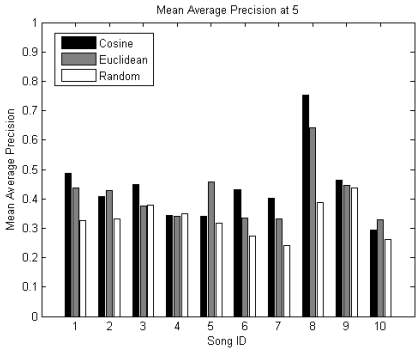


Figure 2. Mean Average Precision at 5 using cosine distance metric, euclidean distance metric, and randomly based metric.

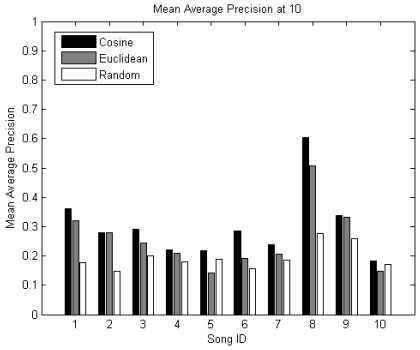


Figure 3. Mean Average Precision at 10 using cosine distance metric, euclidean distance metric, and randomly based metric.

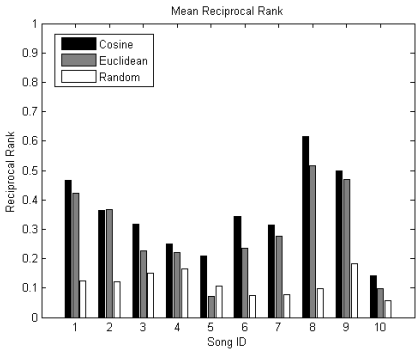


Figure 4. Mean Reciprocal Rank using cosine distance metric, euclidean distance metric, and randomly based metric.

result is again compared with that generated randomly and is shown in Figure 4.

V. Results and Discussion

From the results shown in the previous section, we were able to find that cosine distance metric outperformed the euclidean distance metric. This can be explained by the fact that normalization wasn't performed prior to our evaluation. For example, a document using word 'a' once and 'b' once would be distinct from the document that uses word 'a' twice and word 'b' twice if Euclidean distance metric is applied. However, syntactically these two documents should be regarded as nearly the same. Cosine distance metric considers this fact and thus outperforms the Euclidean distance metric. Also, for most of the test data, the result of our algorithm outperformed the result generated randomly. This indicates that our system was able to find similar documents and those similar documents actually requested the same song. Thus, it is possible to recommend music purely based on textual mining and analysis of blog or specific music related programs. The best performance was shown in song 9. A possible explanation is that the lyric and the melody of the song matches. Song 9 is a ballad song which is quiet and moody. The lyric is about reminiscence of one's past love. The lyric follows the moody melody and thus people who request this song would share a similar situation regarding sad love.

However, as shown in song 10, there were cases where our algorithm didn't perform well. This can be explained by the characteristic of the song. The melody of the song is bright and cheerful. However, the lyric of the song is about waiting for love. Having such characteristic, people whose preference is more dependent on the melody might prefer the song in a cheerful situation while people whose preference is more dependent on the lyric might prefer the song in an moody situation reminiscing love. We believe that such diversity in situation when requesting the song is the reason for the poor result.

An unexpected finding was the relatively poor result for song 6. The lyric and the melody of the song absolutely fits for proposing marriage. Therefore, our expectation was that the music would be requested usually in situations regarding marriage proposal. However, when the documents requesting this song were checked manually, we found out that the song was requested not only in proposing situations, but also in situations when the user was celebrating his/her anniversary. Due to this subtle difference, our system was not able to find similarity between marriage and anniversary, and thus gave a relatively low result. In order to overcome these situations, future work will be discussed in the following section.

Despite some limitations, most of the results outperformed the results obtained when the stories were randomly ranked. Thus, our assumption that people shared similar preference in similar situation proved right and our approach to analyze textual information to gather contextual information showed possibilities.

VI. Conclusion

In this paper, we presented a novel approach to recommending music based on text analysis. Rather than implicitly guessing the contextual information of the users, we showed that it is possible to use documents, written by the users, to extract contextual information explicitly. To this end, we

 ISBN: 978-1-61208-227-1

gathered radio stories written by individuals via the radio channel's bulletin board and performed LSA to identify the semantic meanings of the documents to find similar stories. Our assumption was that if people shared similar situational information, then the music they prefer would be similar. Since each story was associated with a song request, the song linked to the most similar story could be recommended. In order to evaluate the system, we used several metrics to check if similar stories actually requested the same song. The result showed that there was a positive correlation between story similarity and song similarity, and thus it could be possible to recommend music purely based on document analysis. Additionally, to check the quality of the recommendation, we plan to perform an user evaluation test.

One limitation in our experiment was that the documents retrieved were limited to stories written in the Korean radio station's bulletin board. However, the main purpose of this research was to check the validity of the approach in using text mining and analysis to extract the contextual information of the user when recommending music. While not perfect, we showed that in most cases, people requested similar songs in similar situation. Thus, it proved the possibility of performing text analysis when recommending music. To expand our research for worldwide radio listeners remains as a future work.

Along with applying our research to worldwide listeners, we plan to improve the quality of the system. As indicated above, LSA has some limitations. One most crucial limitation is that it is not appropriate for detecting polysemies. Polysemies are words that have multiple meanings. Recent studies have shown that this problem can be tackled by implementing the probabilistic Latent Semantic Analysis (pLSA). Therefore, by performing pLSA to our dataset we expect to achieve a better result.

Another possible improvement can be found in the morpheme analysis tool. The morpheme analysis tool we used, correctly removed stop-words almost completely but was only able to discover the stem words with approximately 80% correctness. This rate went down significantly if the document contained misspelled words, abbreviations, and non-spaced words. We expect that our approach will have better performance if these noise within the stories are handled.

VII. ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0013476).

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, pp. 734–749, 2005.

[2] C. Anderson, *The long tail*, Gramedia Pustaka Utama, 2006.

[3] Y. Hu, Y. Koren, and Volinsky, *Collaborative filtering for implicit feedback datasets*, Proc. of ICDM-08, 8th IEEE ICDM, pp. 263–272, 2008.

[4] International Federation of the Phonographic Industry, *Digital Music Report 2012*, Available at: http://www.ifpi.org/content/library/DMR2012.pdf (Accessed: 10 September 2012).

[5] Kookmin University Korean Language Technology, Available at: http://nlp.kookmin.ac.kr/HAM/kor/index.html (Accessed: 13 October 2012).

[6] T.K. Landauer, P.W. Foltz, and D. Laham, *An introduction to Latent Semantic Analysis*, Discourse Processes, Vol. 25, No. 2–3, pp. 259–284, 1998.

[7] T. Li and M. Ogihara, *Content-based music similarity search and emotion detection*, Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing, Vol. 5, pp. 705–708, 2004.

[8] Q. Li, B.-M. Kim, D.-H. Guan, and D.-W. Oh, *A music recommender based on audio features*, SIGIR 04: Proc. 27th Annual Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 532–533, 2004.

[9] S. Pauws, W. Verhaegh, and M. Vossen, *Fast generation of optimal music playlists using local search*, Proc. 7th Int. Conf. Music Inf. Retrieval, pp. 138–143, 2006.

[10] R. Ragno, C. J. C. Burges, and C. Herley, *Inferring similarity between music objects with application to playlist generation*, Proc. 7th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval, pp. 73–80, 2005.

[11] G. Reynolds, D. Barry, T. Burke, and E. Coyle, *Interacting with large music collections: Towards the use of environmental metadata*, Multimedia and Expo, 2008 IEEE International Conference, pp. 989–992, 2008.

[12] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, *Methods and metrics for cold-start recommendations*, Proc. of the 25th SIGIR, pp. 253–260, ACM Press, 2002.

[13] B. Schwartz, *The paradox of choice: Why more is less*, Harper Perennial, 2005.

[14] J.H. Su, H.H. Yeh, P.S. Yu, and V.S. Tseng, *Music recommendation using content and context information mining*, Intelligent Systems, IEEE, Vol. 25, No. 1, pp. 16–26, 2010.

[15] J. Wang, A. P. de Vries, and M. J. T. Reinders, *Using user-based and item-based collaborative filtering approaches by similarity fusion*, Proc. of the 29th SIGIR, 2006.

[16] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, *Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences*, Proc. ISMIR, 2006.

[17] M. Zhang and N. Hurley, *Avoiding monotony: Improving the diversity of recommendation lists*, Proc. of the ACM RecSys, pp. 723-732, ACM, 2008.

 ISBN: 978-1-61208-227-1

Automated Reference Model Generation and Utilization for Dimensional Control of Large Scale Assemblies and Assembly Processes

Teuvo Heimonen
Optical Measurement Laboratory
Kemi-Tornio University of Applied Sciences
Kemi, Finland
Teuvo.Heimonen@tokem.fi

Markku Manninen
A. M. S. Accuracy Management Services
Oulu, Finland
Markku.Manninen@ams-oulu.com

***Abstract*—A lot of manual user interaction with computer aided design and analysis software is currently needed for dimensional control of large scale assemblies. This user input is required for both selecting the vital entities to be measured and analyzing the results of the measurements. In this paper an automated approach for reference model generation and vital entity selection is presented. The reference model generation is guided by user editable knowledge base and it is based on data extraction from computer aided design data. Prospects to utilize the generated reference model for automated manufacturing accuracy analysis are also outlined. The software implementation made has proved to be feasible for different kinds of dimensional control needs of large scale assemblies, and outweighs the current state of the art both in speed, completeness, and versatility.**

Keywords-accuracy control; large scale assemblies; data extraction; dimensional inspection

I. INTRODUCTION

A modular construction is today a typical way to build large scale objects like ships, submarines, offshore platforms, airplanes, and bridges. In this manufacturing approach, the final product is assembled by joining separate, large subassemblies (blocks) one after another.

One of the main interests in the assembly process of large scale objects is how different blocks can be fit together optimally. If the two blocks to be joined do not fit properly, corrective reworking is needed, causing extra costs and time delays [3]. In order to expedite the joining process, the geometry of the blocks is measured and verified (as-built to as-designed) before the joining.

It is common in modern industry that the dimensions and the shape of the objects to be manufactured are designed computer aided. Thus geometrical design data of parts and products is almost always available in the form of computer aided design (CAD) files. These files can be used directly as sources of geometric information and thus the idea to use CAD data as a reference for as-built to as-designed verifications has been brought forth, see e.g., [1]. However, the CAD data typically contain a lot of unnecessary information from the dimensional accuracy control viewpoint. Thus the extraction of the essential data out of CAD data is needed. Currently, this vital entity selection and extraction is performed by using different kinds of interactive approaches requiring a lot of manual work, caution, and time.

In the case of large scale objects, the implementation of an automated vital entity (e.g., point to be measured) selection seems to remain an extremely challenging and thus an unsolved task. Reasons for this comprise

- The differences in the CAD data and systems: an approach designed for a certain CAD data format or CAD system is typically not feasible with another
- One of a kind manufacturing -property of large scale objects: vital entities are neither the same nor in the same place on the object
- The rules, principles, and practices to choose the vital entities are different in different applications and in different organizations: the expertise to select vital entities is undocumented tacit knowledge of the personnel.

There are few papers addressing automatic vital entity extraction and the generation of a reference model for as-built to as-designed verifications of large scale objects (readers interested in computer aided inspection planning for smaller sized objects are referred to [10]). Manninen et al. [4] seems to be the first and only paper in which a workable implementation (for shipbuilding) was presented. However, even though the basic ideas and principles presented in [4] have been proven to be feasible, it has been noted that the implementation is not flexible enough for different kinds of (complicated) blocks, different kinds of practices for choosing the vital entities, and different kind of measuring and as-built to as-designed analysis needs in different shipyards. The main drawback of the implementation presented in [4] is that it analyzes only planar surfaces of CAD data leading both instability and incomplete reference models in case of CAD models comprising curved surfaces. The implementation is also slow, restricted to only one CAD format (dxf) for input, and comprises no support for automatic manufacturing analysis.

In this paper, a knowledge base guided approach to automatically generate reference models for dimensional manufacturing accuracy analysis of large scale objects is proposed. This reference model generation comprises the reduction of CAD data and selection of vital entities essential and sufficient for dimensional control purposes. The automatic reference model creation and vital entity selection processes are controlled through a set of parameters. The identification information and values of these parameters

 ISBN: 978-1-61208-227-1

form an information source, which is called a knowledge base in this paper.

Some prospects to utilize the generated reference model for manufacturing accuracy analysis are also presented. The main new idea in this sense is the automated manufacturing accuracy analysis. It comprises computation of quality figures for a certain set of user defined structure types and combining these quality figures to form a comprehensive quality database (quality figure tree) of the object manufactured. In addition to be used in analyzing the dimensional accuracy of the object manufactured, this quality data is intended to be utilized for monitoring the manufacturing and assembly processes.

The rest of this paper is structured as follows: In Section II, a new approach is presented at the general level. Some implementation issues are commented on as well. Then, in Section III, some results of our experiments with shipbuilding data are presented, and finally, in Section IV, we conclude by summarizing the results and presenting ideas for further study.

II. DESCRIPTION OF THE APPROACH

In this section, we present our approach at the general level with some examples. First, the general flow of the reference model creation is given, and then, in subsections II.A to II.C, the utilization of the knowledge base, the recognition of vital structures, and the automated manufacturing accuracy analysis are presented.

The creation of the reference model and vital entity selection is purely based on geometric information. The process progresses through the following steps (see also Fig. 1): First, the file containing the CAD data is read into the system and all geometric data is converted to a boundary representation (see e.g., [5, 8]). We utilize Open Cascade Technology [6] for this step, since it provides ready-made interfaces for this operation. Then, each face of the boundary representation is analyzed to find out, which side (nominal side, non-nominal side, thickness side, see Fig. 2) of a structure (e.g., steel plate) the face represents and classified on the basis of this analysis for further processing. After all faces have been "side-classified", the nominal side faces are studied once again in order to find out whether bigger structures can be constructed by stitching a set of faces together. The results of the stitching phase are either planar or curved surfaces. These surfaces are then checked against size limits for acceptable objects, and either accepted or excluded. If needed, the application specific recognition of vital structures can be performed in this phase. Finally, a reference model of "intelligent objects" (see [4]) is created and finalized by dividing the points of the model into vital (to be measured) and non-vital points according to the instructions given in the knowledge base controlling the reference model generation.

A. Utilization of knowledge base

The end-user is able to guide the automatic reference model creation and vital entity selection processes through a set of parameters. The identification information and values of these parameters form an information source, which is called a knowledge base.

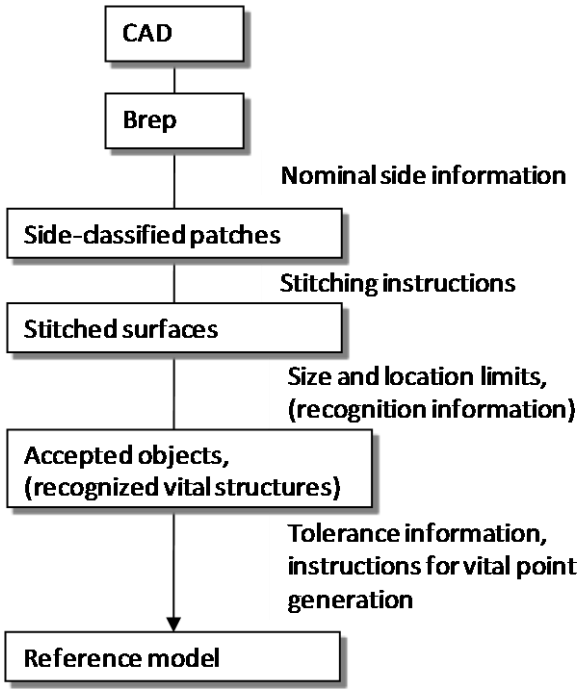


Figure 1. Reference model generation process.

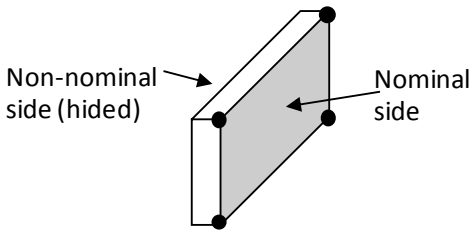


Figure 2. Nominal and non-nominal sides of a steel plate. Besides the nominal and non-nominal sides there are four thickness sides in this example. Vital points (points to be measured), shown here with black dots, should be generated on the nominal side of the plate.

The structure of the current knowledge base has been developed together with several dimensional control and shipbuilding experts. Currently, it comprises two main sections: a general section and an application specific section. The end-user can edit the values of the parameters in the general section but otherwise (identifications, the structure of the knowledge base) the general section of the knowledge base is meant to be edited only by the software engineer. The general section of the knowledge base currently comprises

- Tolerance information to be assigned to the various objects in the reference model
- Information for selecting correct nominal side for each structure to be included into the reference model

- Size and location limit information to control whether a CAD structure is to be accepted to or excluded from the reference model (see Fig. 3 for example of the effect)
- Information where to generate vital points (points to be measured) in some special cases (see Fig. 4 for example of the effect)
- Information whether some of special details, like holes or curved structures are to be included into the reference model (see Fig. 3 and Fig. 4 for examples of the effect)
- Information for automatic alignment of the measurement data to the reference model

The application specific section of the knowledge base is used to define the vital structures to be recognized and to be used in manufacturing accuracy analysis. This section is completely user editable even though some keywords need to be used in order to guarantee the knowledge transfer between the knowledge base and the software. This section includes information for

- Recognition of the vital structures
- Vital structure -wise selection of the quality figures to be computed
- Combining the lower level quality figures to upper level quality figures
- Constructing the quality figure tree of the quality figures computed (see Fig. 6 for an example).

B. *Recognition of vital structures*

Vital structures are different for different large scale assembly applications. Thus application specific knowledge to recognize the vital structures from CAD data is needed. Currently, this information is given in an interchangeable recognition data module of the knowledge base. Thus different applications can be handled easily by changing the contents of this module to be suitable for the application.

In the recognition data module identification (name), recognition, recognition process and quality figure type information is given for each structure wanted to be recognized and further used in the manufacturing accuracy analysis. The recognition information is simple geometric data to classify a structure, and could currently comprise size, location, and orientation limits. The recognition process information is used to control the classification process. It informs the system in which order the recognition information should be applied when trying to classify an input structure. Quality figure type information is a list of quality figures to be computed for the structure after the actual (measured) data are available. A weight, which is used when quality figures are combined in manufacturing accuracy analysis, is given for each quality figure as well.

We have demonstrated the recognition of vital structures in shipbuilding applications. In our demonstrations structures like whole block, decks, bulkheads, stiffeners, block faces, engine foundations, basic plane, centre plane, and special points were recognized. The quality figures computed for these structures comprised flatness, straightness, location, orientation, length, width, height, and cross-measures.

C. *Automated manufacturing accuracy analysis*

In order to pack the information offered by numerous quality figures computed for each vital structure, these quality figures are statistically combined to form a hierarchical, tree-like presentation of the manufacturing quality of the block. This manufacturing quality tree is saved as an xml-file and can thus be browsed afterwards to check any detail of the manufacturing quality (see Fig. 6).

The vital structure types for which the quality figures are computed can be selected by using the knowledge base. For each structure type, which is identified to be recognized, a set of quality figures to be computed can be selected (from a list of available quality figures).

There are two types of quality figures in our approach. The basic quality figures are the lowest level quality figures obtained by comparing a geometric feature (either directly measured or computed by using measured data) to the corresponding design value for the feature in question. Thus the basic quality figures are direct measures of dimensional accuracy of the feature in question. The combined quality figures are obtained by combining two or more quality figures to a one (higher level) quality figure.

Each quality figure has at least value, weight, and location information assigned to it. Combined quality figures have also statistical data available. Usually, the unit of the quality figure is also given but if the quality figure is a combination of values with different units, the unit is not defined.

The value of a basic quality figure is a deviation of the actual data from the design or desired data. The value may, in some cases, also be negative. The closer the value is to zero the better.

A weight is used when combining quality figures in order to emphasise the significance of some measured data or quality figures more than others. As mentioned above, weights can be given quality figure -wise for each structure to be recognized in the application specific section of the knowledge base. The default weight is one for each quality figure meaning that each quality figure will have equal importance.

Two types of location information is presented: the location point, which is typically a centre point of the input data of the quality figure, and the bounding box, which limits the size of the space from where the input data was obtained. Currently, the main purpose of the location information is to identify to which structure the quality figure in question is related to but later the location information can be used to put the quality information data in the correct place in a graphical presentation.

For combined quality figures, several statistical figures (maximum, minimum, weighted arithmetic mean, standard deviation, and mean deviation) are computed and saved. As a value (result) of a combined quality figure either the root mean square error (RMSE) or weighted arithmetic mean is used. The RMSE, which is related to the arithmetic mean (m_x) and the standard deviation (s) of the sample data (x_i)

$$\mathrm{RMSE}^2 = m_x^2 + s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 \quad , (1)$$

where *n* is the number of samples, is used when the basic quality figures are combined. Thus, when the RMSE is used, the value of the combined quality figure includes information of both average error (bias) and deviation of the data. It should be noted that, in our approach, the sample data (x_i) to be used in (1) is derived from actual to design comparisons, and thus, RMSE (not RMS) is indeed obtained.

The weighted arithmetic mean is used, when already combined quality figures are further refined to a higher level quality figures. An example of this is an integration of similar quality figures of different structures of the same type (e.g., flatness quality figures of all block faces of the block). This kind of integration is continued until a quality figure tree, from simple point deviations up to the overall quality of the whole block, has been constructed.

III. EXPERIMENTS

The approach presented has been experimented in several shipbuilding cases. Separate knowledge base instances have been built for several different kinds of block types from two different shipyards. The CAD data provided by the design department of the shipyard have automatically been reduced to form a reference model by using our software modules, and obtained results have been evaluated together with the personnel responsible for the dimensional control in the shipyard. Then the actual blocks have been measured by the measurement group, and finally, the measured data have been aligned to the reference model and quality figures computed. In the following sub-chapters some detailed observations of the experiments are presented. The overall results of the feasibility of the proposed approach are summarized in Section IV.

A. *Usability of different CAD file formats*

As a CAD data format both dxf [2], iges [9], and step [7] formats were evaluated and each format was found out to be equally suitable. No significant difference was observed either in the generated reference model or the vital entities selected to be measured.

B. *Validity of reference models for dimensional control*

The reference models created automatically were suitable for dimensional control purposes. The amount of detail could be controlled by the values of the parameters in the knowledge base, and thus the data extraction process could be tuned to be suitable for the needs of the inspection task in question (see Fig. 3, Fig. 4, and Fig. 5). Some minor problems were detected in the reconstruction of the curved surfaces (e.g., shell plates of the ships) but the deficiencies, which these problems lead to, were more esthetical than practically meaningful from the dimensional control point of view.

The validity of the vital points, i.e. percentage of the vital points that are actually measured, and percentage of the points that are measured but not classified as vital in the reference model, will be studied later in detail when the approach has been adopted into actual use properly. Preliminary studies indicate about 90% result for both cases, which has been judged to be acceptable at this point of development work.

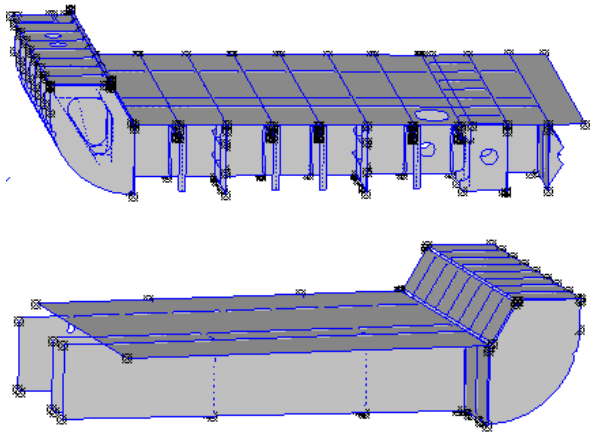


Figure 3. Examples of a knowledge based guided reference model generation: Two halves of the same block imported by using different instructions given in knowledge base. The upper reference model has more details and vital points included (also holes) than the lower model.

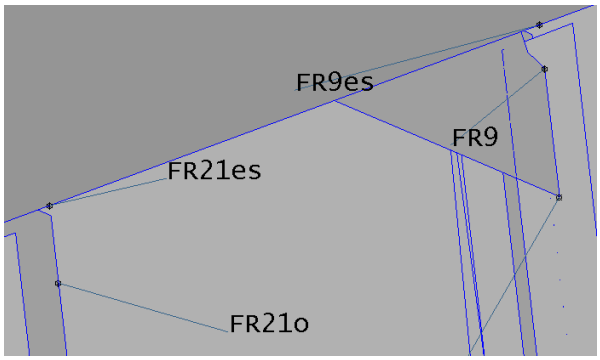


Figure 4. Example of a knowledge based guided vital point generation: "If a bulkhead or a stiffener has a notch (corner cutout), use notch point (FR9), otherwise make an offset point 100 mm from the corner of the bulkhead or the stiffener (FR21o). Create points also on the deck edge in the position where the bulkhead or stiffener goes under the deck (FR9es and FR21es)".

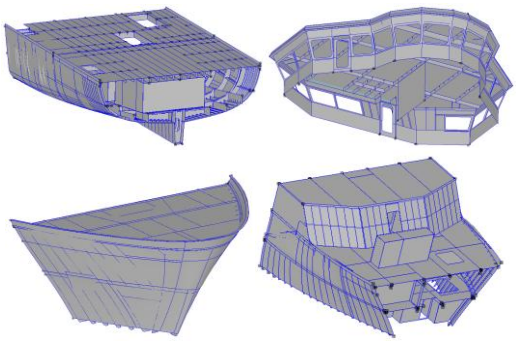


Figure 5. Four examples of generated reference models.

C. Import time comparison

The time to create a reference model depends on the amount of data in the CAD file, the complexity of the 3d structure, and the amount of analysis the knowledge base instructs to be done during the reference model generation process. The reference model generation time with the new approach was compared to the current state-of-the-art approach [4] by creating reference models by using both approaches on the same normal laptop computer. Our approach proved to be significantly faster (Table I).

TABLE I. REFERENCE MODEL GENERATION TIME COMPARISON

CAD-file size [MB]	Generation time [s]	
	Reference [4]	Our approach
2,4	55	4
3,3	65	6
3,7	60	5
4,4	85	6
7,4	240	12
11,3	600	18
21,9	1625	35

D. Data reduction

One purpose of the reference model generation is to reduce the data available in the CAD models. The reference model should comprise all data and structures needed for performing the measurements and visualizing the model judiciously. However, unnecessary data and structures should be excluded in order to offer illustrative and clear model.

All the evaluated models were approved by the measurement personnel and dimensional control management by inspecting the models visually. The amount of data reduced was computed by comparing the number of 3d objects defined in the CAD model and in the generated reference model. The amount of data was reduced typically to less than 2% of the original (Table II).

TABLE II. DATA REDUCTION

CAD-file size [MB]	Number of 3d objects		Percentage
	CAD	REF	REF / CAD
2,4	8515	141	1,66 %
3,3	10875	93	0,86 %
3,7	10500	51	0,49 %
4,4	13687	212	1,55 %
7,4	25431	501	1,97 %
11,3	37890	454	1,20 %
21.9	71707	557	0.78 %

E. Quality figure computation

Quality figure computation was experimented by defining several different structures into the application specific module of the knowledge base. The structures, which were used in all our experiments, comprised block as whole, block face, deck, bulkhead, stiffener, and special points. In some specific experiments some additional structures like foundations, center plane, and basic plane were evaluated. The size, location and direction information was given for the recognition of the structures.

Instead of finding out the actual quality of the blocks measured, the purpose of the quality figure computation experiments was to evaluate the feasibility of the approach. From this point of view, the results obtained from our preliminary experiments were encouraging: The structures were correctly recognized, correct points were used for the computations of different quality figures, and lower level quality figures were successfully combined to the upper level quality figures of the quality figure tree. An example of an automatically generated quality figure tree is shown in Fig. 6.

IV. CONCLUSIONS AND FUTURE WORK

The automated, knowledge base guided approach for reference model generation proposed in this paper seems to offer a step forward for the dimensional control and accuracy analysis of large scale objects. The following advances have been accomplished:

- Possibility to use different CAD formats
- Possibility to import more complicated structures than with the current state-of-the-art approach
- Significantly faster reference model generation than with the current state-of-the-art approach
- Significant data amount reduction (compared to the amount of original CAD data)
- Recognition of user defined structures vital for dimensional accuracy analysis
- Creation of comprehensive quality database advantageous to dimensional accuracy analysis of the object in question and for monitoring the assembly process
- Flexibility and better suitability for different kinds of dimensional control needs by using user editable knowledge base.

The usability of the quality data obtained from the proposed automated manufacturing analysis has not been properly evaluated yet. In order to utilize the data e.g., for monitoring the assembly process and long-term statistical process control further studies are needed. For example, it has to be studied, which statistical quality figures need to be computed and how these figures should be combined in order to properly serve the process monitoring. The current implementation serves as a good starting point for these studies.

In this paper, the studies focusing on the automation of the reference model generation and manufacturing accuracy analysis were reported. Even though very promising results were obtained, it should be noted that some amount of user

 ISBN: 978-1-61208-227-1

interaction is and will be needed. How this interaction (e.g., for editing the knowledge base) should be arranged, must also be considered properly in the future in order to obtain a useful system for the industry of large scale assemblies.

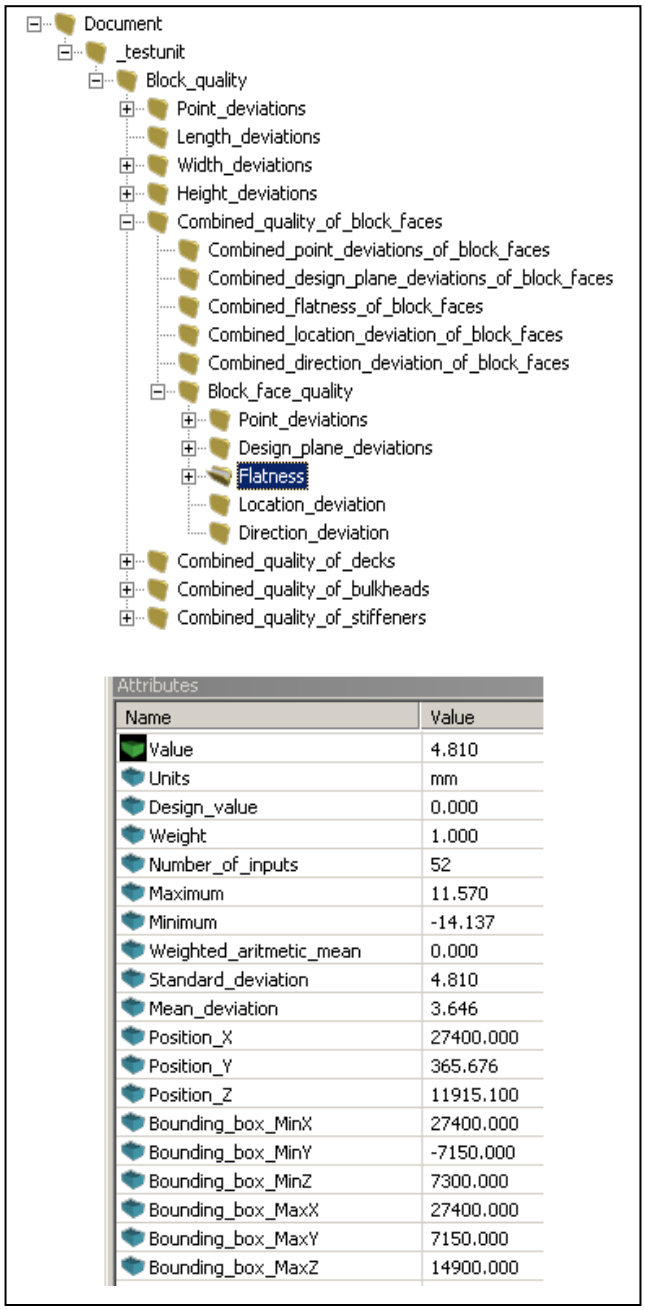


Figure 6. Example of an automatically generated quality figure tree.

ACKNOWLEDGMENT

STX Europe Rauma Shipyard, IHC Merwerde Kinderdijk Shipyard, and measurement service company Prismarit Ltd. are gratefully thanked for the co-operation. Tekes, the Finnish Funding Agency for Technology and Innovation, is acknowledged for the financial support.

REFERENCES

[1] H. Ailisto, “CAD model-based planning and vision guidance for optical 3D co-ordinate measurement”, Technical Research Centre of Finland, VTT Publications 298, 1997. 70 p. + app. 63 p.

[2] Autodesk Inc., Autocad 2012, DXF Reference, February 2011. 262 p. Available from: http://images.autodesk.com/adsk/files/autocad_pdf_dxf-reference_enu.pdf. [7 June 2012].

[3] M. Manninen and J. Jaatinen, ”Productive Method and System to Control Dimencional Uncertainties at Final Assembly Stages in Ship Production”, Journal of Ship Production, vol. 8, no. 4, 1992, pp. 244 - 249.

[4] M. Manninen, J. Linna, and K. Jacobsen, ”Object Oriented Software for CAD Based Dimensional Analysis and Alignment Control of Steel Structures in Hull Assembly”, 12th International Conference on Computer Applications in Shipbuilding. Busan, Korea, 23.-25. August 2005.

[5] M. Mäntylä, An Introduction to Solid Modeling. Computer Science Press, College Park, MD, 1988.

[6] Open Cascade S.A.S., Open CASCADE Technology, 3D modeling & numerical simulation. Available from: http://www.opencascade.org/. [7 June 2012].

[7] SCRA, Step Application Handbook, ISO 10303, Version 3. 175 p. Available from: http://www.uspro.org/documents/STEP_application_hdbk_63006_BF.pdf 2006. [7 June 2012].

[8] I. Stroud, Boundary Representation Modelling Techniques. Springer, 2006.

[9] U.S. Product Data Association, Initial Graphics Exchange Specification IGES 5.3. 621 p. Available from: http://www.uspro.org/documents/IGES5-3_forDownload.pdf. [7 June 2012].

[10] F. Zhao, X. Xu, and S. Q. Xie. ”Survey paper: Computer-Aided Inspection Planning-The state of the art”,. Computers in Industry, vol. 60, issue 7, 2009, pp. 453-466.

 ISBN: 978-1-61208-227-1

ListCreator: Entity Ranking on the Web

Alexandros Komninos
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi 67100, Greece
alexkomn@ee.duth.gr

Avi Arampatzis
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi 67100, Greece
avi@ee.duth.gr

Abstract— In this paper, we present a web application for entity ranking. The application accepts as input a query in natural language and outputs a list of the most relevant entities according to the query. The system uses web documents as data and performs extraction, formatting and ranking of entities in real time. An experiment is conducted to determine the most efficient ranking method among six alternatives. The experiment suggests that the total frequency of an entity in a retrieved set of documents has less to say on the entity's relevance than the number of retrieved documents it occurs in. Furthermore, for small retrieved sets such as the top-10, document rank information seems to play a little role.

Keywords-web entity ranking; entity search; information retrieval

I. INTRODUCTION

Search engines answer user queries by returning ordered lists of documents. In many occasions though, users are not searching for documents but for some more specific information. This information is often *named entities*. The term named entity is used for anything that has a distinct existence and can be characterized by a name, so it can refer to people, companies, products, etc. The need for retrieving named entities as query answers has led to research for systems that can recognize and return this type of information instead of whole documents.

ListCreator [1] is a web application that can answer user queries for entities of three categories: persons, locations and organizations. The application uses as data web documents that match to the submitted query. The ranking of the entities found in these documents is achieved by statistical information retrieval methods, taking advantage of the common information among them. The results are returned to the user as a ranked list of all the relevant entities that the application managed to extract.

The contribution of this paper is twofold. First, we build an online prototype as proof-of-concept for entity ranking using information retrieval methods. Such methods are simple and fast, and therefore suited for an online application. They are also less-limited than ontology-based methods since web documents are used as data. Second, we evaluate several entity ranking methods based on several combinations of statistical quantities corresponding to different hypotheses on language use of document authors and search engine document ranks.

The rest of this paper is organized as follows. In Section II, we review related work. In Section III, we give a detailed description of ListCreator's methods and architecture. In Section IV, we perform a small experiment comparing different methods for ranking entities. Conclusions are drawn in Section V together with directions for further research and improvements.

II. RELATED WORK

Entity ranking has a lot in common with automatic question answering, since the answer to a question is often a named entity or in some cases a list of named entities. An approach that led to good results is using many different text snippets that are expected to contain the desired answer, and using the common information among them to accurately locate it [2]. INEX (INitiative for the Evaluation of XML retrieval) started in 2007 an entity ranking track which was run until 2009. The purpose of this track was the creation of entity ranking systems that could rank relevant entities that had a Wikipedia page. A common approach among many teams was to find a relevant document for each candidate entity and then rank these entities according to the relevance of the document to the query, using document retrieval methods [3][4]. TREC (Text REtrieval Conference) run from 2009 to 2011 a track for related entity finding in the web. The purpose was finding relevant entities to a query that engage in a given relationship with a source entity. The relevance of the candidate entities was determined by many participants by the co-occurrence with the source entity in web documents [5][6].

A different approach is using information extraction techniques to construct structured date from text by extracting facts about entities [7][8]. This requires natural language processing, for example a syntactic parser, and is achieved using machine learning methods. Since applying machine learning to large volumes of text has great computational cost, the above systems constructed a database of relations between entities offline. The database is then queried for relevant entities by the user at runtime. An alternative is using data sets of existing ontologies constructed either manually or automatically using information extraction [9][10].

 ISBN: 978-1-61208-227-1

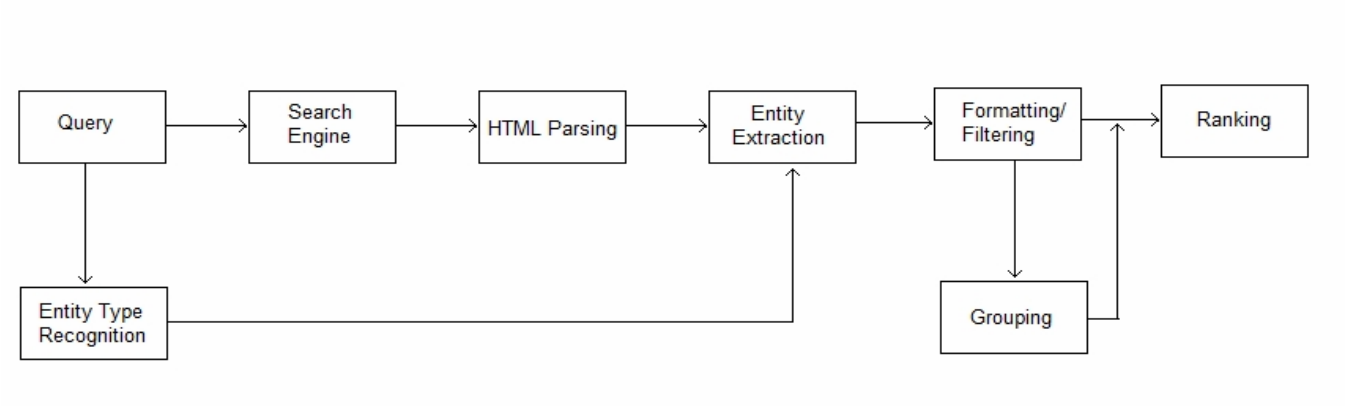


Figure 1 . The System's Components and Dataflow

Our approach is based on information retrieval methods leveraging the large data volume of the web. The difference from INEX and TREC approaches is that a descriptive document for the entities, like Wikipedia pages or personal homepages, or a source entity is not required. The information retrieval methods use statistical measures based on a bag of words model. These methods cannot identify complex relations in sentences like the methods using machine learning, but can process large amounts of text very efficiently and are proven effective by traditional search engines. The advantage over methods using machine learning is that the data can be processed in real time so the results are not limited by the relations recorded in a database. The question answering systems that use common information between different documents are closer to our approach, but they only use term frequency as a measure since their goal is not ranking but verifying the correctness of results produced by an extraction process. We evaluated several methods for ranking, and the results suggest, in contrast to question answering systems, that term frequency is not a strong indication of an entity's relevance, as we will see in Section IV.

III. SYSTEM DESCRIPTION

The system's architecture is depicted in Figure 1. The components for formatting, filtering, grouping and ranking of entities are all coded in JAVA [11]. The user web interface is coded in HTML [12], JavaScript [13], and PHP [14].

A. The Application Website

The central webpage consists of an input form for the user's query and gives the option to determine the type of entity (person, location, organization) that she is searching for. The default option is "auto" which corresponds to automatic recognition of the entity type.

The automatic recognition feature uses a list of about 100 keywords for the location type and about 50 keywords for the organization type. The collection of keywords is based on WordNet categories [15]. The system checks for the appearance of any of those keywords in the submitted query and if they exist it is assumes the user is searching for the corresponding entity type. If none of the keywords appear the system assumes that the user is searching for persons.

The submission of a query calls the main application and the output is presented in the results webpage with the use of PHP. Each result is linked to a corresponding Wikipedia page (if it exists) so that the user can get more information. The results webpage also gives as references links to the web documents that the entities were extracted from. A results page is presented in Figure 2.

B. The Search Engine

The search engine is a very important component of the system since it provides all the data in the form of documents for extracting and ranking the entities. The application essentially functions as a front-end in a search engine. In the current version the search engine used is the Yahoo! BOSS API [16]. Google and Bing were also tested with similar results but Yahoo was chosen because it combines good results with an easy to use API.

The user's query is sent to Yahoo! API without being changed and the results are returned in JSON (JavaScript Object Notation) format. The system asks for only the top-N results. Through some testing we empirically determined that N=10 retrieves enough information while, at the same time, keeps the computational cost low enough for a real time application.

C. Entity Extraction

In this stage, the system recognizes the entities in the documents and determines their type. For this purpose the Stanford NER (Named Entity Recognizer) is used [17]. Stanford NER is a system for entity extraction from text coded in JAVA and distributed with GNU general public license [18] for research and education purposes. The entity recognition is done with a classifier, an algorithm that assigns words in specific categories. The categories supported by the classifier are person, location and organization.

Classification is a supervised machine learning technique. The algorithm uses hand-annotated text to construct statistical rules that can find and determine the category of names in documents. The Stanford NER classifier [19] is based on the statistical model CRF

 ISBN: 978-1-61208-227-1

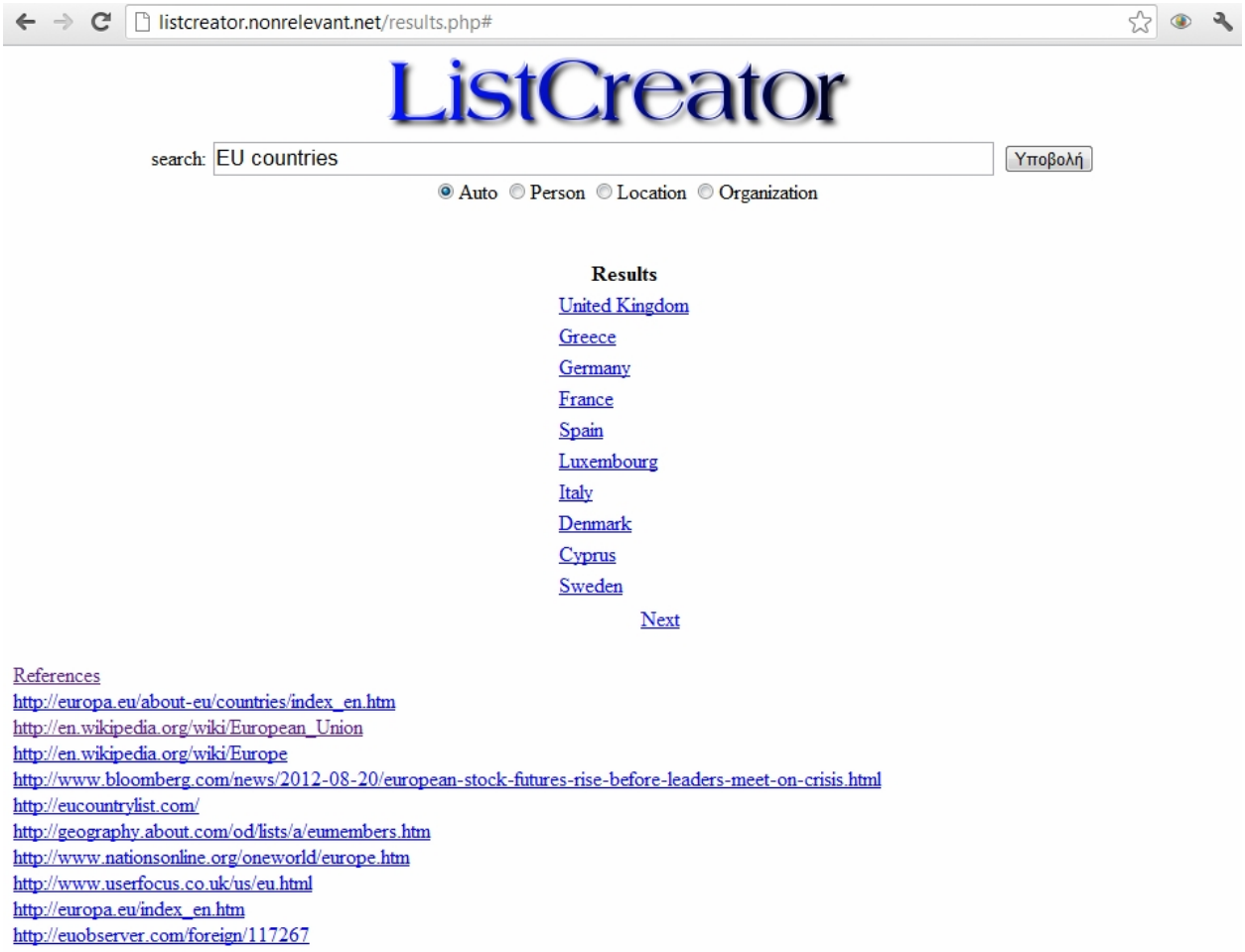


Figure 2. A results Page of the Application

(Conditional Random Field) [20] and comes trained on American and British news articles. The classification process offers some very useful filtering of the entities. The usage of a NER system was considered more suitable for unknown data since it identifies entities by their context in documents, in contrast with a dictionary based approach. It is limited though in the three general entity categories.

In order to extract entities from a web document, the HTML tags have to be removed. For the HTML parsing the JSOUP HTML Parser is used [21]. JSOUP is an open source parser also coded in JAVA that can handle html code with errors

D. Formatting and Filtering

Each entity can appear in a document in many different ways. A person's name for example can first appear with its full name and later be referred with just the last name. In order to achieve a cleaner better ranking in the next stage, the system must recognize which names correspond to the same entity, a task called *co-reference resolution*, and then assign to all of them the same canonical name. The results of this stage are also important for the final presentation since names should appear with all details and avoid listing the same names more than once. The processing of names comes in two steps. In the first step, each entry is formatted and in the second step the names referring to the same entity are grouped taking in consideration the whole set of extracted names.

The basic processing of the first step is converting the names to proper case, i.e. converting the first letter in uppercase and the rest in lowercase. For organization names with less than four letters, all of them are converted to uppercase. Furthermore, the candidate entities are filtered using an exception list. The exception list consists of about 20 entries that correspond to certain names that are often misclassified by Stanford NER. These names are popular websites (Wikipedia, Facebook, Twitter, etc.) that are classified as locations and some acronyms like FAQ, ISBN that are classified as organization. Using this exception list the results from the extraction stage are improved. Another exception list used contains all the country names. This list is checked for search of location type entities because country names appear in large numbers in documents about locations and they can have negative influence on ranking. This exception list is not used when the user is searching for country names.

The grouping of entities that happens in the second step is rule-based and is achieved by comparing each entry with all others. The system checks if an entry forms part of another in word level, and then it is substituted by its complete name. For example, the entries John Kennedy, Kennedy, John F. Kennedy and John Fitzgerald Kennedy are all grouped and substituted by the last form. In order to avoid

 ISBN: 978-1-61208-227-1

grouping into names that may be misspelled, or into a concatenation of two names, the substitution takes place when an entry appears more than once. The grouping step is not applied for queries asking for names of countries, cities and organizations. Country and city names usually do not appear in different forms, while organization names have lots of variance to be grouped with simple rules that often lead to errors.

The above method of grouping gives good results and greatly improves performance, but in some cases the correct grouping of entries cannot be determined. Such is the case of two different candidate entities with the same last name and an entry containing this last name alone. A possible improvement could be the usage of a system that accomplishes co-reference resolution utilizing machine learning but such an approach would increase computational cost.

E. Entity Ranking

The ranking algorithm makes usage of statistical methods of information retrieval. The input in this stage is 10 lists of candidate entities, each one corresponding to the names extracted from each document the search engine provides. The entities are then ranked according to the formula:

$$score = \sum_{j=1}^{df} (N + 1 - r_j)$$

where j is the document an entity appears in, df is the number of the top-N documents that mention an entity, N is the total number of retrieved documents and in the current version is always equal to 10, r is the rank of the retrieved document according to the search engine and has a value from 1 to 10. The formula is based on the preferential voting method *Borda Count*. According to the formula, an entity that appears only in the first document gets 10 points, if it appears on the first and second document, it will get 10 plus 9 points, etc. Entities with higher score are considered more relevant to the query. This ranking formula was chosen after the small experiment that will be described in the next section.

IV. EXPERIMENT

The proposed ranking method tries to solve a problem that resembles the reverse procedure of finding relevant documents to a query. Instead of searching for documents relevant to some terms, it utilizes a small collection of documents (10 in our case) with a common subject and searches for terms (in this case named entities) that are important for this collection. The quantities that were considered useful for the ranking according to the above line of thinking are:

- The total number of occurrences of each entity in the collection of documents (*f*). The higher the frequency of an entity, the more confidence we have in its correctness and importance.
- Document frequency (*df*), which corresponds to the number of distinct documents where each entity occurs. This quantity shows the common information between documents. Assuming that all documents are equally relevant to the submitted query, the names that occur in most documents would also be the most relevant.
- The rank of documents that an entity appears in, according to the search engine (*r*). Taking into account this quantity the documents are no longer treated as equally relevant.

In order to find which of these quantities or which combination of them is more accurate for ranking entities, the following six ranking formulae were compared in the experiment:

$$score = \log(df) \quad (1)$$

$$score = \log(f) \times \log(df) \quad (2)$$

$$score = f \times \log(df) \quad (3)$$

$$score = \sum_{j=1}^{df} (N + 1 - r_j) \quad (4)$$

$$score = \sum_{j=1}^{df} \log(1 + f_j)(N + 1 - r_j) \quad (5)$$

$$score = \sum_{j=1}^{df} f_j (N + 1 - r_j) \quad (6)$$

In all formulae above, j is the document, N is equal to 10, f_j is the number of occurrences of an entity in document j.

There are two opposite hypothesis regarding the frequency of a term and the importance that has for a document [22]. According to the *verbosity hypothesis*, multiple occurrences of a term are not really important because the document is more verbose: the author just used more words to express the same meaning. According to the *scope hypothesis* though, a document's author uses a specific term more times because she has more information to share on this subject.

Equations (1), (2) and (3) do not take into account the ranking of documents, while equations (4), (5) and (6) do. The other difference between the above equations is the weight given to the term frequency of each entity. Equations (1) and (4) are based on the verbosity hypothesis, while (3) and (6) are based on the scope hypothesis. In equations (2) and (5) the logarithm of the term frequency is used. The logarithm in these equations acts as a dampening factor so that the equations represent a middle ground between the two hypotheses.

The evaluation of information retrieval systems is done with some specific measures. For evaluating the performance of the various ranking formulae the measures Precision-at-10 (P@10) and R-Precision were used. P@10 shows the number of relevant answers within the top-10 results. While it does not take into account the ranking of the correct answers, it offers an easy interpretation of results and does not require knowledge of the total of correct answers (recall)

 ISBN: 978-1-61208-227-1

TABLE I. P@10 AND R-PRECISION MEASURES FOR THE SIX RANKING EQUATIONS AVERAGED OVER THE 30 EVALUATION QUERIES.

Ranking Equations	P@10	R-Precision
$\log(df)$	0.4733	0.4209
$\log(f_{tot}) \times \log(df)$	0.4633	0.4306
$f_{tot} \times \log(df)$	0.4433	0.4294
$\sum_{j=1}^{df}(N+1-r_j)$	**0.49**	0.4216
$\sum_{j=1}^{df}\log(1+f_j)(N+1-r_j)$	0.4767	**0.4463**
$\sum_{j=1}^{df} f_j(N+1-r_j)$	0.41	0.4024

to be computed. Furthermore, the p@10 measure is suitable for web retrieval since most users usually check only the top-10 results. A problem with P@10 is that it does not average well across queries, since the number of correct answers has great variance. R-Precision shows the number of relevant answers within the top-R results, where R is the total number of relevant answers in the set. R-precision overcomes the problem of variance in the number of correct answers [23].

Each ranking formula was tested on 30 queries based on the evaluation topics for entity ranking systems from INEX 2009 and TREC 2010. The usage of these topics was not intended to compare the results of this system to those participating on these tracks, but to evaluate on a set of queries with several degrees of difficulty. The queries were slightly modified to be more specific, since they originally were followed by a narrative for more details. Most of them ask for entities that satisfy more than one condition. In order to accept an entity as relevant, it had to satisfy all the conditions of the query. The correctness of the results was manually checked. The experimental results can be seen on Table 1. The query set is on Table 2.

The six ranking methods achieved similar results, so it is not clear which one is better. The P@10 measure indicates that term frequency does not improve ranking results. As the influence of term frequency increases, P@10 decreases, suggesting that verbosity hypothesis works better for entity ranking. Equations (2) and (5) that represent the middle ground, achieve a higher R-Precision. Assuming the user wants to find all relevant results this method will work better. The reason that (4) is used in the prototype is we expect users to be mostly interested in the first 10 results. Further increase of term frequency influence on ranking, as the scope hypothesis suggests, does not offer any improvement. The ranking of documents does not have a great impact, as expected with a small set of 10 documents, but offers some small improvement except for the case of (6).

TABLE II. THE 30 EVALUATION QUERIES USED IN THE EXPERIMENT

Evaluation Queries
Pacific navigators Australia explorers
List of countries in World War Two
Nordic authors known for children's literature
Makers of lawn tennis rackets
National capitals situated on islands
Poets winners of Nobel prize in literature
Formula 1 drivers that won the Monaco Grand Prix
Formula One World Constructors' Champions
Italian Nobel prize winners
Musicians who appeared in the Blues Brothers movies
Swiss cantons where they speak German
US Presidents since 1960
Countries which have won the FIFA world cup
Toy train manufacturers that are still in business
German female politicians
Actresses in Bond movies
Star Trek Captains characters
EU countries
Record-breaking sprinters in male 100-meter sprints
Professional baseball team in Japan
Japanese players in Major League Baseball
Airports in Germany
Universities in Catalunya
German cities that have been part of the hanseatic league
Chess world champions
Recording companies that now sell the Kingston Trio songs
Schools the Supreme Court justices received their undergraduate degrees
Axis powers of World War Two
State capitals of the United States of America
National Parks East Coast Canada US

The experiment also provided some insight in the system's function. First, we noticed the dependency of performance on the quality of retrieved documents. As expected, queries that resulted in many relevant documents had much more precision in results than others where they had fewer relevant documents. Another problem comes with queries that have a small amount of correct answers (e.g., Axis Powers of World War Two). Determining a cut-off threshold on result scoring so that only relevant results may appear on the list is a difficult task [24].

V. CONCLUSIONS AND FUTURE WORK

We presented a prototype of an online application for entity ranking that uses web documents as data and ranks the entities using information retrieval methods. The application uses various components for recognizing the query topic,

 ISBN: 978-1-61208-227-1

retrieving documents, extracting entities and performing co-reference resolution before the ranking takes place. We experimented with and evaluated several combinations of statistical quantities for ranking entities.

The experiments showed that the combination of rank position for source documents along with a measure of the common information among them yields the best results for ranking. The total frequency of entities did not work very well, verifying the verbosity hypothesis. Furthermore, the experiments showed that using the large data volume of the Web along with a search engine for retrieving them, the system has almost no limitations in query handling.

The application currently supports search for persons, locations and organization. The search can be easily expanded to other types of entities like products, books and movie titles by incorporating them to the extraction stage. The ranking method is very fast but the overall speed of the application is currently confined by the extraction stage which uses machine learning. The necessary processing of this stage though could be done in advance by crawling for documents and extracting information in a similar way that search engines create their indices. With this modification the speed of the ranking method will be fully utilized.

REFERENCES

[1] ListCreator. [Online]. Available: http://listcreator.nonrelevant.net [retrieved: September 2012].

[2] J. Lin. "The Web as a Resource for Question Answering: Perspectives and Challenges". *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, 2002.

[3] G. Dermatini, T. Iofciu, and A. P. de Vries. "Overview of the INEX 2008 Entity Ranking Track". *Lecture Notes in Computer Science, Volume 5631/2009*, 2009, pp. 243-252.

[4] G. Dermatini, T. Iofciu, and A. P. de Vries. "Overview of the INEX 2008 Entity Ranking Track". *Lecture Notes in Computer Science, Volume 6203/2010*, 2010, pp. 254-264.

[5] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. "Overview of the TREC 2009 Entity Track". *Proceedings of the Eighteenth Text RErieval Conference (TREC 2009)*,.2009.

[6] K. Balog, A. P. de Vries, and P. Serdyukov. "Overview of the Trec 2010 Entity Track". *Proceedings of the Nineteenth Text RErieval Conference (TREC 2010)*,.2010.

[7] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. "Open Information Extraction: the Second Generation". *Proceedings of the Twenty-second International Joint Conference in Artificial Intelligence (IJCAI'11)*, 2011, pp 3-10..

[8] M.J. Cafarella, C. Re, D. Suciu, and O. Etzioni. "Structured Querying of Web Text Data: A Technical Challenge". *Proceedings of the Third Conference on Innovative Data Systems Research (CIDR 2007)*, 2007, pp.225-234.

[9] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. "NAGA: Searching and Ranking Knowledge". *Proceedings of the Twenty-fourth International Conference on Data Engineering (ICDE 2008)*, 2008, pp. 953-962.

[10] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M.D. Rijke,. "Mapping queries to the Linking Open Data cloud: A case study using DBpedia". In J. Web Semantics. December 2011, pp.418-433.

[11] Java. [Online]. Available: http://www.java.com/en/ [retrieved: September 2012].

[12] HTML 4.01 Specification. [Online]. Available: http://www.w3.org/TR/1999/REC-html401-19991224/ [retrieved: September 2012].

[13] JavaScript. [Online]. Available: https://developer.mozilla.org/en-US/docs/JavaScript [retrieved: September 2012].

[14] PHP. [Online]. Available: http://www.php.net/ [retrieved: September 2012].

[15] Princeton University (2010). WordNet. [Online]. Available: http://wordnet.princeton.edu [retrieved: September 2012].

[16] Yahoo BOSS API. [Online]. Available: http://developer.yahoo.com/search/boss [retrieved: September 2012].

[17] J. R. Finkel, T. Grenager, and C. Manning. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005, pp 363-370.

[18] GNU General Public License. [Online]. Available: http://www.gnu.org/licenses/gpl.html [retrieved: September 2012].

[19] Named Entity Recognition and Information Extraction [Online] http://nlp.stanford.edu/ner/index.shtml [retrieved: September 2012].

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, 2001, pp.282-289.

[21] Jsoup: Java HTML Parser. [Online]. Available: http://jsoup.org [retrieved: September 2012].

[22] S. E. Robertson and S. Walker. "Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval". *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 345–354.

[23] C. Buckley and E. M. Voorhees. "Retrieval Evaluation with Incomplete Information". *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 25-32.

[24] A. Arampatzis, J. Kamps, and Stephen Robertson "Where to Stop Reading a Ranked List? Threshold Optimization using Truncated Score Distributions.". *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 524–531.

 ISBN: 978-1-61208-227-1

Parallel Processing of Very Many Textual Customers' Reviews Freely Written Down in Natural Languages

Jan Žižka and František Dařena
Department of Informatics
FBE, Mendel University in Brno
Brno, Czech Republic
Email: zizka@mendelu.cz, darena@mendelu.cz

***Abstract*—Text mining of hundreds of thousand or millions of documents written in a natural language is limited by the computational complexity (time and memory) and computer performance. Many applications can use only standard personal computers. In this case, the whole data set has to be divided into smaller subsets that can be processed in parallel. This article deals with the problem how to divide the original data set, which represents a typical collection containing two millions of customers' reviews written in English. The main goal is to mine information the quality of which is comparable with information obtained from the whole set despite the fact that the mining is carried out using subsets of the original large data set. The article suggests a method of dividing the set into subsets including a possibility of evaluating the mining results by comparing the unified outputs of individual subsets with the original set. The suggested method is illustrated with a task that searches for significant words expressing the customers' opinions on hotel services. It is shown that there is always a certain boundary under which the subset sizes cannot fall as well as how to experimentally find this border.**

Keywords-text mining; natural language; parallel processing; decision tree; data subset size; computational complexity.

I. INTRODUCTION

Today, it is important to look for methods which speed up document search and reduce classifier training times and errors for very large text data collections [11]. Using a very large set of real data, this paper describes a parallelism-based procedure that improves the deficiency caused by rapidly increasing computational complexity. Collecting and subsequent processing of customer opinions that relate to a specific matter can usually present a valuable form of feedback. Many organizations and companies allow their users or customers to subsequently express opinions or sentiments, which can be later used for improving the provided services or any related activities with the intention to strengthen competitiveness. As a commonplace, the opinions are in many cases written down by way of the Internet as free unformatted (or with a very limited formatting), not very long text reviews using any natural language. Logically, the more reviews expressing various opinions, the better information can be mined and utilized from the data collection. Today's literature, like [9], describes a lot of different possibilities what we can mine from textual data, from clustering and classification to sentiment analysis to computational linguistic topics.

Looking at the high number of reviews as a naturally positive thing, it is necessary to see also the second coin side: Due to the nonlinear increase of computational complexity, the processing of very many textual items can take also very long time, including escalated memory demands.

In this article, we describe their experience with opinion mining from large textual data containing hundreds of thousands to millions of reviews written down by customers of on-line hotel services. The method itself of text mining was published, for example, in [2][13][14][15][16]. However, the mining in question had to face up to the high computational complexity caused by the big data volume. Altogether, there were more than five millions of customer reviews written in more than 50 natural languages. The most of reviews were written in English (almost two millions), following from more than 700,000 to more than 300,000 in French, Spanish, German, and Italian, to mention just the largest data sets. The original task was to mine significant words and phrases representing the customers' opinions concerning the hotel services booked on-line.

Among the main intentions, there was also the investigation how possible was the realistic text-mining using a common personal computer, PC, (as a 64-bits four-kernel processor 2.0 GHz, 8 GB RAM, 64-bits MS Windows 7 Professional) supposing that the hotel service provider had no access to a super-computer. The experiments quickly showed that it was not possible to process the big data sets en bloc, either because of insufficient memory or very long computational times (weeks), even if the mining procedure applied a professional implementation of the decision tree generator *c5/See5* that is based on the entropy minimization, see [7], that worked with RAM very well. However, in the beginning when *c5/See5* needs to read large data, it consumes a lot of memory, too.

After some experiments, the authors had to accept a natural solution based on dividing the whole data set into smaller subsets that could be processed in parallel using several common PC's. Different authors applied parallelism to various problems connected to very large data sets. For example,

 ISBN: 978-1-61208-227-1

Ulmer et al. [12] created a text document-similarity classifier used to detect web attacks in HTTP data streams. They applied a parallel hardware approach because a sequential algorithm could not process a real-time data stream above certain data volumes. The parallel approach was also used for text feature selection – the process was parallelized and demonstrated using a cluster formed with several computers [6]. For data divided into several broad domains with many sub-category levels, separate classifiers of the same type could be trained on different subspaces in parallel. An improvement in subspace learning was accompanied by a very significant reduction in training times for all types of used classifiers [11]. Lertnattee and Theeramunkong [5] parallelized and distributed the process of text classification separately in each dimension. Classifiers learned from large training documents with a small number of classes on each dimension, and the best classifiers for each dimension were then combined. Both learning and classification phases run in parallel. Hao and Lu [3] developed a modular version of the k-nearest neighbor algorithm (k-NN) which was a faster and more efficient method for large-scale text categorization by direct modular classification without reducing the precision of the classification. The algorithm decomposed the large-scale text categorization problem into a number of smaller two-class subproblems and combined all of the individual modular k-NN classifiers info one classifier.

As the experiments described further showed, it was not negligible how large the subsets were (how many reviews they contained) because a dictionary of each subset was logically not identical, some significant words were not in all the data parts, or their significance – based on the frequency representation – markedly changed. Such a behavior of the textual data can be expected due to the high sparsity of vectors representing individual reviews.

In the following sections, a reader can find the English data description (Section II), the design of experiments (Section III), the results and their interpretation (Section IV) and, finally, conclusions (Section V).

II. Characteristics of the Examined Text Data

The investigated textual data represented usual customers' reviews written freely in natural languages, without following any specific structure or form. In all of the languages, the hotel service customers were satisfied or dissatisfied with the same or very similar, typical things (cleanness, price, personal willingness or helpfulness, noise, price, hotel position, food, and so like).

The results presented in this article come from the largest data set that contained the customers' reviews written in English, however, there were no significant differences in other 'big' languages (from the data volume point of view) mentioned in the Introduction section. It is necessary to emphasize the fact that not all authors of English reviews were English native speakers – the natural reason was that people all over the world use English, 'international English', as a universal communication means. As a result, the reviews contained many imperfections coming from lower knowledge of English or mistyping. This language incorrectness brings certain consequences like the artificial extension of the word list (dictionary) where a word can have many variations but only one is correct, for example, '*behavoir, behavior*', '*acomodation, accommodation, accomodation, acommodation*', '*noise, nois*', and so like. Such imperfections could be subsequently corrected by spell-checkers, however, without a human control (that could be for large data impossible) the result would not be guaranteed – fully automatic check-spelling can introduce additional errors. One possibility could be applying a spell-checker during writing a review but it would also need spell-checkers for all acceptable languages. Similarly, the customers used often also interjections like '*goooooood, goood*', '*aaarrrgh-hhh*', '*uuugly*', and so like, to express their dis/satisfaction with the service.

Sometimes, the English text contained also non-English terms when a customer could not remember a word, for example '*albergo*', which in Italian means '*hotel*'. In some cases, there were reviews written in two languages but they were assigned to English because customers wanted to express the opinion in their native language, however, if their native language belonged to a group of 'small' languages, as for example Czech, the opinion contained also its (not always correct) English version – one could not expect that hotel managers in, say, South America knew Czech. Here are some original examples of reviews without corrections:

- *breakfast and the closeness to the railwaystation were the only things that werent bad*
- *did not spend enogh time in hotel to assess*
- *it was somewhere to sleep*
- *very little !!!!!!!!!*
- *breakfast, supermarket in the same building, kitchen in the apartment (basic but better than none)*
- *no complaints on the hotel*

Overall, the English dictionary generated from the English group of reviews contained some 200,000 words in almost 2,000,000 reviews, which could be represented by matrix having two millions rows and 200,000 columns – a really large matrix containing ca 4×10^{11} numbers where each number meant the frequency of a word in a review.

The dictionary contained only *words*, which means that all numbers, punctuation symbols, or any special marks were excluded. Sometimes, the dictionary contained peculiar 'words' like '*t*' but it resulted from the preprocessing of the original words like *didn't* after using the apostrophe as one of delimiters, therefore *didn't* was transformed into two 'words' *didn* and *t*. In addition, all characters were transformed into the lower-case representation to avoid having more versions of the same term – even if there could be

 ISBN: 978-1-61208-227-1

some loss of information, for example *Rose* (a hotel name) and *rose* (a plant); however, such cases were extremely rare, without influencing the results.

Each review was transformed into vector, which is a standard representation method. This representation contains all possible dimensions (that is, words in the dictionary), however, because of the word number per review and the word number in the dictionary, the vectors are extremely sparse, containing zeros in most of word positions because the minimum review length was one word (for example, *Excellent!!!*), the maximum was 167 words, and the average length of a review was 19 words. The vector sparsity was typically around 0.01%, that is, on average, a review contained only 0.01% of the words in the dictionary.

Comparing those word numbers with the dictionary size, it is clear that the vectors were very sparse, containing mostly zeroes for the word frequencies. Still, a human reader could say what reviews were positive, negative, mixed, neutral, or non-classable, and what terms were significant from the positive or negative standpoint, attitude, or sentiment: *noisy, quiet, smell, helpful personnel, good but small food portion, dirty rooms, nice hotel position*, and so like. As the previous research showed, see [15], an overwhelming majority of words were insignificant, only some 300 of terms played the significant role from the classification point of view (either a negative or positive review); the huge majority from almost 200,000 dictionary words had no function. That vector sparsity influenced the results of dividing the original data set into smaller subsets to decrease the computational complexity.

III. Experiments for Finding the Subset Size

The experiments were aimed at finding the optimal subset size for the main data set division. The *optimum* was defined as *obtaining the same results from the whole data set and the individual subsets*. Such a non-mathematical definition ideally meant that each subset should provide the same significant words that would have the same significance for categorizing reviews into correct classes; here, positive and negative opinions where the review positivity or negativity was given by a customer. As the subsets contained different, randomly selected reviews, the similarity of the subset results were defined as an average value.

The *word significance* was defined as *the number of times when a decision tree asked what was a word frequency in a review*. Obviously, the most significant words are tested every time for each classification query, which is, for example, quite typical for a word in the tree root, even if there could also be other words tested in 100% cases. Usually, the word frequency tests on lower tree levels do not check the words so often as on the higher levels due to the wide tree branching. The words included in the tree are in fact the relevant attributes from the classification point of view; other words are irrelevant and could be calmly omitted – but it is not how people create sentences understandable for them.

Therefore, if there is a word from the whole data set R in the root, most of the n subsets (ideally all) should have the same word in their roots. Similarly, the same rule can be applied to other words included in the trees on levels approaching the leaves. Then we could say that each subset represents the original set perfectly. In reality, the decision trees generated for each review subset r_i more or less mutually differ because they are created from different reviews. In addition, a tree generated from a subset r_i may contain also at least one word that is not in the tree generated from R. Each tree provides a set w_{r_i} of significant words. The union w_r of the sets of significant words w_{r_i} should give a resulting set that should ideally have the same words as in the whole review set R with the word set w_R provided by the tree generated for R:

$$r_i \subset R,\ w_{r_i} \subset w_R\ , \tag{1}$$

$$w_r = \bigcup_{i=1}^{n} w_{r_i}\ , \tag{2}$$

$$\text{therefore ideally, } w_R = w_r\ , \tag{3}$$

for $i = 1,\ \ldots,\ n$.

Thus, the question is: How many subsets should the whole review set R be divided into so that the unified results from all r_i's provide (almost) the same result as from R? Intuitively, if each r_i would contain just one review, the result can be bad because the individual reviews are typically very different even if they refer to the same thing: *bad accommodation, not good accommodation, horrible accommodation, we were not satisfied with the accommodation, excellent accommodation, relatively good accommodation*, and so like. The only shared word is *accommodation*, however, it itself is not either positive or negative, it is simply neutral. The adequate decision trees would be very different.

If those reviews would be grouped into one common set, the adequate tree would be also very different from the previous individual trees and, moreover, it would represent certain generalization, that is, knowledge. Provided that a computer cannot process the whole set R, the intuitively best way would be to create n as large subsets r_i as possible so that the computer could process its r_i as quickly as possible without the preliminary depletion of memory. Then, having n computers, the reviews could be processed during the time acceptable by a user.

Obviously, it is not easy to find a general solution because the result depends on particular data. The authors selected the data described above because it corresponded to many similar situations: a lot of short reviews concerning just one topic.

 ISBN: 978-1-61208-227-1

Firstly, the original set R was too big to be processed as a whole: two millions reviews. For a given PC model, the authors looked for the maximal size of R using a random selection from the whole original set. Selections containing more than 300,000 reviews crashed because of insufficient PC memory (8 GB RAM). The sets with 300,000 reviews (and more) were not ready after a week, therefore the computations had to be canceled.

In this place, it is worth to remark the computational complexity of the *c5/See5* decision tree type. In [10], the authors mention the time complexity of the c4.5 (a forerunner of c5) decision tree generator. The upper boundary is $O(m \cdot n^2)$, where m is the size of the training data (the number of matrix rows) and n is the number of attributes (the number of words in the dictionary).

The subset containing 200,000 reviews (10% of the whole original data) consumed almost 85,000 seconds of elapsed time (approximately 24 hours), therefore it was accepted as the largest processable R. Similarly, there were successively created smaller R's: 100,000, 50,000, and 20,000 (plus other sizes, but there were no big differences in the results between R's with similar sizes). Each of R was processed to obtain its particular significant words.

After that, each of the generated R's was randomly divided into smaller r's so that the individual sizes of each r_i represented 10%, 20%, 25%, 30%, 40%, and 50% of its adequate 'parent' R.

In the second step, every data set was preprocessed using the commonly known method called *bag-of-words*, see for example [8]. The reason was that linguistic preprocessing was impossible due to the too large data volume and not the same method for any language. In addition, all words appearing only once in the whole data set were removed which decreased the number of words, n, in the dictionary almost to a half, and the computational complexity even more because $O(m \cdot n^2)$ depends strongly on n^2.

The words were represented by their frequencies in reviews. As it was mentioned above, in each vector there were mostly zeros. The subsequent experiments tried the more advanced representation called *tf-idf* (term frequency times inverted document frequency, see for example [8]), however, the results were not better (maybe because the sizes of reviews were very similar – typically tens of word).

For each R and r_i, the third step gradually generated the decision trees to reveal the significant words as the relevant attributes for the classification to the positive or negative opinion class. Typically, the results looked similarly like this: *100% location, 80% friendly, 79% not, 73% excellent, 68% helpful, 63% closeness, 63% helpfulness, 63% friendliness, 62% comfortable, 62% spacious, ..., 5% facilities, 5% and, 4% nothing, 3% on, 2% door, 2% with, 2% to, 2% so, 1% in*, and so on (in this example, there were 167 significant/relevant words in the tree; in other cases, it was similar). The first word always represented the root – the tree asked the *location* frequency always, *friendly* frequency in 80%, and so on. The percentage value plays here the role of the *significance weight* because the frequencies of words that are closer to the root contribute more to the entropy decrease than frequencies of words on levels closer to leaves.

The basic result was always given by an R set. The lists of significant/relevant words generated for individual r_i's, where $r_i \subset R$, were compared with the basic result. The authors were interested in the fact how much each significant word in r_i corresponded to the same word in R from the percentage point of view, S_R – that is, a word in the R tree had its percentage equal to S_R. The sets of significant words generally contained a lot of the same words, even if there were also words that were not included in all r_i trees. For the r_i's common percentage of a given significant word, S_r, it was taken the average value:

$$S_r = \frac{1}{n} \sum_{i=1}^{n} S_{r_i} \ , \tag{4}$$

where n is the number of r's (subsets of R) and S_{r_i} is the percentage of the word in the i-th subset r_i.

Then, it was possible to compare S_R's with S_r's for each word and subset. As it was expected, dividing an R set into less but larger r_i subsets provided better results – the correspondence between R and its r's was closer to the ideal than in the case of more smaller subsets r. On the other hand, smaller subsets were processed noticeably faster than the larger ones. One of the reasons was the fact that each r_i contained only part of the total dictionary generated from R – consequently, smaller dictionaries of r_i's decreased also the computational complexity $O(.)$ The results of mining significant words are demonstrated in the following section.

IV. Results of Experiments

To compare results provided by the review sets and subsets having various number of items, the authors used a method that is illustrated in the following graphs Figure 1, Figure 2, and Figure 3.

On the horizontal axis, there are significant words generated by the trees. The graph does not show all significant words because of insufficient space; only the words having the higher percentage value are here used.

The vertical axis y shows the correspondence between the percentage of the significant words in the relative R set and the average percentage of the relevant r_i subsets. The whole set R contains all the significant words which means that the y value is always 1.0 (that is, 100%). In other words, the occurrence of significant words w_i in R is given by a simple equation:

$$y_R(w_i) = 1.0 \ . \tag{5}$$

On the other hand, some words in some r_i's could be missing. In the case of individual r_i's, the occurrence of

 ISBN: 978-1-61208-227-1

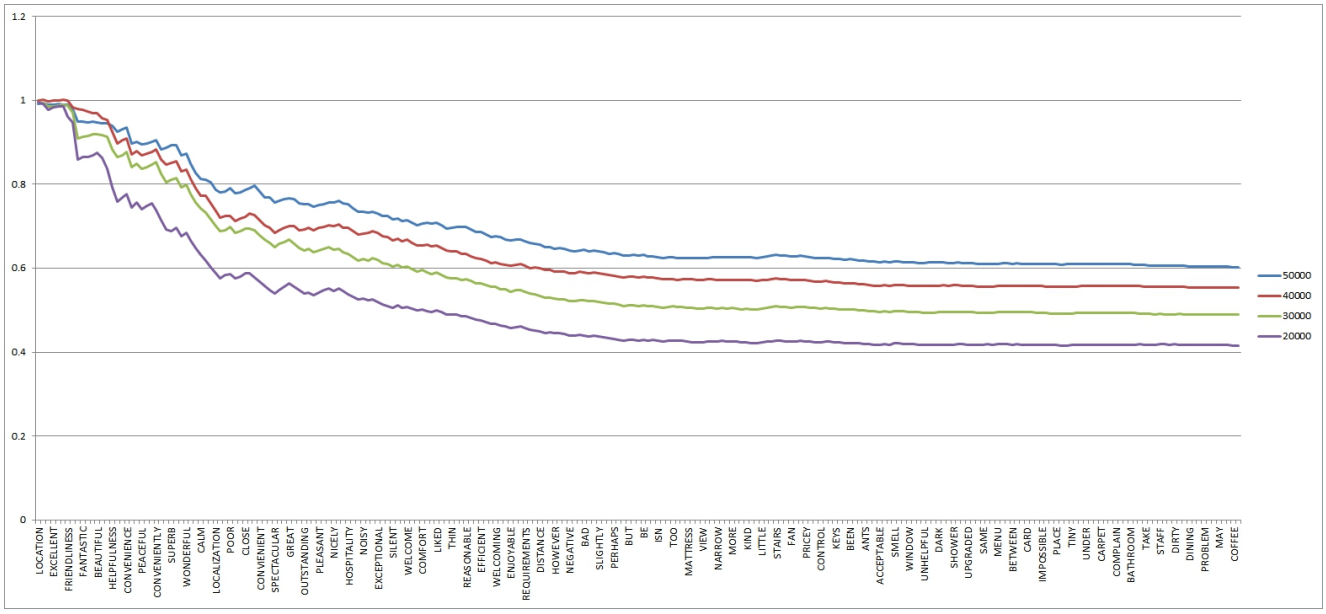


Figure 1. The whole set R with 200,000 reviews divided into subsets r_i having gradually 50,000, 40,000, 30,000, and 20,000 reviews

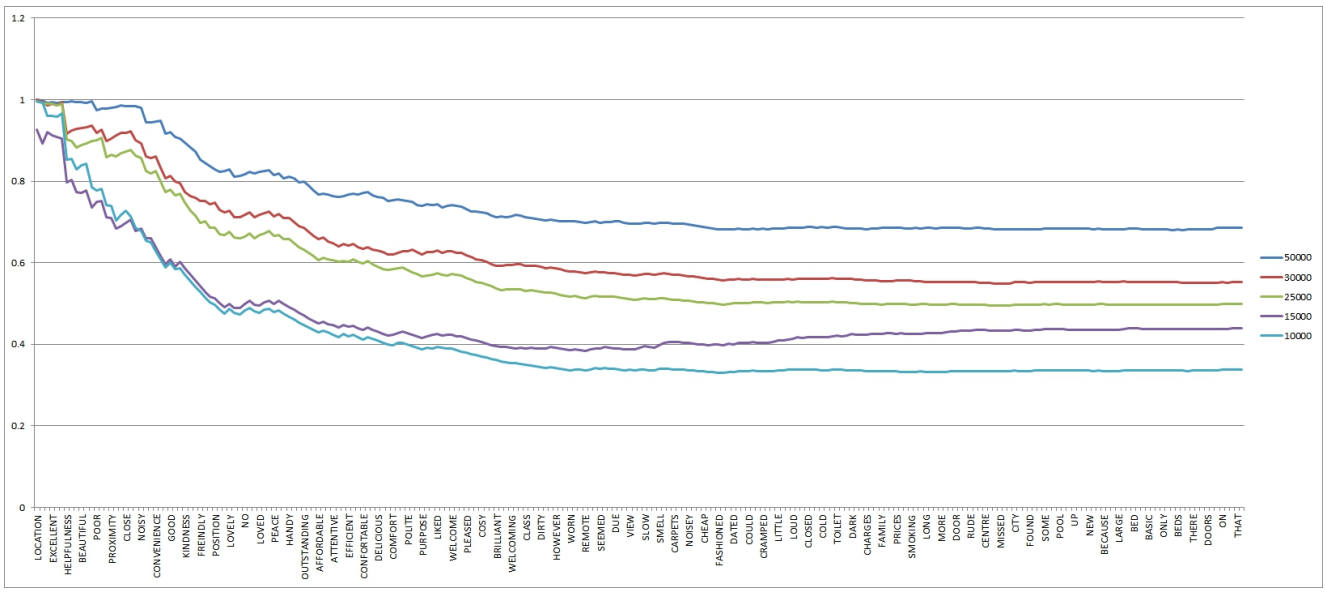


Figure 2. The whole set R with 100,000 reviews divided into subsets r_i having gradually 50,000, 30,000, 25,000, 15,000, and 10,000 reviews

significant words is expressed using the following formula, where for a word w_i the value on the y axis is calculated as:

$$y_r(w_i) = \frac{\sum_{j=1}^{i} S_r(w_j)}{\sum_{j=1}^{i} S_R(w_j)} \ , \qquad (6)$$

where i is the serial number of a word, S_R is the percentage of usage of a word w_j given by the decision tree and created for the complete data set R, and S_r is the average percentage of usage of a word w_j by the decision tree created for every subset $r \subset R$. The same word can have different percentage values in different subsets r as well as in the relative set R, therefore $y_r(w_i) \neq y_R(w_i)$. Ideally, all significant words should be at the same tree position having the same weight; then, $\forall i, y_r(w_i) = y_R(w_i) = 1.0$.

Equation 6 measures the agreement between the percentage weight of a word w_j in the tree generated for R and the average value in the trees generated for all r_i's, where $r_i \subset R$. For example, if a whole set R would be randomly divided into $n = 5$ subsets, and a certain

 ISBN: 978-1-61208-227-1

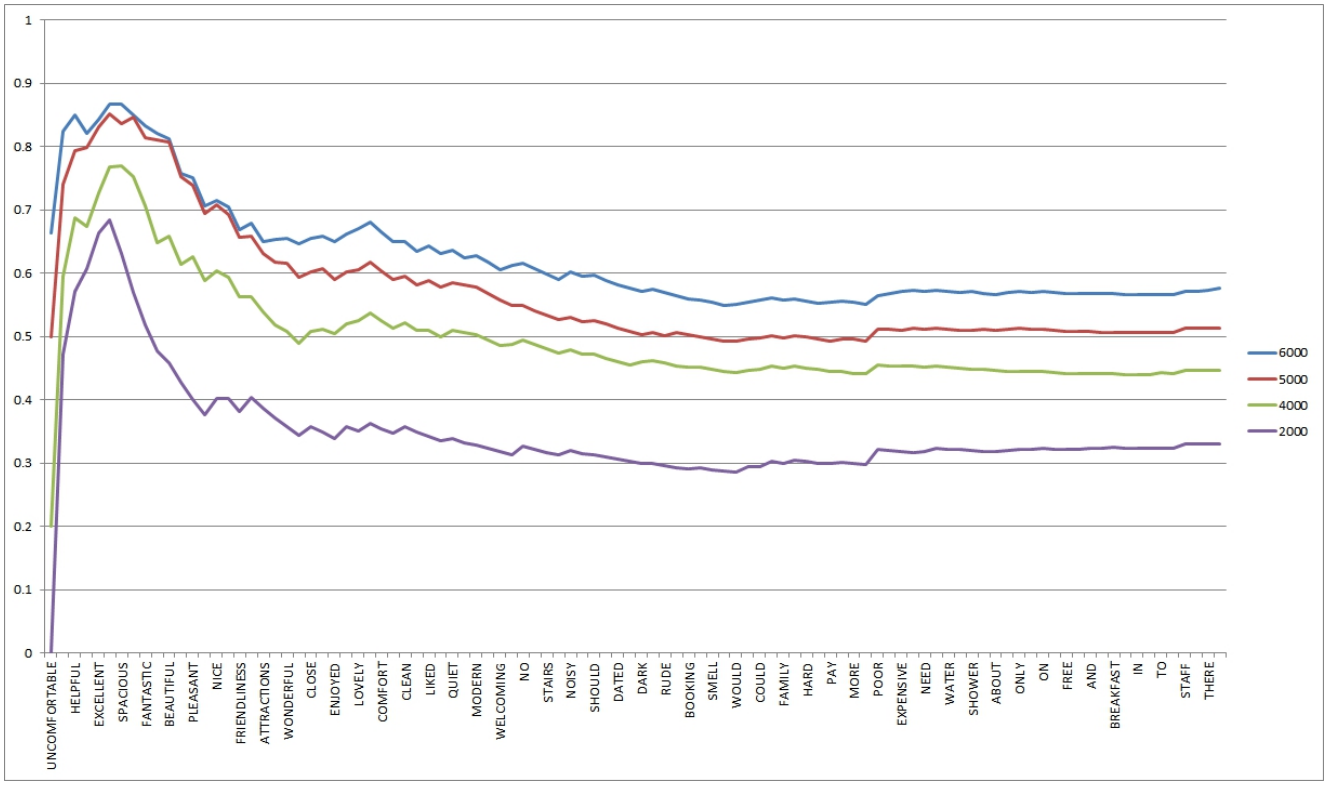


Figure 3. The whole set R with 20,000 reviews divided into subsets r_i having gradually 6,000, 5,000, 4,000, and 2,000 reviews

word $w_j = excellent$ would have its percentage weight $S_R(w_j) = 73\%$, then if all r_i's would have the same word weight $S_{r_i}(w_j) = 73\%$ for $i = 1, \ldots, 5$, the agreement is perfect, that is, $y(excellent) = 1.0$; otherwise, the results provided by the subsets may differ from the whole set.

The graphs in Figure 1, Figure 2, and Figure 3 show how the averaged results of subsets r_i agree with the results obtained from the complete review sets R for individual words that are at the top of the percentage list. Each individual curve represents an average result for r_i's that have a certain number of reviews (see also the graphs legends).

For example, Figure 1 illustrates the situation when R contains 200,000 reviews. After dividing R into four subsets $r_i, i = 1, 2, 3, 4$, where each r_i has 50,000 randomly selected reviews, it is possible to see that the correspondence (computed using Equation 6) is better than 80% for the first 13 significant words with high percentage weights. Then the similarity gradually decreases, but never under 40%. The curves also show that dividing R into 10 subsets r_i's (20,000 reviews per r_i) provides worse results than for the less number of larger r's.

Similarly, Figure 2 illustrates the situation for R containing 100,000 reviews, and Figure 3 for 20,000 reviews (here are the results markedly much worse – in addition, no r_i contained the R's root word *uncomfortable*). The experiments were carried out for various subset sizes and whole sets, however, the results were quite consistent, therefore they are not here illustrated all – only the three most characteristic ones.

V. Conclusion

Interestingly and predictably, the graphs illustrate the fact that the higher number of smaller subsets provide altogether worse results than the lower number of larger ones. Naturally, the complete review set R provides the best result as one extreme, and subsets (singletons) of R, containing just single reviews, give the worst results as the contrary extreme (not shown here because it is not interesting – at least, no one would process 2,000,000 reviews using 2,000,000 computers in parallel).

When the R data volume is too large to be processed using one PC, it has to be divided into smaller subsets r. It is probably not a big surprise that the smaller subsets should be as large as possible, however, the authors needed an empirical proof that randomly divided original sets R into subsets can provide similar (if not identical) results by unifying the results of all individual subsets using some large real-world data. Also, it was necessary to test what subset sizes could be used to obtain reliable results within a reasonable time (max. several hours, not many days).

 ISBN: 978-1-61208-227-1

The *result reliability* is a rather 'fuzzy' concept; it naturally depends on a user what he or she would accept as *reliable*. However, in reality, users mostly have no choice – standard PC's do not enable processing of such large data volumes, thus it is very useful to know how the data having the similar properties as the one analyzed here should be prepared for the parallel processing that radically decreases the computation complexity (both time and memory).

ACKNOWLEDGMENT

The research work published in this paper was supported by the Research program of the Czech Ministry of Education VZ-MSM-6215648904.

REFERENCES

[1] http://www.rulequest.com/see5-info.html, September 2012.

[2] F. Dařena and J. Žižka, "Text Mining-based Formation of Dictionaries Expressing Opinions in Natural Languages." In: Proceedings of the 17^{th} International Conference Mendel-2011, June 15-17, 2011, Brno, Czech Republic. No. 1, pp. 374–381, Brno Technical University Press, 2011.

[3] H. Zhao and B. L. Lu, "A modular k-nearest neighbor classification method for massively parallel text categorization." Lecture Notes in Computer Science, Vol. 3314, 2004, pp. 867-872.

[4] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews." In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'04, August 22-25, 2004, Seattle, Washington, USA. ACM, 2004.

[5] V. Lertnattee and T. Theeramunkong, "Parallel text categorization for multi-dimensional data." Lecture Notes in Computer Science, Vol. 3320, 2004, pp. 38-41.

[6] M. J. Meena, K. R. Chandran, A. Karthik, and A. V. Samuel, "An enhanced ACO algorithm to select features for text categorization and its parallelization." Expert Systems with Applications, Vol. 39, No. 5, 2012, pp. 5861-5871.

[7] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning, 1992.

[8] F. Sebastiani, "Machine learning in automated text categorization." ACM Computation Survey 34, 1, March 2002, pp. 1-47.

[9] A. N. Srivastava and M. Sahamimph, Text Miming: Classification, Clustering, and Applications. Chapmann and Hall/CRC, New York, USA, 2009.

[10] J. Su and H. Zhang, "A Fast Decision Tree Learning Algorithm." In: Proceedings of the 21^{st} National Conference on Artificial Intelligence and the 18^{th} Innovative Applications of Artificial Intelligence Conference, July 16-20, Boston (MA), USA. AAAI Press, 2006.

[11] N. Tripathi, M. Oakes and S. Wermter, "A fast subspace text categorization method using parallel classifiers." Lecture Notes in Computer Science, Vol. 7182, No. 2, 2012, pp. 132-143.

[12] C. Ulmer, M. Gokhale, B. Gallagher, P. Top, and T. Eliassi-Rad, "Massively parallel acceleration of a document-similarity classifier to detect web attacks." Journal of Parallel and Distributed Computing, Vol. 71, No. 2, 2011, pp. 225-235.

[13] J. Žižka and F. Dařena, "Automatic Sentiment Analysis Using the Textual Pattern Content Similarity in Natural Language." Lecture Notes in Computer Science No. 6231, Springer, Heidelberg, Germany, 2010, pp. 224-231.

[14] J. Žižka and F. Dařena, "Mining Textual Significant Expressions Reflecting Opinions in Natural Languages." In: Proceedings of the 11^{th} International Conference Intelligent Systems Design and Applications. Cordoba, Spain, November 22-24, 2011, pp. 136-141.

[15] J. Žižka and F. Dařena, "Mining Significant Words from Customer Opinions Written in Different Natural Languages." Lecture Notes in Computer Science No. 6836, Springer, Heidelberg, Germany, 2011, pp. 211–218.

[16] J. Žižka and V. Rukavitsyn, "Automatic Categorization of Reviews and Opinions of Internet E-Shopping Customers." International Journal of Online Marketing, Vol. 1 No. 2, IGI Global, USA, 2011, pp. 68-77.

 ISBN: 978-1-61208-227-1

A New Algorithm for Accurate Histogram Construction

Zeineb Dhouioui
Computer Science Department
ISG
41, rue de la Liberté, 2000 Le Bardo Tunisia
dhouioui.zeineb@hotmail.fr

Wissem Labbadi
Computer Science Department
ISG
41, rue de la Liberté, 2000 Le Bardo Tunisia
wissem.labbadi@isg.rnu.tn

Jalel Akaichi
Computer Science Department
ISG
41, rue de la Liberté, 2000 Le Bardo Tunisia
jalel.akaichi@isg.rnu.tn

***Abstract*—**Many commercial relational database systems use histograms to summarize data sets and also to determine the frequency distribution of attribute values. Based on this distribution, a database system estimates query result sizes within query optimization useful in effective information retrieval. Moreover, histograms are beneficial for judging whether the quality of the source is reliable or not; therefore, they enable us/ one to decide whether to keep this source in the information retrieval or remove it. Each histogram contains commonly an error which affects the accuracy of the estimation. This work surveys the state of the art on the problem of identifying optimal histograms, studies the effectiveness of these optimal histograms in limiting error propagation in the context of query optimization, and proposes a new algorithm for accurate histogram construction. As a result, we can conclude that theoretical results are confirmed in practice. In fact, the proposed histogram generates a low error.

Keywords- Optimal histograms; query result size estimation; error; query optimization; data summarization.

I. INTRODUCTION

Information retrieval is a science that studies how to respond effectively to a request by finding the appropriate information in a huge stream of data. The need to use information retrieval systems (IRS) has increased with the rapid evolution of information technology and communication and also with the proliferation of computer data and their sources. Diversity and heterogeneity of information sources require the use of IRS that must meet user expectations and needs and provide relevant information among the mass of available information in the shortest time and with reduced cost. However, in front of the fast growth of data databases have witnessed an exponential increase of stored data which make it increasingly difficult to control and effectively manage the potentially flow of information.

A straightforward way to satisfy an information need is to send the query to all sources, and to get results from each one, which are then provided to the user. However, this strategy is not efficient in front of the big number of dispersed sources. This simple method incurs unnecessary cost and an additional waiting time when sending the query to sources not containing the required information [1]. For this reason, more efficient information retrieval techniques that can extract relevant information from large scale distributed sources are needed in order to satisfy users' requirements in the shortest waiting time. The idea suggested for overcoming this problem is to associate to each source a summary which is a compact representation of its content. Then, the interestingness of a given data source with respect to the user requirements, expressed into a query, is assessed by processing this query against the summary. This operation is a simple match and doesn't need to send the query to the considered source and manipulate its big mass of data. Data summary techniques provide concise and complete representation of data and are now considered as accurate tools to handle huge databases, in particular when precise values of data are not needed.

Many commercial DBMSs [3] maintain a variety of types of histograms to summarize the contents of the database relation by approximating the distribution of values in the relation attributes and based on them estimate sizes and value distributions in query results. A histogram approximates the distributions by grouping the data values into buckets. This grouping into buckets loses information. This loss of information engenders errors in estimates based on these histograms. The resulting estimation-errors directly or transitively affect the accuracy of the resulting estimates and hence, degrade the dramatically the performance of the applications using these estimates. This effect may be devastating in the most cases. For multi-join queries that are processed as a sequence of many join operations, the transitive effect of error propagation among the intermediate results on the estimates derived for the complete query may be destructive even if the original errors are small. Motivated by the fact that inaccurate estimations can lead to wrong decisions, we propose in this paper an efficient algorithm, called CM, for accurate histogram constructions. The survey is organized as follows. Both theoretical and effective experiments are done using two datasets.

The remainder of this paper is organized as follows: Section 2 presents the basic definitions on histograms. Section 3 provides an overview of several earlier and some more recent classes of histograms that are close to optimal and effective in many estimation problems. In Section 4, we present the proposed histogram the CM Histogram. In

 ISBN: 978-1-61208-227-1

Section 5, we formulate the main problem. In Section 6, we propose HConst Algorithm for accurate histogram constructions. In Section 7, we present the result of a set of experiments that compare the existing histograms to our algorithm in term of estimation accuracy. Finally, Section 8 concludes and outlines some of the open problems in this area.

II. BACKGROUND

Histograms are widely used as a data summarization approach. They present an efficient and powerful way to capture the data distribution and also estimate the query result. Histograms are composed of a set of buckets where each bucket contains the frequency of occurrence of each attribute value histogram on attribute X of relation R divides the data distribution into buckets.

We assume the following parameters [2]:

- Domain D of X: it is the set of all possible values of X
- Value set V ⊆ D: is the set of values of D present in R
- Frequency f_i of $v_i \in V$ is the number of tuples $t \in R$

The data distribution DD of the attribute X in the relation R is the set of pairs which comprises the attribute value and its frequency: $DD_{i=1...D,\ D \le |X|} = \{(v_i, f_i) \ ... \ (v_D, f_D)\}$

TABLE I. EXAMPLE OF A DATA DISTRIBUTION

V_i	f_i
180	2
250	1
260	1
270	2
320	1
345	1
380	1
410	1
450	3
490	1
550	1

III. STATE OF THE ART

In this section, we present several earlier and other relative recent histograms, listed in the literature and considered as optimal in estimating range query result sizes.

A. Trivial Histogram

This kind of histograms [3] is simple because it is based on uniform distribution assumption.

It is composed of one single bucket where the approximate frequency is identical for all attribute values [4].

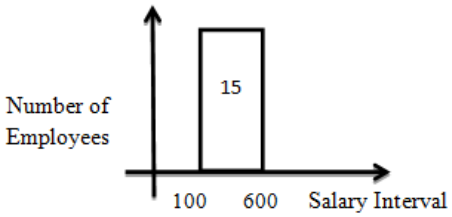


Figure 1. Data Distribution with Trivial Histogram

This type of histogram is based on the principle of uniform distribution. Therefore, the appearance of all values is equally likely; this means that each value appears in the data set a single time.

In this example, we have one single bucket in which the frequency of each value is identical to others.

Trivial histograms have usually a large error rate in query estimation; we will prove this with a selection query.

In our work, we will focus in select queries which allow us to select records according to a specific criterion. Take this example of the selection query and find the number of the employees who have a salary upper than 450.

```
Query:  Select count (*)
From employees
Where salary > 450;
```

According to the histogram, we have at worst 15 values that can be greater than 450 but actually, we have two values greater than 550 and 490.

Subsequently, the corresponding absolute error is: Eabs=|2-15|=13.This is considered a very large error rate.

B. Equi-width Histogram

The idea is to divide the data distribution into buckets. The same width is maintained in all buckets. We apply Equi-depth at the proposed dataset:

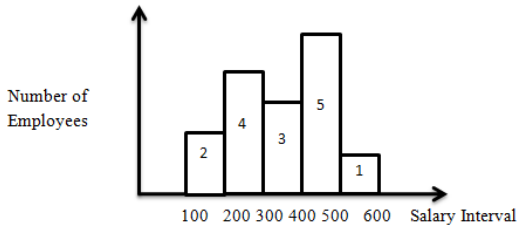


Figure 2. Data Distribution with Equi-width

The height of each bucket presents the total of the frequencies of all attribute values falling in this bucket [5].

The problem in this type of histogram lies in the precision because the error rate is large. This will be proved with the previous query. In the worst case, the maximum error rate of the selection query is half of the height of the bucket.

This case is called unlucky distribution of attribute values where the tallest bucket contains almost 100% of the tuples; then, the error rate is equal to 0.5.

 ISBN: 978-1-61208-227-1

In our example, we have 2 values superior than 450, but according to the histogram, we have 6 values: five values in the interval [400-500] and one value in the interval [500-600].Consequently the absolute error:

Eabs=|2-6|=4

We can conclude that the problem is in the height of buckets, hence the idea of creating a new type of histogram: the equi-height histogram described subsequently.

C. *Equi-depth Histogram*

In equi-depth histogram [4], called also equi-height, the sum of the frequencies in each bucket is the same.

For the construction of this histogram, we must first sort the attribute values in an ascending order to obtain a height balanced histogram. The representation of the attribute values from the previous example consists in a histogram containing seven buckets having equal heights. The threshold per bucket is equal to two as it is shown in the following figure:

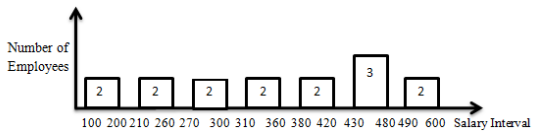


Figure 3. Data distribution with Equi-depth Histogram

The equi-height is more accurate than the equi-width as we will prove with the previous query.

According to the previous query, we can estimate the absolute error:

E_{abs}=|2-4|=2

D. *V-optimal Histogram*

V-optimal Histograms are also called Variance-optimal [4].The basic idea of this histogram is to minimize the weighted variance inside each bucket [6].

The weight here is the number of attribute values in the j^{th} bucket.

$$\sum_{j=1}^{\beta} nj\ vj \quad (1)$$

Where:

- j is the number of buckets.
- n_j is the number of entities in the j^{th} bucket.
- v_j is the variance between the values of the entities in bucket.
- β is the maximum number of buckets

We apply V-optimal histogram to our example; the corresponding absolute error is equal to:

E_{abs} =|2-5|=3

E. *MaxDiff Histogram (Maximum Difference)*

In a MaxDiff histogram, there is a bucket boundary between the adjacent values which have the maximum difference [3]. We compute [7] the difference between $f(v_{i+1})*S_{i+1}$ and $f(v_i)*S_i$.

Where:

- S_i is the spread of attribute value v_i

  $S_i = v_{i+1}-v_i$ (2)
- $f(v_i)* S_i$ is the area of v
- $f(v_i)$:frequency of v_i

We apply max-diff histogram to our example; we separate the adjacent values with a large change in the area.

TABLE II. COMPUTING THE SPREAD, AREA AND Δ AREA

Value	180	250	260	270	320	345	380	410	450	490	550
Frequency	2	1	1	2	1	1	1	1	3	1	1
Spread	70	10	10	50	25	35	35	40	40	60	-
Area	140	10	10	100	25	35	35	40	120	60	-
Δ Area	130	0	90	75	10	0	5	80	60	-	-

TABLE III. Max-diff Histogram

Bucket	Frequency
[100-180]	2
[200-260]	2
[270-300]	2
[320-400]	3
[410-460]	4
[480-600]	2

We compute the absolute error corresponding to the previous query:

E_{abs}=|2-4|=2

F. *Compressed Histogram*

In this type of histogram, we assign the n highest source values in n individual bucket and we apply the equi-heigth histogram on the rest [6].

n is the number of values that exceeds the sum of all the frequencies, Sum_F, that is divided by the number of buckets B:

$$n > \frac{SumF}{B} \quad (3)$$

$n > \frac{15}{7}$=2.14

Hence, we affect the frequencies which are upper than 2.14 to an individual bucket.

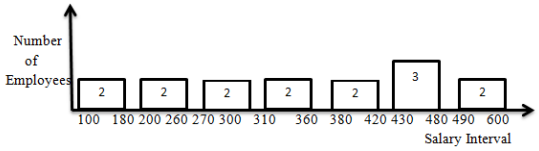


Figure 4. Data Distribution with Compressed Histogram

It looks like end-biased histogram since it distinguishes the highest value of the others, but it differs in the organization of the remaining values.It is an improvement of equi-depth.

The most frequent value is 450 which belong to the interval [430-480].The corresponding absolute error is equal to:

E_{abs} =|2-5|=3

G. Error Metrics

The authors in [8] present the error metrics; we define some concepts useful to compute the error:

- $q_{i:}$ is a query
- A_i: the real size
- A_i': the estimated size found using the histogram
- N: the number of queries.

There are many types of error metrics; the first one is the absolute error computed by the formula:

$$E_{abs} = |A_i - A_i'| \quad (4)$$

The average absolute error is the ratio between the absolute error and the number of [9].

$$AE_{abs} = \sum_{i=1}^{N} Eabs / N \quad (5)$$

And the second type is the relative error:

$$E_{rel} = E_{abs} / A_i = |A_i - A_i'| / A_i \quad (6)$$

The average relative error is the ratio between the relative error and the number of queries [9].

$$AE_{rel} = \sum_{i=1}^{N} Erel / N \quad (7)$$

We can also use average standard deviation which can be computed directly from the histogram.

We find similarly the SSE Sum of Squared Error [8].

For an interval I [i , j] the SSE is calculated by the following formula:

$$SSE([i,j]) = \sum_{k=i}^{k=j} (F[k] - AVG([i,j]))^2 \quad (8)$$

$$AVG(i,j) = \sum_{k=i}^{k} \frac{F[k]}{j-i+1} \quad (9)$$

Where:

F[k]: frequency of the element k

IV. CM HISTOGRAM

The problem of histogram construction is primordial in many tasks in databases. Therefore, many researches have been developed extensively in the past since histograms are characterized by their popularity, accuracy and simplicity in representing the data distribution efficiently.

The idea behind all histograms is to reduce the error and to reasonably consume a small space.

Many algorithms were proposed in the past; they differ in how the values are assigned to buckets.

In this work, we propose a new approach which is an algorithmic solution: Hconst (Histogram construction) Algorithm which tends to find and construct the optimal histogram: CM histogram.

This naming comes from the idea to ameliorate the existing version of compressed histogram by the principle of Maximum-difference histogram.

This approach reconciles the benefits of Max-diff histogram. We developed an experimental evaluation to underline the effectiveness and the accuracy of our algorithm and to prove that the error is lower than existing techniques.

One of the drawbacks of previous techniques in their accuracy is that the error rate is large, so we attempt to overcome this problem by introducing Hconst algorithm to construct CM histogram.

Our algorithm is applicable to minimize the error rate in query optimization.

We propose an improved version of compressed histogram; we will demonstrate that the concept of Max-diff extends to optimize the compressed histogram.

We realize the effectiveness and the advantages of Max-diff histogram to develop an approach and find a compromise between efficiency, accuracy and applicability.

Our idea is based on the principle of compressed histogram that affects the most frequent value in an individual bucket and applies equi-depth histogram on the remaining values.

The idea of CM histogram is assigning the highest frequency, i.e., the more occurring attribute value in an individual bucket. We apply Max-diff histogram to the remaining values.

According to the literature review, a lot of researches have shown with the experimental studies that Max-diff histogram is more accurate, and that is why we have the idea of applying Max-diff.

Remember that Max-diff histogram minimizes the maximum difference of adjacent source values [10].

Nigel Srikanth [11] has shown that max-diff histogram uses efficiently the Central Processing Unit CPU and memory. Kyung [12] stated that this histogram allows grouping the closest frequency since it inserts a boundary between values that have a maximum difference; so the estimation of the query size is more correct and precise.

V. PROBLEM FORMULATION

Consider [13], a relational table R which comprises n attributes $X_1 ... X_{n.}$

D: domains of attributes $X_1 ... X_{n.}$

Given a data set, find the histogram H associated with the attribute X with the smallest error:

Min Error (H)

So, our need is to find an efficient algorithm for constructing an accurate histogram.

The accuracy of the histogram relates to the accuracy of each bucket.

VI. HCONST ALGORITHM

We observe that any histogram contains an error; this error is due to the loss of information in their summary. There remains a need to find the optimal histogram; motivated by this, we propose a new technique to construct a histogram: CM histogram with a smaller error. Hconst algorithm fails to respond to this challenge.

Definition

Consider two histograms Hi and Hj which represent the distribution of a given dataset; we say that Hi is more accurate than Hj if and only if:

Error (Hi) < Error (Hj)

 ISBN: 978-1-61208-227-1

Theorem

A Max-difference bucket with a height h1 provides estimation more accurate than an Equi-depth bucket with a height h2 for all h1≤ h2.

Proof

From a set V of values with their corresponding set F of frequencies, we construct a maximum difference histogram M and an equi-depth histogram E, supposedly composed of a same number N of buckets.

Let (H_i^M)i=1 to N and (H_i^E)i=1 to N be the respective heights of the buckets (B_i^M)i=1 to N and (B_i^E)i=1 to N that compose the two histograms M and E.

Suppose that:

$H_i^E \geq H_i^M$ for a given $1 \leq i \leq N$.

To prove that estimation of the Histogram M is better than that of histogram E, it is sufficient to prove that:

Error $(B_i^E) \leq$ Error (B_i^M).

This inequality is verified using SSE metric since M, the max-diff histogram is already constructed by minimizing the variance, as this kind of histogram tends to group closer values, whereas equi-depth controls just the sum of the frequencies by bucket.

Hence $E_{abs}(M) \leq E_{abs}(E)$

We can conclude that for each value A, the estimation determined using Max-diff histogram is more accurate than equi-depth which justifies our choice of applying max-diff histogram instead of equi-depth.

For the case h1 > h2, we improve the accuracy of Max-diff by using the following algorithm:

Hconst Algorithm

Input: frequencies of each the attribute value

Output: the accurate histogram

1. Begin
2. Find (n, freqV,B)
3. maxDiff (remaining values, maxDiff histogram)

Optimization phase

4. For each Maxdiff bucket
5. If (exceptional bucket=True)
6. If H(BucketI-1) < heightMax
   BucketI-1← minVal
7. Else
8. If H(BucketI+1) < heightMax
   BucketI+1← maxVal
9. Return CM histogram
10. End;

This algorithm takes as input the different frequencies of each attribute value; later, there will be a call to the procedure Find to determine the highest frequencies; and then, there will be a call to the procedure max-Diff.

In the optimization phase, we reduce the height of the exceptional buckets.

We mean with Exceptional bucket whether the height bucket of max-diff histogram is greater than the height of equi-depth.

This phase proceeds as follows:

If the height of the previous bucket is lower than the maximal height; we migrate the minimum value in the bucket; else we migrate the maximum value to the next bucket.

We can change the location more than one value as we have not exceeded the maximal height.

As output, the result of the proposed algorithm will be an efficient and accurate histogram.

TABLE IV Time Complexity

Algorithm	Time Complexity
Procedure Find	O(N)
Procedure maxDiff	O(M)
Function Boundry	O(M)
Algorithm Hconst	O(N)

Where:

- N attribute value
- M remaining values

VII. EXPERIMENTAL RESULT

We attempt to prove that the theoretical results are confirmed in practice.

We investigated the effectiveness of the different histogram types cited above for estimating range query result sizes. The absolute errors due to the different histograms, as a function of the number of the bucket, are omputed based on a selectivity query applied on two data distribtions on the attribute salary from real database: National League Baseball Salaries for the years 2003 and 2005 to compare the performance of existing histograms; we assign the same number of buckets to different histograms.

The typical behavior of the histogram errors for the selection query applied respectively on the dataset for the year 2005 and 2003 are illustrated respectively in figure 5 and 6, with the number of bucket indicated on the x-axis and the absolute error indicated on the y- axis

As a visualization tool, we will use MATLAB, an example of a selection query on the National League Baseball Salaries dataset of the year 2005:

Select count (*)
Where salary =1000000;

The real frequency of the value 1000000 = 5

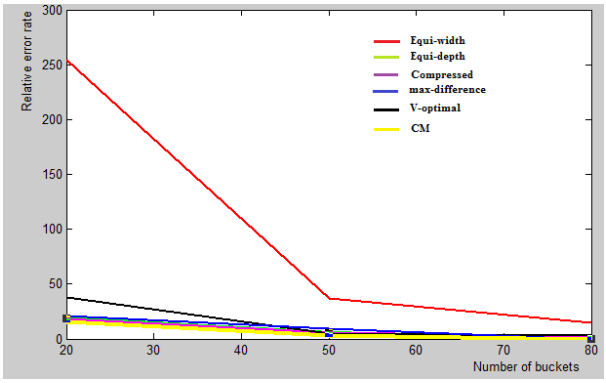


Figure 5. The Absolute Error of the first Dataset

We apply the same query on the second dataset corresponding to the year 2003 to better show the accuracy of our technique:

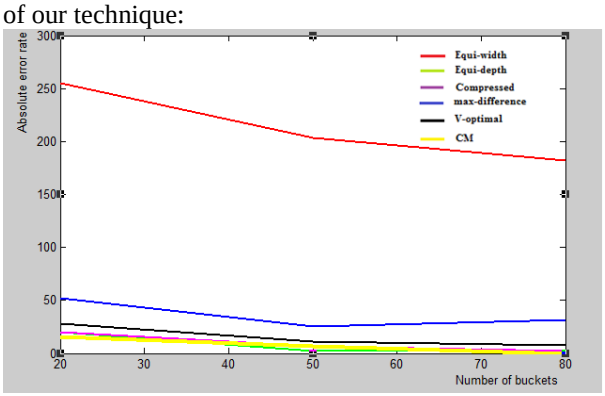


Figure 6. The Absolute Error of the second Dataset

In those plots, the number of buckets is varied. The error generated is proportional to the number of buckets. As shown in the two figures, the accuracy can be reached when increaing the number of buckets for all histogram types and the compressed, max-diff and v-optimal histograms are significantly better than the others that they show the least errir for different number of buckets. Moreover, the equi-width histogram exhibits the worst accuracy.

We show the efficiency of different traditional histograms, namely equi-width, equi-depth, compressed, V-opt, Max-diff and CM histogram obtained using our algorithm.

The results from the experiments show that the absolute error generated by our method is lower than the absolute error from existing histograms. This is a consequence of the fact that attribute values in our histogram are closer.

Those results do not only confirm our theoretical results presented in the previous chapter, but also confirm that the accuracy of our method is superior to that of previous histograms.

VIII. CONCLUSION

The use of histograms is widespread especially in approximating frequency distributions in data bases thanks to their simplicity and accuracy.

In our work, we studied and discussed various kinds of existing histograms; in addition to that, we have introduced a new technique for histogram constructing using an algorithm called Hconst.

We have also proposed a theorem to justify our technique.

Furthermore, we can deduce from the experimental comparisons that the histogram reduces the error; accordingly, we can confirm, and based on those experiments on a real database, that the quality of the histogram improves.

The identification of the optimal histogram remains an open field. As several new research opportunities appear, we will try to identify optimal histograms for different types of queries such as joins and non-equality joins, to limit not only the absolute error but also other metrics of error, to determine the appropriate number of buckets to build the optimal histogram and to find the histogram that can handle uncertain data.

And finally, we want to treat the problem of data stream which is the transmission of the flow of data that changes over time. Existing database systems do not process data streams efficiently; and this makes this area a popular search field [13].

REFERENCES

[1] C. Yu, G. Philip, and W. Meng, “Distributed top-N query processing with possibly uncooperative local systems,” In Proc, 29th VLDB Conference, 2003, pp 117-128.

[2] K. Chakrabarti, G. Minos, R. Rajeev, and S. Kyuseok, “Approximate Query processing using wavelets,” The VLDB Journal Vol. 10 Issue 2-3, 2001.

[3] V. Poosala, and Y. Ioannidis, “Improved Histograms for selectivity estimation of range predicates,” In Proc, 23rd VLDB Conference, 1997.

[4] Y. Ioannidis, “Query optimization,” ACM Computing Surveys, symposium issue on the 50th Anniversary of ACM, Vol. 28, 1996, pp. 121-123.

[5] S. Joseph, “Adaptive histogram algorithms for approximating frequency queries in dynamic data streams, “ world comp, 2011.

[6] V. Poosala, Y. Ioannidis, P. Haas, and E. Shekita, “Improved histograms for selectivity estimation of range predicates,” International ACM SIGMOD Conference, pp. 294-305, 1996.

[7] Y. Liu, “Data preprocessing,” Department of Biomedical, Industrial and Human Factors Engineering Wright State University, 2010.

[8] V. Jagadish, J. Hui, C. Beng, and T. Kian-Lee, “Global optimization of histograms,” ACM SIGMOD Vol. 30 Issue 2, 2001.

[9] X. Lin, and Q. Zhang, “Error minimization for approximate computation of range aggregates,” Proceedings of the Eighth International Conference on Database Systems for Advanced Applications IEEE Computer Society, 2003.

[10] B. Zina, J. Liu, B. Omran , L. Huian , M. Jesse, B. Chavali , and O. Robert, “Use and Maintenance of Histograms for Large Scientific Database Access Planning: A Case Study of a Pharmaceutical Data,” Repository Journal of Intelligent Information Systems, 2004.

 ISBN: 978-1-61208-227-1

[11] E. Nigel, and A. Srikanth, “Query planning using a maxdiff histogram, “ Microsoft Corporation, 2000.

[12] H. Kyoung, “Query Size Estimation through Sampling,” phd thesis, North Carolina State University, 2005.
(http://repository.lib.ncsu.edu/ir/handle/1840.16/1274)

[13] V. Jagadish, V. Poosala, K. Nick, S. Ken, S. Muthukrishnan, and S. Torsten, “Optimal Histograms with Quality Guarantees,” VLDB Proceedings, 1998.

[14] P. Pawluk, “Stream Databases,” PhD thesis, York University Department of Computer Science and Engineering, 2006.
(http://www.bth.se/fou/cuppsats.nsf/all/e1571a10dc340e51c12571bc005ddc43/$file/prpa05-master_thesis.pdf).

 ISBN: 978-1-61208-227-1